

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*

Supplementary Material

This supplementary material is organized as follows. In Sections S1 and S2, we present additional numerical results from the simulation and the application, respectively. In Section S3, we present the proofs for Section 3.1, Lemmas 1-2 and Theorem 1.

S1 Simulation Results

In the simulation, we considered four data generation schemes. Table S1 lists detailed settings of the four schemes, and Figure S1 illustrates the corresponding data structures.

Tables S2-S4 present the performances of the estimation methods. Specifically, Tables S2-S3 present the estimated value function $V_{\mathcal{I}}^{\pi}(\mathcal{G})$ of the two

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

Table S1: Settings of Schemes 1-4.

	Settings		Properties			
	S_{k+1}	X_{k+1}	(A1)&(A2)	$X_{k+1} \perp S_{k+1}$	$X_{k+1} \perp S_{k+1} \mathcal{F}(T_k)$	model(3.1)
Scheme 1	$S1$	$X1$	✓	✓	✓	✓
Scheme 2	$S2$	$X2$	✓	×	✓	✓
Scheme 3	$S2$	$X3$	✓	×	×	✓
Scheme 4	$S3$	$X2$	✓	×	×	×

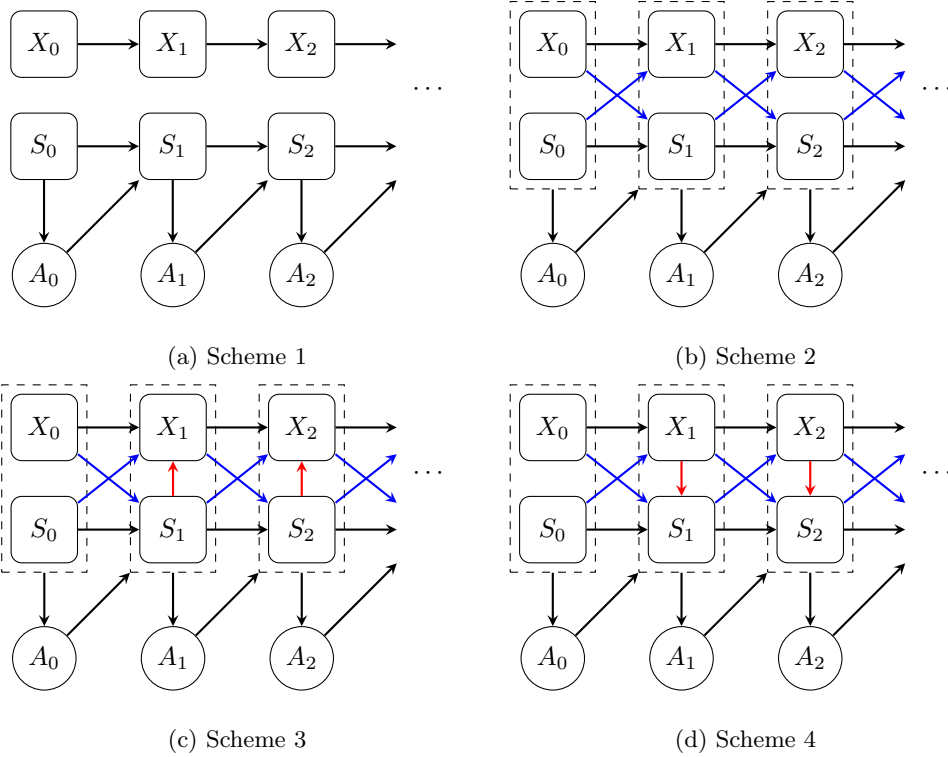


Figure S1: Data structures of schemes 1-4.

target policies when $\lambda_0(x) = 0.5x$, behavior policy $b(a; s, x)$ follows distribution (b1), and the observed data are generated by Schemes 1-3. The estimation results when $\lambda_0(x) = 1.5$ and $0.5x^2$, and $b(a; s, x)$ follows dis-

tributions (b2) and (b3), are given in Table S4. Within each scenario, the discount factor is set as $\gamma = 0.7$ and 0.9 separately, the reference distribution is set as $\mathcal{G}(s, x) = 0.5(5s + 1)x$ with $-0.2 \leq s \leq 0.2$ and $0 \leq x \leq 1$, and four types of trajectories are considered by setting the number of trajectories $n = 100, 200$ with the length of trajectory $K = 10$, and setting $K = 100, 200$ when $n = 10$. In each case, we generate 600 simulation replicates to approximate the bias (Bias), standard deviation (SD), estimated standard error (SE), and empirical coverage probability of 95% confidence intervals (CP) for the proposed estimators, as well as the bias and standard deviation for the MDP and SMDP estimators. Table S4 presents the policy evaluation results for $\pi(s, x) = I(s < x - 1)$ when the baseline hazard $\lambda_0(x) = 1.5$ and $0.5x^2$, and the behaviour policy $b(a; s, x)$ follows (b2) and (b3). For brevity, we only report the results when the datasets are generated by scheme 3 and the discount is set as $\gamma = 0.7$. Simulation results in the other scenarios, (i.e, under schemes 1,2,4 and $\gamma = 0.9$), are similar and thus omitted.

To investigate the robustness of the proposed estimates, Table S5 summarizes policy evaluation results for $\pi(a; s, x) = aI(s < 0) + (1 - a)I(s \geq 0)$ and $\pi(a; s, x) = aI(s < x + 1) + (1 - a)I(s \geq x + 1)$ when the observation process model is misspecified, including the bias (Bias), standard deviation

(SD), estimated standard error (SE), and empirical coverage probability of 95% confidence intervals (CP) for standardized estimators and modulated estimators, and the bias and standard deviation of the naive estimators.

Table S2: Policy evaluation results for $\pi(a; s, x) = aI(s < 0) + (1 - a)I(s \geq 0)$.

γ	n	K	MDP		SMDP		Proposed			
			<i>Bias</i>	<i>SD</i>	<i>Bias</i>	<i>SD</i>	<i>Bias</i>	<i>SD</i>	<i>SE</i>	<i>CP</i>
Scheme 1										
0.7	10	100	-0.241	0.111	-0.003	0.070	0.001	0.119	0.100	0.945
	10	200	-0.235	0.074	-0.002	0.042	-0.002	0.073	0.069	0.941
	100	10	-0.234	0.110	0.000	0.067	0.001	0.115	0.096	0.923
	200	10	-0.240	0.073	-0.003	0.045	-0.001	0.078	0.068	0.947
0.9	10	100	-1.081	0.382	-0.217	0.283	0.002	0.330	0.274	0.927
	10	200	-1.046	0.249	-0.215	0.164	0.037	0.207	0.191	0.925
	100	10	-1.045	0.363	-0.289	0.253	0.054	0.298	0.265	0.906
	200	10	-1.066	0.252	-0.226	0.177	0.022	0.220	0.191	0.933
Scheme 2										
0.7	10	100	-0.159	0.077	-0.281	0.071	-0.005	0.101	0.082	0.912
	10	200	-0.154	0.053	-0.270	0.049	-0.001	0.065	0.059	0.947
	100	10	-0.164	0.084	-0.275	0.073	-0.010	0.102	0.084	0.925
	200	10	-0.168	0.051	-0.274	0.048	-0.013	0.068	0.062	0.930
0.9	10	100	-0.849	0.259	-1.091	0.281	-0.192	0.301	0.251	0.928
	10	200	-0.835	0.189	-1.005	0.213	0.005	0.209	0.178	0.945
	100	10	-0.855	0.282	-1.043	0.310	-0.027	0.268	0.259	0.948
	200	10	-0.863	0.189	-1.037	0.206	-0.035	0.200	0.194	0.942
Scheme 3										
0.7	10	100	-0.157	0.075	-0.141	0.073	-0.004	0.100	0.082	0.912
	10	200	-0.155	0.052	-0.278	0.050	0.014	0.065	0.058	0.948
	100	10	-0.162	0.078	-0.286	0.073	-0.009	0.102	0.082	0.925
	200	10	-0.166	0.051	-0.287	0.047	-0.012	0.065	0.061	0.928
0.9	10	100	-0.776	0.266	-1.055	0.283	-0.029	0.288	0.249	0.928
	10	200	-0.845	0.185	-1.047	0.209	0.003	0.200	0.177	0.944
	100	10	-0.865	0.278	-1.060	0.304	-0.028	0.275	0.253	0.941
	200	10	-0.873	0.186	-1.073	0.205	-0.036	0.182	0.188	0.938

S2. ADDITIONAL APPLICATION RESULTS

Table S3: Policy evaluation results for $\pi(a; s, x) = aI(s < x - 1) + (1 - a)I(s \geq x - 1)$.

γ	n	K	MDP		SMDP		Proposed			
			<i>Bias</i>	<i>SD</i>	<i>Bias</i>	<i>SD</i>	<i>Bias</i>	<i>SD</i>	<i>SE</i>	<i>CP</i>
Scheme 1										
0.7	10	100	-0.172	0.112	-0.147	0.069	-0.001	0.119	0.103	0.942
	10	200	-0.171	0.077	-0.148	0.041	-0.005	0.086	0.074	0.941
	100	10	-0.168	0.121	-0.143	0.073	-0.007	0.119	0.105	0.938
	200	10	-0.174	0.081	-0.148	0.047	-0.006	0.084	0.077	0.953
0.9	10	100	-0.928	0.422	-0.536	0.284	0.014	0.296	0.281	0.941
	10	200	-0.914	0.264	-0.517	0.184	-0.001	0.217	0.197	0.950
	100	10	-0.915	0.398	-0.507	0.281	-0.019	0.299	0.276	0.944
	200	10	-0.915	0.247	-0.520	0.183	-0.018	0.218	0.199	0.938
Scheme 2										
0.7	10	100	-0.007	0.082	-0.177	0.049	-0.010	0.100	0.092	0.942
	10	200	-0.004	0.053	-0.176	0.031	-0.005	0.075	0.066	0.945
	100	10	-0.008	0.079	-0.175	0.049	-0.009	0.108	0.093	0.944
	200	10	-0.015	0.055	-0.178	0.033	-0.013	0.071	0.069	0.952
0.9	10	100	-0.214	0.279	-0.444	0.205	-0.085	0.279	0.269	0.938
	10	200	-0.193	0.176	-0.439	0.133	-0.040	0.208	0.190	0.953
	100	10	-0.197	0.250	-0.430	0.192	-0.074	0.280	0.267	0.958
	200	10	-0.211	0.180	-0.438	0.142	-0.056	0.202	0.194	0.938
Scheme 3										
0.7	10	100	-0.015	0.084	-0.200	0.049	-0.011	0.101	0.093	0.941
	10	200	-0.012	0.055	-0.199	0.033	-0.006	0.076	0.066	0.944
	100	10	-0.016	0.078	-0.198	0.048	-0.011	0.110	0.094	0.938
	200	10	-0.023	0.057	-0.202	0.034	-0.014	0.070	0.069	0.948
0.9	10	100	-0.262	0.276	-0.499	0.211	-0.102	0.293	0.270	0.934
	10	200	-0.246	0.174	-0.514	0.132	-0.052	0.207	0.191	0.955
	100	10	-0.258	0.241	-0.514	0.197	-0.088	0.283	0.266	0.939
	200	10	-0.270	0.180	-0.524	0.132	-0.067	0.193	0.192	0.941

S2 Additional Application Results

In this section, we present additional application results to compare the observed stochastic policy, the observed fixed policy, and the optimal pol-

Table S4: Policy evaluation results for $\pi(a; s, x) = aI(s < x - 1) + (1 - a)I(s \geq x - 1)$, when $\lambda_0(x) = 1.5$ and $0.5x^2$, and $b(a; s, x)$ follows (b2) and (b3).

			MDP		SMDP		Proposed			
	n	K	$Bias$	SD	$Bias$	SD	$Bias$	SD	SE	CP
$\lambda_0(x)$			behaviour policy ($b1$)							
1.5	10	100	-0.017	0.046	0.451	0.080	-0.017	0.060	0.059	0.944
	10	200	-0.016	0.030	0.459	0.054	-0.008	0.038	0.042	0.953
	100	10	-0.016	0.049	0.462	0.088	-0.012	0.062	0.059	0.931
	200	10	-0.017	0.034	0.458	0.061	-0.009	0.046	0.042	0.944
$0.5x^2$	10	100	0.008	0.159	-0.123	0.121	0.033	0.260	0.215	0.919
	10	200	-0.060	0.086	-0.129	0.065	0.006	0.162	0.141	0.930
	100	10	-0.058	0.138	-0.126	0.102	-0.035	0.256	0.219	0.922
	200	10	-0.074	0.081	-0.137	0.060	-0.012	0.153	0.141	0.928
$b(a; s, x)$			baseline hazard $\lambda_0(x) = 0.5x$							
$(b2)$	10	100	-0.009	0.089	-0.209	0.052	-0.003	0.107	0.097	0.934
	10	200	-0.019	0.059	-0.200	0.035	-0.009	0.070	0.067	0.944
	100	10	0.046	0.085	-0.203	0.053	-0.003	0.106	0.094	0.955
	200	10	-0.019	0.055	-0.199	0.032	-0.003	0.077	0.069	0.925
$(b3)$	10	100	-0.023	0.074	-0.198	0.048	-0.008	0.095	0.085	0.948
	10	200	-0.032	0.050	-0.205	0.033	-0.001	0.064	0.060	0.967
	100	10	-0.013	0.074	-0.205	0.046	-0.004	0.092	0.085	0.959
	200	10	-0.017	0.050	-0.200	0.029	-0.006	0.067	0.061	0.947

icy. The policies are estimated by 2488 visit records from 352 patients (approximately 33% of the total study subjects) randomly selected from the full dataset, while the policy evaluation procedure uses the remaining 5166 records from 704 patients.

S2. ADDITIONAL APPLICATION RESULTS

Table S5: Policy evaluation results for $\pi(a; s, x) = aI(s < 0) + (1 - a)I(s \geq 0)$ and $\pi(a; s, x) = aI(s < x + 1) + (1 - a)I(s \geq x + 1)$ when the observation process model is misspecified.

γ	n	K	Classic RL		SMDP		Proposed			
			$Bias$	SD	$Bias$	SD	$Bias$	SD	SE	CP
$\pi(a; s, x) = aI(s < 0) + (1 - a)I(s \geq 0)$										
0.7	10	100	-0.199	0.081	-0.293	0.071	-0.001	0.094	0.086	0.941
	10	200	-0.200	0.051	-0.295	0.045	-0.008	0.070	0.063	0.936
	100	10	-0.210	0.088	-0.295	0.074	-0.007	0.110	0.096	0.938
	200	10	-0.210	0.062	-0.294	0.051	-0.013	0.077	0.072	0.950
0.9	10	100	-0.962	0.268	-1.172	0.306	-0.010	0.268	0.270	0.941
	10	200	-0.968	0.176	-1.169	0.204	-0.033	0.212	0.199	0.956
	100	10	-0.995	0.300	-1.183	0.312	-0.042	0.341	0.318	0.952
	200	10	-0.992	0.219	-1.176	0.226	-0.069	0.227	0.247	0.958
$\pi(a; s, x) = aI(s < x + 1) + (1 - a)I(s \geq x + 1)$										
0.7	10	100	-0.021	0.084	-0.164	0.051	-0.005	0.109	0.097	0.955
	10	200	-0.009	0.053	-0.163	0.033	-0.008	0.081	0.070	0.941
	100	10	-0.030	0.098	-0.168	0.053	-0.009	0.124	0.108	0.944
	200	10	-0.029	0.059	-0.166	0.034	-0.011	0.082	0.078	0.961
0.9	10	100	-0.182	0.279	-0.360	0.223	-0.074	0.298	0.291	0.959
	10	200	-0.152	0.182	-0.325	0.137	-0.056	0.201	0.206	0.947
	100	10	-0.189	0.282	-0.351	0.207	-0.079	0.318	0.305	0.947
	200	10	-0.177	0.189	-0.343	0.141	-0.057	0.227	0.221	0.964

To obtain the observed stochastic policy, we first fit a logistic regression model with A_k being the (binary) response and (S_k, X_k) being the explanatory variables. Let the estimated probability of $A_k = a$ given (s, x)

denoted by $\pi_{os}(a; s, x)$, and we refer to π_{os} as the observed stochastic policy. Moreover, by setting a probability threshold of 0.48, we further obtain a deterministic function mapping from (s, x) to $\{0, 1\}$, and denote it as the observed fixed policy $\pi_{of}(a; s, x)$. Finally, the optimal policy is estimated using double-fitted Q-learning (Shi et al., 2022), where the corresponding Bellman equation can be derived by a straightforward extension of Lemma 2. Plots of the obtained policies are shown in Figure S2. From the plots, it can be seen that 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 intervals. In contrast, under the estimated optimal policy, there is no such clear linear relationship among actions, states, and gap times.

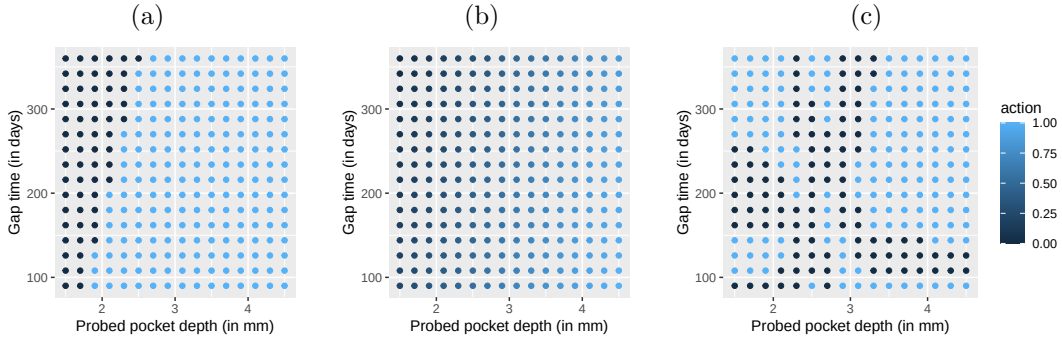


Figure S2: Plots of policies, including (a) observed fixed policy π_{of} ; (b) observed stochastic policy π_{os} ; (c) estimated optimal policy π_{op} .

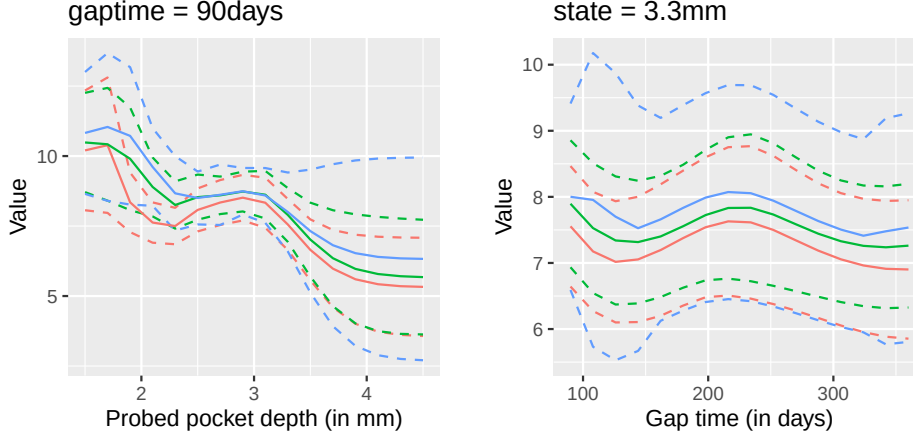


Figure S3: The estimated value (solid line) and confidence intervals (dashed lines) for observed fixed policy π_{of} (red line), observed stochastic policy π_{os} (green line), and estimated optimal policy π_{op} (blue line).

Apart from the results presented in Section 4, we also compare the confidence intervals of the value functions for these policies using the modulated estimation method. The estimated results are presented in Figure S3. Together with Figure S2, we find that changes in the action assignments under the optimal policy, compared with the observed policies, indeed improve the policy's value. For example, while the observed policies both suggest that all patients with $X_{k-1} \leq 90$ days should revisit more frequently (i.e $A_k = 1$), the estimated optimal policy π_{op} suggests that patients with a gap time of 90 days and probed pocket depth values less than 2.7 mm can schedule their subsequent visit at six months or later (i.e $A_k = 0$). Correspond-

ingly, the value of the optimal policy at a gap time of 90 days and probed pocket depth values below 2.2 mm is significantly higher than that of the two observed policies.

S3 Proofs

S3.1 Proofs for the equivalence between (A2) and (A2')

Firstly, if the distribution $\mathcal{P}_X$ in (A2) is absolutely continuous with respect to the Lebesgue measure, there exists a density function $p_X(x; S_{k+1}, S_k, X_k, A_k)$ satisfying $p_X(x; S_{k+1}, S_k, X_k, A_k)dx = d\mathcal{P}_X(x; S_{k+1}, S_k, X_k, A_k)$. Then under (A2), the hazard of X_{k+1} conditional on $\{\{S_j, X_j, A_j, T_j\}_{j \leq k}, S_{k+1}\}$ takes the form

$$\begin{aligned} & \lim_{\Delta \rightarrow 0+} \frac{1}{\Delta} P(x \leq X_{k+1} \leq x + \Delta | X_{k+1} \geq x, \{S_j, X_j, A_j, T_j\}_{j \leq k}, S_{k+1}) \\ &= \frac{p_X(x; S_{k+1}, S_k, X_k, A_k)}{1 - \mathcal{P}_X(x; S_{k+1}, S_k, X_k, A_k)}. \end{aligned}$$

which is a function of x and $\{S_{k+1}, S_k, X_k, A_k\}$. Denote the hazard as $\lambda(x; S_{k+1}, S_k, X_k, A_k)$. Then $N(t) = \sum_{j \geq 1} I(T_j \leq t)$ is a right-continuous

counting process satisfies

$$\begin{aligned}
& \lim_{\Delta \rightarrow 0+} \frac{1}{\Delta} P(N(t+\Delta) - N(t-) = 1 | \mathcal{F}(t-), S_{N(t-)+1}) \\
&= \lim_{\Delta \rightarrow 0+} \frac{1}{\Delta} P(N(t+\Delta) - N(t-) = 1 | N(t-), \{S_j, X_j, A_j, T_j\}_{j \leq N(t-)}, S_{N(t-)+1}) \\
&= \sum_{k=1}^{\infty} I(N(t-) = k) \lim_{\Delta \rightarrow 0+} \frac{1}{\Delta} P(N(t+\Delta) - N(t-) = 1 | N(t-) = k, \\
&\quad \{S_j, X_j, A_j, T_j\}_{j \leq N(t-)}, S_{N(t-)+1}) \\
&= \sum_{k=1}^{\infty} I(N(t-) = k) \lim_{\Delta \rightarrow 0+} \frac{1}{\Delta} P(t \leq T_k \leq t + \Delta | T_k < t, \{S_j, X_j, A_j, T_j\}_{j \leq k}, S_{k+1}) \\
&= \sum_{k=1}^{\infty} I(N(t-) = k) \lim_{\Delta \rightarrow 0+} \frac{1}{\Delta} P(t - T_k \leq X_k \leq t - T_k + \Delta | X_{k+1} \geq t - T_k, \\
&\quad T_k < t, \{S_j, X_j, A_j, T_j\}_{j \leq k}, S_{k+1}) \\
&= \sum_{k=1}^{\infty} I(N(t-) = k) \lambda(t - T_k; S_{k+1}, S_k, X_k, A_k) \\
&= \lambda(t - T_{N(t-)}; S_{N(t-)+1}, S_{N(t-)}, X_{N(t-)}, A_{N(t-)}),
\end{aligned}$$

and thus assumption (A2') follows directly.

On the other hand, if a right-continuous counting process $\{N(t)\}_{t \geq 0}$ with jumps at $\{T_k\}_{k \geq 1}$ satisfies assumption (A2'), then

$$\begin{aligned}
E[dN(t) | \mathcal{F}(t-), S_{k+1}] &= E[dN(t) | N(t-), \{S_j, X_j, A_j, T_j\}_{j \leq k}, S_{k+1}] \\
&= \lambda(t - T_k; S_{k+1}, S_k, X_k, A_k) dt,
\end{aligned}$$

for $T_k < t \leq T_{k+1}$. With similar arguments, it can be obtained that the

corresponding gap times $\{X_{k+1}\}_{k \geq 0}$ also satisfies

$$\begin{aligned}
& \lim_{\Delta \rightarrow 0+} \frac{1}{\Delta} P(x \leq X_{k+1} < x + \Delta | X_{k+1} \geq x, \mathcal{F}(t-), S_{k+1}) \\
&= \lim_{\Delta \rightarrow 0+} \frac{1}{\Delta} P(x \leq X_{k+1} < x + \Delta | X_{k+1} \geq x, \{S_j, X_j, A_j, T_j\}_{0 \leq j \leq k}, S_{k+1}) \\
&= \lambda(x; S_{k+1}, S_k, X_k, A_k).
\end{aligned}$$

Thus the distribution of X_{k+1} conditional on $\{S_j, X_j, A_j, T_j\}_{0 \leq j \leq k}$ and S_{k+1} takes the form of

$$P(X_{k+1} \leq x | \{S_j, X_j, A_j, T_j\}_{j \leq k}, S_{k+1}) = \exp\left\{-\int_0^x \lambda(u; S_{k+1}, S_k, X_k, A_k) du\right\},$$

which satisfies the the Markov and time-homogeneous property defined in (A2). ■

S3.2 Proof of Lemmas 1-2

For notational brevity, in the following subsections, we assume that π is a deterministic policy and denote $\pi(s, x) \in \mathbb{A}$ as the selected action given state $s \in \mathbb{S}$ and gap time $x \in \mathbb{R}^+$. All derivations and results can be straightforwardly generalized to the case of stochastic policies.

Proof of Lemma 1. Under assumptions (A1)-(A3), the existence of $Q_{\mathcal{I},k}^\pi(s, x, a, t)$ follows directly from the boundedness of r and $\lim_{j \rightarrow \infty} T_j = \infty$. Following the reformulation of $Q_{\mathcal{I},k}^\pi(s, x, a, t)$ discussed in Section 3.1,

we further rewrite $Q_{\mathcal{I},k}^\pi$ by

$$\begin{aligned}
& \sum_{j \geq k+1} E^\pi \{ \gamma^{T_j - T_k} R_{\mathcal{I}}(T_j) | S_k = s, X_k = x, A_k = a, T_k = t \} \\
&= \lim_{M \rightarrow \infty} \sum_{j=k+1}^M E^\pi \{ \gamma^{X_{k+1} + \dots + X_j} R_{\mathcal{I}}(T_j) | S_k = s, X_k = x, A_k = a, T_k = t \} \\
&= \lim_{M \rightarrow \infty} \left[E \{ \gamma^{X_{k+1}} R_{\mathcal{I}}(T_{k+1}) | S_k = s, X_k = x, A_k = a, T_k = t \} + \sum_{j=k+2}^M E^\pi \{ \right. \\
&\quad \left. \gamma^{X_{k+1} + \dots + X_{j-1}} \gamma^{X_j} R_{\mathcal{I}}(T_j) | S_k = s, X_k = x, A_k = a, T_k = t \} \right] \\
&= \lim_{M \rightarrow \infty} \left(E \{ \gamma^{X_{k+1}} R_{\mathcal{I}}(T_{k+1}) | S_k = s, X_k = x, A_k = a, T_k = t \} + \sum_{j=k+2}^M E^\pi [\right. \\
&\quad \left. \gamma^{X_{k+1} + \dots + X_{j-1}} E \{ \gamma^{X_j} R_{\mathcal{I}}(T_j) | S_{j-1}, X_{j-1}, A_{j-1}, S_k, X_k, A_k, T_k \} | S_k = s, \right. \\
&\quad \left. X_k = x, A_k = a, T_k = t \} \right)
\end{aligned}$$

When the observation times are regularly spaced, $E \{ \gamma^{X_{k+1}} R(T_{k+1}) | S_k = s, X_k, A_k, T_k \}$ can be simplified to $\gamma E \{ R(T_{k+1}) | S_k, X_k, A_k, T_k \}$, and thus the time-homogeneous property of the Q-function follows from the conventional conditional mean independence assumption (Shi et al., 2022). However, when the observation times are irregularly spaced and outcome-dependent, stronger assumptions are required to handle the conditional expectation of $\gamma^{X_{k+1}} R(T_{k+1}) / \lambda(X_{k+1}; Z_k)$. Therefore, in Section 2.1, we propose a modified conditional mean independence assumption (A3) specifying the conditional independence between X_{k+1} and $R(T_{k+1})$. By the proposed assumptions

(A1)-(A3) and the definition of $R_{\mathcal{I}}$,

$$\begin{aligned}
& E\{\gamma^{X_{k+1}} R_{\mathcal{I}}(T_{k+1}) | S_k = s, X_k = x, A_k = a, T_k\} \\
&= E\left\{ \frac{\gamma^{X_{k+1}} R(T_{k+1})}{\lambda(X_{k+1}; S_{k+1}, S_k, X_k, A_k)} \middle| S_k = s, X_k = x, A_k = a, T_k \right\} \\
&= E\left[\frac{\gamma^{X_{k+1}} E\{R(T_{k+1}) | S_{k+1}, X_{k+1}, S_k, X_k, A_k, T_k\}}{\lambda(X_{k+1}; S_{k+1}, S_k, X_k, A_k)} \middle| S_k = s, X_k = x, A_k = a, T_k = t \right] \\
&= E\left[\frac{\gamma^{X_{k+1}} r(S_k, X_k, A_k, S_{k+1}, X_{k+1})}{\lambda(X_{k+1}; S_{k+1}, S_k, X_k, A_k)} \middle| S_k = s, X_k = x, A_k = a, T_k = t \right] \\
&= \iint_{\mathbb{S} \times [0, \infty)} \frac{\gamma^{x'} r(s, x, a, s', x')}{\lambda(x'; s', s, x, a)} d\mathcal{P}(s', x'; s, x, a) =: g_1(s, x, a),
\end{aligned}$$

and

$$\begin{aligned}
& E\{\gamma^{X_j} R_{\mathcal{I}}(T_j) | S_{j-1}, X_{j-1}, A_{j-1}, S_k, X_k, A_k, T_k\} \\
&= E\left[\frac{\gamma^{X_j} E\{R(T_j) | S_j, X_j, S_{j-1}, X_{j-1}, A_{j-1}, S_k, X_k, A_k, T_k\}}{\lambda(X_j; S_j, S_{j-1}, X_{j-1}, A_{j-1})} \middle| S_{j-1}, X_{j-1}, A_{j-1}, \right. \\
&\quad \left. S_k, X_k, A_k, T_k \right] \\
&= E\left\{ \frac{\gamma^{X_j} r(S_{j-1}, X_{j-1}, A_{j-1}, S_j, X_j)}{\lambda(X_j; S_j, S_{j-1}, X_{j-1}, A_{j-1})} \middle| S_{j-1}, X_{j-1}, A_{j-1} \right\} \\
&= g_1(S_{j-1}, X_{j-1}, A_{j-1}).
\end{aligned}$$

for $j \geq k + 2$. Thus, $Q_{L,k}^\pi(s, x, a, t)$ can be further rewritten as

$$\begin{aligned}
& \lim_{M \rightarrow \infty} \left(g_1(s, x, a) + \sum_{j=k+2}^M E^\pi \left[\gamma^{X_{k+1} + \dots + X_{j-1}} g_1\{S_{j-1}, X_{j-1}, \pi(S_{j-1}, X_{j-1})\} \middle| \right. \right. \\
& \quad \left. \left. S_k = s, X_k = x, A_k = a, T_k = t \right] \right) \\
&= \lim_{M \rightarrow \infty} \left(g_1(s, x, a) + E \left[\gamma^{X_{k+1}} g_1\{S_{k+1}, X_{k+1}, \pi(S_{k+1}, X_{k+1})\} \middle| S_k = s, X_k = x, \right. \right. \\
& \quad \left. \left. A_k = a, T_k = t \right] + \sum_{j=k+3}^M E^\pi \left\{ \gamma^{X_{k+1} + \dots + X_{j-2}} E \left[\gamma^{X_{j-1}} g_1\{S_{j-1}, X_{j-1}, \pi(S_{j-1}, X_{j-1})\} \middle| \right. \right. \\
& \quad \left. \left. S_{j-2}, X_{j-2}, A_{j-2}, S_k, X_k, A_k, T_k \right] \middle| S_k = s, X_k = x, A_k = a, T_k = t \right\} \right)
\end{aligned}$$

Using similar arguments, we have

$$\begin{aligned}
& E \left[\gamma^{X_{k+1}} g_1\{S_{k+1}, X_{k+1}, \pi(S_{k+1}, X_{k+1})\} \middle| S_k = s, X_k = x, A_k = a, T_k \right] \\
&= \iint_{\mathbb{S} \times [0, \infty)} \gamma^{x'} g_1\{s', x', \pi(s', x')\} d\mathcal{P}(s', x'; s, x, a) =: g_2^\pi(s, x, a),
\end{aligned}$$

which is also independent of $\{T_k, k\}$. For $j \geq 2$, define

$$g_{j+1}^\pi(s, x, a) \equiv \iint_{\mathbb{S} \times [0, \infty)} \gamma^{x'} g_j^\pi\{s', x', \pi(s', x')\} d\mathcal{P}(s', x'; s, x, a).$$

Then, we have

$$\begin{aligned}
& Q_{\mathcal{I},k}^\pi(s, x, a, t) \\
&= \lim_{M \rightarrow \infty} \left(g_1(s, x, a) + g_2^\pi(s, x, a) + \sum_{j=k+3}^M E^\pi [\gamma^{X_{k+1} + \dots + X_{j-2}} \right. \\
&\quad \left. g_2^\pi \{S_{j-2}, X_{j-2}, \pi(S_{j-2}, X_{j-2})\} | S_k = s, X_k = x, A_k = a, T_k = t] \right) \\
&= \lim_{M \rightarrow \infty} \left(g_1(s, x, a) + g_2^\pi(s, x, a) + g_3^\pi(s, x, a) + \sum_{j=k+4}^M E^\pi [\gamma^{X_{k+1} + \dots + X_{j-3}} \right. \\
&\quad \left. g_3^\pi \{S_{j-3}, X_{j-3}, \pi(S_{j-3}, X_{j-3})\} | S_k = s, X_k = x, A_k = a, T_k = t] \right) \\
&= \dots \\
&= \lim_{M \rightarrow \infty} \left\{ g_1(s, x, a) + \sum_{j=2}^{M-k} g_j^\pi(s, x, a) \right\} = g_1(s, x, a) + \sum_{j \geq 2} g_j^\pi(s, x, a).
\end{aligned}$$

Therefore, $Q_{\mathcal{I},k}^\pi(s, x, a, t) = Q_{\mathcal{I},0}^\pi(s, x, a, 0)$ for all $k \geq 0$, $t \geq 0$, $s \in \mathbb{S}$,

$x \in \mathbb{R}^+$ and $a \in \mathbb{A}$. ■

Proof of Lemma 2 (Bellman equation). Following the discussion of assumption (A3) presented in Section 2.1, from assumptions (A1)-(A3), we obtain that $E[R_{\mathcal{I}}(T_{k+1}) | \{S_j, X_j, A_j\}_{j \leq k}] = E\{R_{\mathcal{I}}(T_{k+1}) | S_k, X_k, A_k\}$. Then

$Q_{\mathcal{I}}^{\pi}(S_k, X_k, A_k)$ can be rewritten as

$$\begin{aligned}
& E^{\pi} \left\{ \sum_{j \geq k+1} \gamma^{X_{k+1} + \dots + X_j} R_{\mathcal{I}}(T_j) \middle| S_k, X_k, A_k \right\} \\
&= E \left\{ \gamma^{X_{k+1}} R_{\mathcal{I}}(T_{k+1}) \middle| S_k, X_k, A_k \right\} + E^{\pi} \left[\gamma^{X_{k+1}} E^{\pi} \left\{ \sum_{j \geq k+2} \gamma^{X_{k+2} + \dots + X_j} R_{\mathcal{I}}(T_j) \middle| \right. \right. \\
&\quad \left. \left. S_{k+1}, X_{k+1}, A_{k+1}, S_k, X_k, A_k \right\} \middle| S_k, X_k, A_k \right] \\
&= E \left\{ \gamma^{X_{k+1}} R_{\mathcal{I}}(T_{k+1}) \middle| S_k, X_k, A_k \right\} + E^{\pi} \left[\gamma^{X_{k+1}} E^{\pi} \left\{ \sum_{j \geq k+2} \gamma^{X_{k+2} + \dots + X_j} R_{\mathcal{I}}(T_j) \middle| \right. \right. \\
&\quad \left. \left. S_{k+1}, X_{k+1}, A_{k+1} \right\} \middle| S_k, X_k, A_k \right] \\
&= E \left\{ \gamma^{X_{k+1}} R_{\mathcal{I}}(T_{k+1}) \middle| S_k, X_k, A_k \right\} + E^{\pi} \left\{ \gamma^{X_{k+1}} Q_{\mathcal{I}}^{\pi}(S_{k+1}, X_{k+1}, A_{k+1}) \middle| S_k, X_k, A_k \right\} \\
&= E \left\{ \gamma^{X_{k+1}} R_{\mathcal{I}}(T_{k+1}) \middle| S_k, X_k, A_k \right\} + E \left[\gamma^{X_{k+1}} Q_{\mathcal{I}}^{\pi} \{ S_{k+1}, X_{k+1}, \pi(S_{k+1}, X_{k+1}) \} \middle| S_k, X_k, A_k \right].
\end{aligned}$$

Therefore, the Bellman equation given in Lemma 2 holds for all $k \geq 1$. ■

S3.3 Proof of Theorem 1

For notational brevity, let $\sum_{i,k}$ denote $\sum_{i=1}^n \sum_{k=0}^{K_i-1}$, let $Q_{\mathcal{I}}^{\pi}(s, x, \pi)$ denote $Q_{\mathcal{I}}^{\pi}\{s, x, \pi(s, x)\}$, let $Q(s, x, \pi; \theta)$ denote $Q\{s, x, \pi(s, x); \theta\}$, and let $\xi(s, x, \pi; \theta)$ denote $\xi\{s, x, \pi(s, x); \theta\}$. For a set of l -dimensional vectors $\mathcal{V} = \{v_{i,k} \in \mathbb{R}^l\}_{1 \leq i \leq n, 0 \leq k \leq K_i-1}$, let $array(\mathcal{V})$ denote a $N_K \times l$ dimension array, whose $(\sum_{j < i} K_j + k + 1)$ th row is $N_K^{1/2} v_{i,k}$, and taking the form

$$array(\mathcal{V}) = N_K^{1/2} (v_{1,0}, \dots, v_{1,K_1-1}, \dots, v_{i,k}, \dots, v_{n,0}, \dots, v_{n,K_n-1})^{\top}.$$

The following notations and conditions are required in the proof of Theorem 1. Let $p_S(s'; s, x, a)$ and $p_X(x'; s', s, x, a)$ denote the density functions of S_{k+1} and X_{k+1} that correspond to $\mathcal{P}_S(\cdot|s, x, a)$ in (A1) and $\mathcal{P}_X(\cdot|s', s, x, a)$ in (A2), respectively. Then, the joint density function of $\{S_{k+1}, X_{k+1}\}$ given $S_k = s, X_k = x$, and $A_k = a$ is $p(s', x'; s, x, a) = p_S(s'; s, x, a)p_X(x'; s', s, x, a)$. Let $\Lambda(p, c)$ denote a class of p -smooth functions defined in Huang (1998). Then (A4) is a smoothness condition commonly used in the nonparametric estimation to control the approximation power of sieve bases (Shi et al., 2022).

(A4) There exists some $p, c > 0$ such that $r'(\cdot, \cdot, a), p(\cdot, \cdot; s, x, a) \in \Lambda(p, c)$ with respect to $(s, x) \in \mathbb{S} \times \mathbb{R}^+$ for any $a \in \mathbb{A}$.

Following the discussion in Section 2.2, the transition trajectories of states and gap times $\{\tilde{S}_k = (S_k^\top, X_k)^\top\}_{k \geq 0}$ form a time-homogeneous Markov chain under assumptions (A1) and (A2). Suppose $\{(S_k^\top, X_k)^\top\}_{k \geq 0}$ has a unique invariant joint distribution with density function $\mu(\cdot, \cdot)$ under some behavior policy b . Similarly, when the reference distribution $\mathcal{G}$ is absolutely continuous with respect to the Lebesgue measure, there exists a density function $v_0(s, x)$ satisfying $d\mathcal{G}(s, x) = v_0(s, x)dsdx$, which is also a joint distribution for the states and gap time. Moreover, define $\bar{u}_\pi(s, x, a) = E(\mathcal{U}_{\pi, i, k+1} | S_{i, k} = s, X_{i, k} = x, A_{i, k} = a)$. The following two

assumptions follow from Lemmas A2 and A3 of Shi et al. (2022) and ensure the existence of $\widehat{\theta}_\pi$.

(A5) $\mu(\cdot, \cdot)$ and $v_0(\cdot, \cdot)$ are uniformly bounded away from 0 and ∞ .

(A6) There exists a constant $\bar{c} > 0$, such that $\lambda_{\min}(E[\sum_{k=0}^{K_i-1} \{\xi_{i,k}^{\otimes 2} - \bar{u}_\pi(S_{i,k}, X_{i,k}, A_{i,k})^{\otimes 2}\}]) \geq \bar{c}$, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix.

The following assumption is required for constructing asymptotic approximation results when $K_i \rightarrow \infty$. Specifically, the geometric ergodicity in A7(i) is part of the sufficient conditions for the Markov chain central limit theorem to hold, and A7(ii) guarantees the validity of the estimating equations for the renewal process model (3.1) when the number of trajectories is limited.

(A7) When $K_i \rightarrow \infty$, (i) the Markov chain $\{(S_{ik}^\top, X_{ik})^\top\}_{k \geq 0}$ is geometrically ergodic, and (ii) there exists a probability measure $v(\cdot)$ such that for any measurable set B_z on $\mathbb{R}^q$, $\{Z_{i,k}\}_{k \geq 0}$ satisfying $K_i^{-1} \sum_{k=0}^{K_i-1} I(Z_{i,k} \in B_z) \xrightarrow{p} v(B_z)$.

For $j = 0, 1, 2$, define $S_{z,j}(x; \beta) = N_K^{-1} \sum_{i,k} I(X_{i,k+1} \geq x) \exp(\beta^\top Z_{i,k}) Z_{i,k}^{\otimes j}$, with $a^{\otimes j} = 1, a, aa^\top$ for $j = 0, 1, 2$. The convergency of $S_{z,j}(x; \beta)$ holds under assumption (A6). So we denote the limit by $s_{z,j}(x; \beta)$ for each x and β , and denote $s_{z,j}(x; \beta_0)$ as $s_{z,j}(x)$ for notational brevity. By Slutsky's the-

orem, we can further denote $\bar{z}(x) = s_{z,1}(x)/s_{z,0}(x)$ as the limit of $\bar{Z}(x; \beta_0)$ where $\bar{Z}(x; \beta) = S_{z,1}(x; \beta)/S_{z,0}(x; \beta)$. Conditions (C1)-(C5) are regular conditions commonly adopted in the modulated renewal process studies (Pons and de Turckheim, 1988; Lin and Fine, 2009; Lin et al., 2013), and here are used to construct asymptotic properties of $\hat{\mathcal{P}}_X$ and $\hat{\lambda}$ under model (3.1) proposed in Section 3.2. Condition (C6) ensures the existence of $\sigma_{\pi, \mathcal{G}}^2$.

(C1) There exists a compact closure $\bar{\mathcal{B}}$ surrounding β_0 , such that $s_{z,0}(\beta, x)$ is bounded away from zero on $\bar{\mathcal{B}} \times [0, \tau]$.

(C2) $\Omega_z = \int_0^\tau \{s_{z,0}^{-1}(x)s_{z,2}(x) - \bar{z}^{\otimes 2}(x)\}s_{z,0}(x)d\Lambda_0(x)$ is positive definite.

(C3) $N_K^{1/2}\{ES_{z,j}(x, \beta_0) - s_{z,j}(x)\}$ is finite for $j = 0, 1, 2$ and for each x in a dense subset of $[0, \tau]$ including 0 and τ .

(C4) For any $\epsilon > 0$, $N_K^{-1} \sum_{i,k} \|Z_{i,k}\|_2^2 I\{\|Z_{i,k}\|_2^2 > N_K \epsilon\} \exp(\beta_0^\top Z_{i,k}) \Lambda_0(\tau) \xrightarrow{p} 0$.

(C5) $N_K^{1/2}\{S_{z,j}(x, \beta_0) - s_{z,j}(x, \beta_0)\}$ is tight on $[0, \tau]$ for $j = 0, 1, 2$ and $N_K \geq 1$.

(C6) Ω_π defined in (S3.10) satisfies $\lambda_{\min}(\Omega_\pi) > c_\omega$ for some constant $c_\omega > 0$.

The proof of Theorem 1 is obtained by three steps. In the first step, we give the asymptotic properties of $\hat{\mathcal{P}}_X(x; z)$ and $\hat{\lambda}$ under the modulated

renewal process model (3.1). In the second step, we use the asymptotic expansions of $\{\widehat{\mathcal{P}}_X(x; z) - \mathcal{P}_X(x; z)\}$ and $\{\widehat{\lambda}(x; z) - \lambda(x; z)\}$ to obtain the asymptotic properties of $\widehat{\theta}_\pi$. Lastly, we give the asymptotic properties of $N_K^{1/2} \sigma_{\pi, \mathcal{G}}^{-1} \{\widehat{V}_{\mathcal{I}}^\pi(\mathcal{G}) - V_{\mathcal{I}}^\pi(\mathcal{G})\}$ and finishes the proof.

Step 1. Asymptotic properties of $\widehat{\mathcal{P}}_X$ and $\widehat{\lambda}$.

Define $M_{i,k+1}(x) = \int_0^x \{dN_{i,k+1}(u) - I(X_{i,k+1} \geq u) \exp(\beta_0^\top Z_{i,k}) d\Lambda_0(u)\}$.

Then Lemma S1 gives the asymptotic properties of $\widehat{\beta}$ and $\widehat{\Lambda}_0(x)$.

Lemma S1. *Under assumption (A7) and conditions (C1)-(C5), we have:*

(i) $N_K^{1/2}(\widehat{\beta} - \beta_0)$ converges weakly to a normal distribution $\mathcal{N}(0, \Omega_z^{-1})$, and can be asymptotically written as

$$N_K^{1/2}(\widehat{\beta} - \beta_0) = N_K^{-1/2} \sum_{i,k} \Omega_z^{-1} \underbrace{\int_0^\tau \{Z_{i,k} - \bar{z}(u)\} dM_{i,k+1}(u)}_{\phi_{i,k,z}} + o_p(1).$$

(ii) $N_K^{1/2}\{\widehat{\Lambda}_0(x) - \Lambda_0(x)\}$ converges weakly to a mean zero Gaussian process, and can be asymptotically written as

$$N_K^{1/2}\{\widehat{\Lambda}_0(x) - \Lambda_0(x)\} = N_K^{-1/2} \sum_{i,k} \left[\int_0^x s_{z,0}^{-1}(u) dM_{i,k+1}(u) - \int_0^x \bar{z}(u)^\top d\Lambda_0(u) \right. \\ \left. \Omega_z^{-1} \phi_{i,k,z} \right] + o_p(1)$$

for each $x \in [0, \tau]$.

Furthermore, by Taylor's expansion, we can rewrite $N_K^{1/2}\{\widehat{\Lambda}(x; z) -$

$\Lambda(x; z)\}$ as

$$\begin{aligned}
& N_K^{1/2} \{\widehat{\Lambda}_0(x) \exp(\widehat{\beta}^\top z) - \Lambda_0(x) \exp(\beta_0^\top z)\} \\
&= N_K^{1/2} \{\widehat{\Lambda}_0(x) - \Lambda_0(x)\} \exp(\beta_0^\top z) + \Lambda_0(x) \exp(\beta_0^\top z) N_K^{1/2} (\widehat{\beta} - \beta_0)^\top z + O_p(N_K^{-1/2}) \\
&= N_K^{-1/2} \sum_{i,k} \exp(\beta_0^\top z) \left[\int_0^x s_{z,0}^{-1}(u) dM_{i,k+1}(u) + \int_0^x \{z - \bar{z}(u)\}^\top d\Lambda_0(u) \Omega_z^{-1} \phi_{i,k,z} \right] \\
&\quad + o_p(1),
\end{aligned}$$

which is also $O_p(1)$ for each fixed x and z . Recall that $\mathcal{P}_X(x; z) = 1 - \exp\{-\Lambda(x; z)\}$. Therefore, by Taylor expansion, $N_K^{1/2} \{\widehat{\mathcal{P}}_X(x; z) - \mathcal{P}_X(x; z)\}$ is asymptotically equivalent with

$$N_K^{-1/2} \sum_{i,k} \exp\{-\Lambda(x; z) + \beta_0^\top z\} \left[\int_0^x s_{z,0}^{-1}(u) dM_{i,k+1}(u) + \int_0^x \{z - \bar{z}(u)\}^\top d\Lambda_0(u) \Omega_z^{-1} \phi_{i,k,z} \right]. \tag{S3.1}$$

To show the asymptotic properties of $\widehat{\lambda}(x; z)$, we first rewrite

$$\begin{aligned}
& \widehat{\lambda}(x; z) - \lambda(x; z) \\
&= \widehat{\lambda}_0(x) \exp(\widehat{\beta}^\top z) - \lambda_0(x) \exp(\beta_0^\top z) \\
&= \{\widehat{\lambda}_0(x) - \lambda_0(x)\} \exp\{\beta_0^\top z\} + \lambda_0(x) \{\exp(\widehat{\beta}^\top z) - \exp(\beta_0^\top z)\} + r_n(x; z)
\end{aligned}$$

with $r_n(x; z) = \{\widehat{\lambda}_0(x) - \lambda_0(x)\} \{\exp(\widehat{\beta}^\top z) - \exp(\beta_0^\top z)\}$. For the first term,

we rewrite

$$\begin{aligned}
& \widehat{\lambda}_0(x) - \lambda_0(x) \\
&= \int_0^\tau \frac{1}{b_n} K\left(\frac{u-x}{b_n}\right) d\{\widehat{\Lambda}_0(u) - \Lambda_0(u)\} + \underbrace{\int_0^\tau \frac{1}{b_n} K\left(\frac{u-x}{b_n}\right) d\Lambda_0(u)}_{\lambda_0^*(x)} - \lambda_0(x) \\
&= \{\widehat{\lambda}_0(x) - \lambda_0^*(x)\} + \{\lambda_0^*(x) - \lambda_0(x)\}
\end{aligned}$$

Following Theorem 4.2.2 of Ramlau-Hansen (1983), we can obtain that $(N_K b_n)^{1/2} \{\widehat{\lambda}_0(x) - \lambda_0^*(x)\}$ converges weakly to a normal distribution for each $x \in [0, \tau]$, and therefore $|\widehat{\lambda}_0(x) - \lambda_0^*(x)| = O_p(N_K^{-1/2} b_n^{-1/2})$. For the term $\lambda_0^*(x) - \lambda_0(x)$, we give the following lemma for kernel-based estimates.

Lemma S2. *For any function $g(x)$, define $g^*(x) = \int_0^\tau b_n^{-1} K\{(u-x)/b_n\} g(u) du$.*

Then $|g^*(x) - g(x)| = O(b_n^2)$ if $g(\cdot)$ has a bounded second-order derivative in a neighborhood of x .

By Lemma S2, $|\lambda_0^*(x) - \lambda_0(x)| = O(b_n^2)$ if $\lambda_0(\cdot)$ has a bounded second derivative in a neighborhood of x . Together with Lemma S1(ii) and Lemma

S2, we have $N_K^{1/2}\{\widehat{\lambda}_0(x) - \lambda_0(x)\}$ is asymptotically equivalent with

$$\begin{aligned}
 & \int_0^\tau \frac{1}{b_n} K\left(\frac{u-x}{b_n}\right) d\left[N_K^{1/2}\{\widehat{\Lambda}_0(u) - \Lambda_0(u)\}\right] + O(N_K^{1/2}b_n^2) \\
 &= \int_0^\tau \frac{1}{b_n} K\left(\frac{u-x}{b_n}\right) N_K^{-1/2} \sum_{i,k} \left\{ s_{z,0}^{-1}(u) dM_{i,k+1}(u) - \bar{z}^\top(u) d\Lambda_0(u) \Omega_z^{-1} \phi_{i,k,z} \right\} \\
 &+ O(N_K^{1/2}b_n^2). \\
 &= N_K^{-1/2} \sum_{i,k} \left\{ \int_0^\tau \frac{1}{b_n} K\left(\frac{u-x}{b_n}\right) s_{z,0}^{-1}(u) dM_{i,k+1}(u) - \lambda_0(x) \bar{z}^\top(x) \Omega_z^{-1} \phi_{i,k,z} \right\} \\
 &+ O(N_K^{1/2}b_n^2 + b_n^2), \tag{S3.2}
 \end{aligned}$$

when $\bar{z}(\cdot)$ has a bounded second-order derivative in a neighborhood of x .

On the other hand, by Taylor expansion and Lemma S1(i), we have

$$\begin{aligned}
 & N_K^{1/2} \lambda_0(x) \{\exp(\widehat{\beta}^\top z) - \exp(\beta_0^\top z)\} \\
 &= \lambda_0(x) \exp(\beta_0^\top z) z^\top N_K^{1/2} (\widehat{\beta} - \beta_0) + O_p(\|\widehat{\beta} - \beta_0\|_2^2 N_K^{1/2}) \\
 &= \lambda(x; z) z^\top \Omega_z^{-1} N_K^{-1/2} \sum_{i,k} \phi_{i,k,z} + o_p(1), \tag{S3.3}
 \end{aligned}$$

which is $O_p(1)$ by the martingale central limit theorem.

Lastly, combining (S3.2) and (S3.3), and with condition $b_n^5 N_K \rightarrow 0$, we have $r_n = O_p(N_K^{-1} b_n^{-1/2})$, $\widehat{\lambda}(x; z) - \lambda(x; z) = O_p(N_K^{-1/2} b_n^{-1/2})$, and

$$\begin{aligned}
 & N_K^{1/2} \{\widehat{\lambda}(x; z) - \lambda(x; z)\} \\
 &= N_K^{-1/2} \sum_{i,k} \left[\exp(\beta_0^\top z) \int_0^\tau \frac{1}{b_n} K\left(\frac{u-x}{b_n}\right) s_{z,0}^{-1}(u) dM_{i,k+1}(u) \right. \\
 &\quad \left. + \lambda(x; z) \{z - \bar{z}(x)\}^\top \Omega_z^{-1} \phi_{i,k,z} \right] + O_p(N_K^{1/2}b_n^2 + N_K^{-1/2}b_n^{-1/2}). \tag{S3.4}
 \end{aligned}$$

Step 2. Asymptotic properties of $\hat{\theta}_\pi$.

Firstly, by assumption (A4), Lemma 1 of Shi et al. (2022) and Section 2.2 of Huang (1998), there exists $\theta_\pi^* \in \mathbb{R}^{mL}$ that satisfies

$$\sup_{s \in \mathbb{S}, x \in \mathbb{R}^+, a \in \mathbb{A}} |Q_{\mathcal{I}}^\pi(s, x, a) - \xi(s, x, a)^\top \theta_\pi^*| \leq C_1 L^{-p/(d+1)} \quad (\text{S3.5})$$

for some constant C_1 . Define $\hat{D}_\pi = N_K^{-1} \sum_{i,k} \xi_{i,k} (\xi_{i,k} - \hat{\mathcal{U}}_{\pi,i,k+1})^\top$. Then, by the definition of $\hat{\theta}_\pi$, we can rewrite $\hat{D}_\pi(\hat{\theta}_\pi - \theta_\pi^*)$ as

$$\begin{aligned} & \frac{1}{N_K} \sum_{i,k} \xi_{i,k} \left\{ \gamma^{X_{i,k+1}} \hat{R}_{\mathcal{I},i}(T_{i,k+1}) - (\xi_{i,k} - \hat{\mathcal{U}}_{\pi,i,k+1})^\top \theta_\pi^* \right\} \\ &= \frac{1}{N_K} \sum_{i,k} \xi_{i,k} \left\{ \gamma^{X_{i,k+1}} \hat{R}_{\mathcal{I},i}(T_{i,k+1}) - Q_{\mathcal{I}}^\pi(S_{i,k}, X_{i,k}, A_{i,k}) + \int \gamma^x Q_{\mathcal{I}}^\pi(S_{i,k+1}, x, \pi) d\hat{\mathcal{P}}_X(x; Z_{i,k}) + \epsilon_{i,k,1} \right\} \\ &= \frac{1}{N_K} \sum_{i,k} \xi_{i,k} \left\{ \underbrace{\gamma^{X_{i,k+1}} R_{\mathcal{I},i}(T_{i,k+1}) - Q_{\mathcal{I}}^\pi(S_{i,k}, X_{i,k}, A_{i,k}) + \int \gamma^x Q_{\mathcal{I}}^\pi(S_{i,k+1}, x, \pi) d\mathcal{P}_X(x; Z_{i,k})}_{\phi_{i,k,1}} + \epsilon_{i,k,1} \right\} \\ &+ \frac{1}{N_K} \sum_{i,k} \xi_{i,k} \int \gamma^x Q_{\mathcal{I}}^\pi(S_{i,k+1}, x, \pi) d\{\hat{\mathcal{P}}_X(x; Z_{i,k}) - \mathcal{P}_X(x; Z_{i,k})\} \\ &- \frac{1}{N_K} \sum_{i,k} \xi_{i,k} \gamma^{X_{i,k+1}} R_{\mathcal{I},i}(T_{i,k+1}) \left\{ \frac{\hat{\lambda}(X_{i,k+1}; Z_{i,k}) - \lambda(X_{i,k+1}; Z_{i,k})}{\hat{\lambda}(X_{i,k+1}; Z_{i,k})} \right\}, \end{aligned}$$

where

$$\epsilon_{i,k,1} = \{Q_{\mathcal{I}}^\pi(S_{i,k}, X_{i,k}, A_{i,k}) - \xi_{i,k}^\top \theta_\pi^*\} + \int \gamma^x \{\xi(S_{i,k+1}, x, \pi)^\top \theta_\pi^* - Q_{\mathcal{I}}^\pi(S_{i,k+1}, x, \pi)\} d\hat{\mathcal{P}}_X(x; Z_{i,k}).$$

For the first term, rewrite $\epsilon_{i,k,1}$ as $\epsilon_{i,k,1,1} + \epsilon_{i,k,1,2}$, where

$$\begin{aligned} \epsilon_{i,k,2,1} &= Q_{\mathcal{I}}^\pi(S_{i,k}, X_{i,k}, A_{i,k}) - \xi_{i,k}^\top \theta_\pi^* + \int \gamma^x \{\xi(S_{i,k+1}, x, \pi)^\top \theta_\pi^* - Q_{\mathcal{I}}^\pi(S_{i,k+1}, x, \pi)\} d\mathcal{P}_X(x; Z_{i,k}), \\ \epsilon_{i,k,2,2} &= \int \gamma^x \{\xi(S_{i,k+1}, x, \pi)^\top \theta_\pi^* - Q_{\mathcal{I}}^\pi(S_{i,k+1}, x, \pi)\} d\{\hat{\mathcal{P}}_X(x; Z_{i,k}) - \mathcal{P}_X(x; Z_{i,k})\}. \end{aligned}$$

By (S3.5), $\epsilon_{i,k,2,1}$ is bounded by $2C_1 L^{-p/(d+1)}$. For $\epsilon_{i,k,2,2}$, since $N_K^{1/2}\{\widehat{\Lambda}_0(x) - \Lambda_0(x)\}$ converges to a Gaussian process with zero mean and continuous sample paths (see Theorem 3.3 in Pons and de Turkheim (1988)), then, together with the convergence of $N_K^{1/2}(\widehat{\beta} - \beta_0)$, it is not hard to obtain that $N_K^{1/2}\{\widehat{\mathcal{P}}_X(x; z) - \mathcal{P}_X(x; z)\}$ also converges weakly to a stochastic process with zero mean and continuous sample paths for each fixed z . This implies that $\sup_{x \in [0, \tau]} |\widehat{\mathcal{P}}_X(x; z) - \mathcal{P}_X(x; z)| \xrightarrow{p} 0$ for each fixed z . Then, $\epsilon_{i,k,2,2}$ is also bounded by $C_1 L^{-p/(d+1)}$ when N_K is large enough, and thus $\max_{i,k} |\epsilon_{i,k,2}| \leq 3C_1 L^{-p/(d+1)}$. The following lemma gives the boundedness of $\xi_{i,k}$ and $\phi_{i,k,1}$.

Lemma S3. *Under the assumptions given in Theorem 1, we have*

(i) $E(\xi_{i,k}^\top \xi_{i,k}) = O(L)$, $\max_{1 \leq i \leq n, 0 \leq k \leq K_i - 1} |\phi_{i,k,1}| \leq C_\phi$ for some $C_\phi > 0$.

(ii) As $N_K \rightarrow \infty$, $\lambda_{\max}(N_K^{-1} \sum_{i,k} \xi_{i,k} \xi_{i,k}^\top) = O_p(1)$, $\lambda_{\min}\{N_K^{-1} \sum_{i,k} \xi_{i,k} \xi_{i,k}^\top\} \geq$

$\bar{c}_1/3$ with probability of 1, $\lambda_{\max}\{E(N_K^{-1} \sum_{i,k} \xi_{i,k} \xi_{i,k}^\top)\} = O(1)$, and

$\lambda_{\min}\{E(N_K^{-1} \sum_{i,k} \xi_{i,k} \xi_{i,k}^\top)\} \geq \bar{c}_1/2$ for some constant $\bar{c}_1 > 0$.

With arguments similar to the proof of Theorem 1 in Shi et al. (2022), we can use Markov's inequality, Cauchy-Schwarz, and Lemma S3(i) to obtain that $N_K^{-1} \sum_{i,k} \xi_{i,k} \phi_{i,k,1} = O_p(L^{1/2} N_K^{-1/2})$. Then the first term is asymp-

totically equivalent with

$$\frac{1}{N_K} \sum_{i,k} \xi_{i,k} \phi_{i,k,1} + O_p(L^{-\frac{p}{d+1}}). \quad (\text{S3.6})$$

Using the asymptotic expression given in (S3.1), the second term can be further written as

$$\begin{aligned} & \frac{1}{N_K} \sum_{j,l} \xi_{j,l} \int_0^\tau \gamma^x Q_{\mathcal{I}}^\pi(S_{j,l+1}, x, \pi) d\{\widehat{\mathcal{P}}_X(x; Z_{j,l}) - \mathcal{P}_X(x; Z_{j,l})\} \\ &= \underbrace{\int_0^\tau \frac{1}{N_K} \sum_{j,l} \xi_{j,l} \gamma^x Q_{\mathcal{I}}^\pi(S_{j,l+1}, x, \pi) \exp\{-\Lambda(x; Z_{j,l}) + \beta_0^\top Z_{j,l}\} s_{z,0}^{-1}(x) N_K^{-1} \sum_{i,k} dM_{i,k+1}(x)}_{G_{\pi,1}(x)} \\ & \quad + \underbrace{\frac{1}{N_K} \sum_{j,l} \int_0^\tau \xi_{j,l} \gamma^x Q_{\mathcal{I}}^\pi(S_{j,l+1}, x, \pi) \{Z_{j,l} - \bar{z}(x)\}^\top dP_x(x; Z_{j,l}) \Omega_z^{-1}}_{G_{\pi,3}} \frac{1}{N_K} \sum_{i,k} \phi_{i,k,z} \\ & \quad - \underbrace{\int_0^\tau \int_x \frac{1}{N_K} \sum_{j,l} \xi_{j,l} \gamma^u Q_{\mathcal{I}}^\pi(S_{j,l+1}, u, \pi) \exp\{-\Lambda(u; Z_{j,l}) + 2\beta_0^\top Z_{j,l}\} d\Lambda_0(u) s_{z,0}^{-1}(x) \frac{1}{N_K} \sum_{i,k} dM_{i,k+1}(x)}_{G_{\pi,2}(u)} \\ & \quad - \underbrace{\frac{1}{N_K} \sum_{j,l} \int_0^\tau \xi_{j,l} \gamma^x Q_{\mathcal{I}}^\pi(S_{j,l+1}, x, \pi) \exp(\beta_0^\top Z_{j,l}) \{Z_{j,l} \Lambda_0(x) - \int_0^x \bar{z}(u) d\Lambda_0(u)\}^\top dP_X(x; Z_{j,l})}_{G_{\pi,4}} \\ & \quad \times \Omega_z^{-1} \frac{1}{N_K} \sum_{i,k} \phi_{i,k,z} \\ &= \int_0^\tau \{G_{\pi,1}(x) - \int_x^\tau G_{\pi,2}(u) d\Lambda_0(u)\} s_{z,0}^{-1}(x) \frac{1}{N_K} \sum_{i,k} dM_{i,k+1}(x) + (G_{\pi,3} - G_{\pi,4}) \Omega_z^{-1} \frac{1}{N_K} \sum_{i,k} \phi_{i,k,z} \\ &= \frac{1}{N_K} \sum_{i,k} \underbrace{\left[\int_0^\tau \{g_{\pi,1}(x) - \int_x^\tau g_{\pi,2}(u) d\Lambda_0(u)\} s_{z,0}^{-1}(x) dM_{i,k+1}(x) + (g_{\pi,3} - g_{\pi,4}) \Omega_z^{-1} \phi_{i,k,z} + \epsilon_{i,k,2} \right]}_{\varphi_{i,k,2}}, \end{aligned}$$

where $g_{\pi,k} = E(G_{\pi,k})$ for $k = 1, 2, 3, 4$, and

$$\begin{aligned} \epsilon_{i,k,2} &= \int_0^\tau [\{G_{\pi,1}(x) - g_{\pi,1}(x)\} - \int_x^\tau \{G_{\pi,2}(u) - g_{\pi,2}(u)\} d\Lambda_0(u)] s_{z,0}^{-1}(x) \\ &\quad dM_{i,k+1}(x) + \{(G_{\pi,3} - g_{\pi,3}) - (G_{\pi,4} - g_{\pi,4})\} \Omega_z^{-1} \phi_{i,k,z}. \end{aligned}$$

By Lemma S2(ii), for any sequence $\{u_{i,k}\}_{1 \leq i \leq n, 0 \leq k \leq K_i-1}$ that satisfies $\max_{i,k} |u_{i,k}| \leq C_u$ for some positive constant C_u , we have $\|N_K^{-1} \sum_{i,k} \xi_{i,k} u_{i,k}\|_2 = O_p(1)$ and $\|E(N_K^{-1} \sum_{i,k} \xi_{i,k} u_{i,k})\|_2 = O(1)$. Thus, by the boundedness of $Q_{\mathcal{I}}^\pi(\cdot, \cdot, \cdot)$ (see the proof of Lemma S3(i)), $g_{\pi,k} = O(1)$, $G_{\pi,k} = O_p(1)$ for all k . Then, similar with the arguments for the first term, we can obtain that $N_K^{-1} \sum_{i,k} \varphi_{i,k,2} = O_p(L^{1/2} N_K^{-1/2})$. Furthermore, by assumption (A7), $|\sup_x \{G_{\pi,k}(x) - g_{\pi,k}(x)\}| = o_p(1)$ for $k = 1, 2$, and $|G_{\pi,k} - g_{\pi,k}| = o_p(1)$ for $k = 3, 4$. Then, $N_K^{-1} \sum_{i,k} \epsilon_{i,k,2} = o_p(L^{1/2} N_K^{-1/2})$, and the second term is asymptotically equivalent with

$$\frac{1}{N_K} \sum_{i,k} \varphi_{i,k,2} + o_p(L^{1/2} N_K^{-1/2}) \quad (\text{S3.7})$$

For the third term, by the asymptotic properties of $\widehat{\lambda}$ and Lemma S3(ii),

we have

$$\begin{aligned}
& N_K^{-1} \sum_{i,k} \xi_{i,k} \gamma^{X_{i,k+1}} R_{\mathcal{I},i}(T_{i,k+1}) \left\{ \frac{\widehat{\lambda}(X_{i,k+1}; Z_{i,k}) - \lambda(X_{i,k+1}; Z_{i,k})}{\lambda(X_{i,k+1}; Z_{i,k})} + O_p\left(\frac{1}{N_K b_n}\right) \right\} \\
&= \underbrace{\int_0^\tau \frac{1}{N_K} \sum_{i,k} \xi_{i,k} \gamma^{X_{i,k+1}} R_{\mathcal{I},i}(T_{i,k+1}) \{\lambda_0(X_{i,k+1})\}^{-1} \frac{1}{b_n} K\left(\frac{u - X_{i,k+1}}{b_n}\right) s_{z,0}^{-1}(u)}_{G_5(u)} \\
& \quad + \underbrace{N_K^{-1} \sum_{j,l} dM_{j,l+1}(u) + \frac{1}{N_K} \sum_{i,k} \xi_{i,k} \gamma^{X_{i,k+1}} R_{\mathcal{I},i}(T_{i,k+1}) \{Z_{i,k} - \bar{z}(X_{i,k+1})\}^\top}_{G_6} \\
& \quad \Omega_z^{-1} N_K^{-1} \sum_{j,l} \phi_{j,l,z} + O_p(b_n^2 + \frac{1}{N_K b_n}).
\end{aligned}$$

Specifically, when $r(s, x, a, s', \cdot)$ and $\lambda_0(\cdot)$ have bounded second-order derivatives in a neighborhood of u , $E\{G_5(u)\}$ can be rewritten as

$$\begin{aligned}
& E \left\{ \frac{1}{N_K} \sum_{i,k} \xi_{i,k} \gamma^{X_{i,k+1}} R_{\mathcal{I},i}(T_{i,k+1}) \{\lambda_0(X_{i,k+1})\}^{-1} \frac{1}{b_n} K\left(\frac{u - X_{i,k+1}}{b_n}\right) \right\} \\
&= E \left\{ \frac{1}{N_K} \sum_{i,k} \xi_{i,k} \int_0^\infty \frac{1}{b_n} K\left(\frac{x - u}{b_n}\right) \frac{\gamma^x r(S_k, X_k, A_k, S_{k+1}, x)}{\lambda(x; Z_{ik}) \lambda_0(x)} d\mathcal{P}_X(x; Z_{i,k}) \right\} \\
&= E \left[\frac{1}{N_K} \sum_{i,k} \xi_{i,k} \int_0^\infty \frac{1}{b_n} K\left(\frac{x - u}{b_n}\right) \gamma^x r(S_k, X_k, A_k, S_{k+1}, x) \{\lambda_0(x)\}^{-1} \exp\{ \right. \\
& \quad \left. - \Lambda(x; Z_{ik}) \} dx \right] \\
&= \int_0^\infty \frac{1}{b_n} K\left(\frac{x - u}{b_n}\right) E \left[\frac{1}{N_K} \sum_{i,k} \xi_{i,k} \gamma^x r(S_k, X_k, A_k, S_{k+1}, x) \{\lambda_0(x)\}^{-1} \exp\{ \right. \\
& \quad \left. - \Lambda(x; Z_{ik}) \} \right] dx \\
&= E \left[\underbrace{\frac{1}{N_K} \sum_{i,k} \xi_{i,k} \gamma^u r(S_k, X_k, A_k, S_{k+1}, u) \{\lambda_0(u)\}^{-1} \exp\{-\Lambda(u; Z_{ik})\}}_{g_5(u)} \right] + O(L^{1/2} b_n^2).
\end{aligned}$$

Here, the last equation is obtained by Lemma S2, Lemma S3(ii), and the boundedness property:

$$\max_{i,k} |\gamma^x r(S_k, X_k, A_k, S_{k+1}, x) \{\lambda_0(x)\}^{-1} \exp\{-\Lambda(x; Z_{ik})\}| \leq C_r \exp(-C_\Lambda)/C_{\lambda_0},$$

where C_r denote the maximum bound of $r(s, x, a, s', x')$ given in Assumption (A3), $0 < C_{\lambda_0} < 1$ be a constant such that $C_{\lambda_0} \leq \inf_x \lambda_0(x) \leq \sup_x \lambda_0(x) \leq C_{\lambda_0}^{-1}$, and $0 < C_\Lambda < 1$ be a constant such that $C_\Lambda \leq \inf_x \Lambda(x; Z_{ik}) \leq \sup_x \Lambda(x; Z_{ik}) \leq C_\Lambda^{-1}$ wpa 1. Moreover, define $g_6 = E(G_6)$ and rewrite the third term in the expansion of $\widehat{D}_\pi(\widehat{\theta}_\pi - \theta_\pi^*)$ as

$$N_K^{-1} \sum_{i,k} \left\{ \underbrace{\int_0^\tau g_5(u) s_{z,0}^{-1}(u) dM_{i,k+1}(u) + g_6 \Omega_z^{-1} \phi_{i,k,z} + \epsilon_{i,k,3}}_{\varphi_{i,k,3}} \right\} + O_p(b_n^2 + \frac{1}{N_K b_n}),$$

where $\epsilon_{i,k,3} = \int_0^\tau \{G_5(u) - g_5(u)\} s_{z,0}^{-1}(u) dM_{i,k+1}(u) + (G_6 - g_6) \Omega_z^{-1} \phi_{i,k,z}$. Here,

similar with $\varphi_{i,k,1}$ and $\varphi_{i,k,2}$, we also have $N_K^{-1} \sum_{i,k} \varphi_{i,k,3} = O_p(L^{1/2} N_K^{-1/2})$.

For the residuals $\epsilon_{i,k,3}$, note that G_5 is a kernel estimator with $\|G_5 - g_5\|_2 = O_p(L^{1/2} N_K^{-1/2} b_n^{-1/2} + L^{1/2} b_n^2)$. Then with arguments similar to $G_{\pi,k}$, we

have $N_K^{-1} \sum_{i,k} \epsilon_{i,k,3} = O_p(L^{1/2} N_K^{-1} b_n^{-1/2} + L^{1/2} N_K^{-1/2} b_n^2) + o_p(L^{1/2} N_K^{-1/2}) = o_p(L^{1/2} N_K^{-1/2})$ when $N_K b_n \rightarrow \infty$ and $b_n \rightarrow 0$ as $N_K \rightarrow \infty$. Therefore, the

third term is asymptotically equivalent with

$$N_K^{-1} \sum_{i,k} \varphi_{i,k,3} + O_p(b_n^2 + N_K^{-1} b_n^{-1}) + o_p(L^{1/2} N_K^{-1/2}) \quad (\text{S3.8})$$

Combining three asymptotic expansions (S3.6)-(S3.8), $\widehat{\theta}_\pi - \theta_\pi^*$ is asymp-

totically equivalent with

$$\begin{aligned} & \widehat{D}_\pi^{-1} \left[\frac{1}{N_K} \sum_{i,k} (\xi_{i,k} \phi_{i,k,1} + \varphi_{i,k,2} - \varphi_{i,k,3}) + O_p(L^{-\frac{p}{d+1}} + b_n^2 + b_n^{-1} N_K^{-1}) \right], \\ & = \widehat{D}_\pi^{-1} \left[A_\xi^\top \phi_1 + A_I^\top (\varphi_2 - \varphi_3) + O_p(L^{-\frac{p}{d+1}} + b_n^2 + N_K^{-1} b_n^{-1}) \right], \end{aligned}$$

where $\phi_1 = \text{array}(\{\phi_{i,k,1}\}_{1 \leq i \leq n, 0 \leq k \leq K_i-1})$, $\varphi_2 = \text{array}(\{\varphi_{i,k,2}\}_{1 \leq i \leq n, 0 \leq k \leq K_i-1})$, $\varphi_3 = \text{array}(\{\varphi_{i,k,3}\}_{1 \leq i \leq n, 0 \leq k \leq K_i-1})$, $A_\xi = \text{array}(\{\xi_{i,k}\}_{1 \leq i \leq n, 0 \leq k \leq K_i-1}) \in \mathbb{R}^{N_K \times mL}$ and $A_I = \text{array}(\{1\}_{1 \leq i \leq n, 0 \leq k \leq K_i-1})$. The boundedness of $A_\xi^\top \phi_1$ and $A_I^\top (\varphi_2 - \varphi_3)$ directly follows from Lemma S3 and the boundedness of $\Lambda(x; z)$. To show the boundedness of $\widehat{D}_\pi^{-1}$, we give the following Lemma.

Lemma S4. *Under the assumptions given in Theorem 1, we have:*

- (i) $\|D_\pi\|_2 = O(1)$, $\|D_\pi^{-1}\|_2 \leq 3\bar{c}^{-1}$.
- (ii) As $N_K \rightarrow \infty$, $\|\widehat{D}_\pi - D_\pi\|_2 = O_p\{L^{1/2} N_K^{-1/2} \log(N_K)\}$, $\|\widehat{D}_\pi^{-1}\|_2 \leq 6\bar{c}^{-1}$, and $\|\widehat{D}_\pi^{-1} - D_\pi^{-1}\|_2 = O_p\{L^{1/2} N_K^{-1/2} \log(N_K)\}$, wpa 1.

By Cauchy-Schwarz inequality, we can obtain from Lemma S4 that

$$\|D_\pi^{-1}\{A_\xi^\top \phi_1 + A_I^\top (\varphi_2 - \varphi_3)\}\|_2 = O_p(L^{1/2} N_K^{-1/2}), \quad \|(\widehat{D}_\pi^{-1} - D_\pi^{-1})\{A_\xi^\top \phi_1 +$$

$A_I^\top(\varphi_2 - \varphi_3)\}_{\|_2} = O_p\{LN_K^{-1}\log(N_K)\}$, and thus,

$$\begin{aligned}
\widehat{\theta}_\pi - \theta_\pi^* &= D_\pi^{-1}\{A_\xi^\top\phi_1 + A_I^\top(\varphi_2 - \varphi_3)\} + (\widehat{D}_\pi^{-1} - D_\pi^{-1})\{A_\xi^\top\phi_1 + A_I^\top(\varphi_2 - \varphi_3)\} \\
&\quad + O_p(L^{-\frac{p}{d+1}} + b_n^2 + b_n^{-1}N_K^{-1}) \\
&= D_\pi^{-1}\{A_\xi^\top\phi_1 + A_I^\top(\varphi_2 - \varphi_3)\} + O_p\{LN_K^{-1}\log(N_K)\} + O_p(L^{-\frac{p}{d+1}} + b_n^2 \\
&\quad + b_n^{-1}N_K^{-1}). \tag{S3.9}
\end{aligned}$$

Step 3. Asymptotic properties of $\widehat{V}_T^\pi(\mathcal{G})$.

To study the asymptotic properties of $\widehat{V}_T^\pi(\mathcal{G})$, we first define $\sigma_{\pi,\mathcal{G}}^2 = \zeta_{\pi,\mathcal{G}}^\top D_\pi^{-1} \Omega_\pi (D_\pi^\top)^{-1} \zeta_{\pi,\mathcal{G}}$, with $\zeta_{\pi,\mathcal{G}} = \sum_{a \in \mathbb{A}} \int \pi(a|s, x) \xi(s, x, a) d\mathcal{G}(s, x)$ and

$$\Omega_\pi = E\left[\{A_\xi^\top\phi_1 + A_I^\top(\varphi_2 - \varphi_3)\}\{A_\xi^\top\phi_1 + A_I^\top(\varphi_2 - \varphi_3)\}^\top\right]. \tag{S3.10}$$

By Lemma S3(ii) and the boundedness of $\phi_{i,k,1}$, $\varphi_{i,k,2}$ and $\varphi_{i,k,3}$, it can be obtained that $\|\Omega_\pi\|_2 = O(1)$. On the other hand, by condition (C6) and Lemma S4(i), Ω_π and D_π^{-1} satisfy $\lambda_{\min}(\Omega_\pi) \geq c_\omega$ and $\lambda_{\min}\{D_\pi^{-1}(D_\pi^\top)^{-1}\} \geq c_D$ for some constant $c_\omega > 0$ and $c_D > 0$ respectively. Therefore, $\sigma_{\pi,\mathcal{G}}^2 \geq c_\omega c_D \|\zeta_{\pi,\mathcal{G}}\|_2^2$.

Recall that $\widehat{V}_T^\pi(\mathcal{G}) = \zeta_{\pi,\mathcal{G}}^\top \widehat{\theta}_\pi$. Then together with (S3.5) and (S3.9),

$\sigma_{\pi, \mathcal{G}}^{-1} \{\widehat{V}_{\mathcal{I}}^{\pi}(\mathcal{G}) - V_{\mathcal{I}}^{\pi}(\mathcal{G})\}$ is asymptotically equivalent with

$$\begin{aligned}
& \sigma_{\pi, \mathcal{G}}^{-1} \left[\zeta_{\pi, \mathcal{G}}^{\top} (\widehat{\theta}_{\pi} - \theta_{\pi}^*) - \int \{Q_{\mathcal{I}}^{\pi}(s, x, \pi) - \xi(s, x, \pi)^{\top} \theta_{\pi}^*\} d\mathcal{G}(s, x) \right] \\
&= \sigma_{\pi, \mathcal{G}}^{-1} \zeta_{\pi, \mathcal{G}}^{\top} \left[D_{\pi}^{-1} \{A_{\xi}^{\top} \phi_1 + A_I^{\top} (\varphi_2 - \varphi_3)\} + O_p\{LN_K^{-1} \log(N_K) + L^{-\frac{p}{d+1}} + b_n^2 + \right. \\
&\quad \left. b_n^{-1} N_K^{-1}\} \right] + O_p(L^{-\frac{p}{d+1}} \|\zeta_{\pi, \mathcal{G}}\|_2^{-1}) \\
&= \sigma_{\pi, \mathcal{G}}^{-1} \zeta_{\pi, \mathcal{G}}^{\top} D_{\pi}^{-1} \{A_{\xi}^{\top} \phi_1 + A_I^{\top} (\varphi_2 - \varphi_3)\} + O_p\{LN_K^{-1} \log(N_K) + L^{-\frac{p}{d+1}} (1 + \\
&\quad \|\zeta_{\pi, \mathcal{G}}\|_2^{-1}) + b_n^2 + b_n^{-1} N_K^{-1}\} \\
&= \sigma_{\pi, \mathcal{G}}^{-1} \zeta_{\pi, \mathcal{G}}^{\top} D_{\pi}^{-1} \{A_{\xi}^{\top} \phi_1 + A_I^{\top} (\varphi_2 - \varphi_3)\} + o_p(N_K^{-1/2}),
\end{aligned}$$

when $L \ll N_K^{1/2}/\log(N_K)$, $L^{\frac{2p}{d+1}} \gg N_K(1 + \|\zeta_{\pi, \mathcal{G}}\|_2^{-1})^2$, $N_K b_n^4 \rightarrow 0$ and $N_K b_n^2 \rightarrow \infty$. Lastly, using the martingale central limit theorem (McLeish, 1974, corollary 2.8) and following arguments in section E.5 of Shi et al. (2022), we have $N_K^{1/2} \sigma_{\pi, \mathcal{G}}^{-1} \{\widehat{V}_{\mathcal{I}}^{\pi}(\mathcal{G}) - V_{\mathcal{I}}^{\pi}(\mathcal{G})\} \xrightarrow{d} N(0, 1)$. ■

S3.4 Proofs of Lemmas S1–S4

Here we provide proofs of Lemmas S1–S4 given in S3.3.

Proof of Lemma S1. We first introduce some notation. For $1 \leq i \leq n$, $0 \leq k \leq K_i - 1$, denote $\{X_{i,k+1}, Z_{i,k}, N_{i,k+1}\}$ as $\{X_{(g)}, Z_{(g)}, N_{(g)}\}$ for $g = \sum_{l < i} K_l + k + 1$. Then for $j = 0, 1, 2$, $S_{z,j}(x; \beta)$ can be rewritten as $S_{z,j}(x; \beta) = N_K^{-1} \sum_{g=1}^{N_K} I(X_g \geq x) \exp(\beta^{\top} Z_{(g)}) Z_{(g)}^{\otimes j}$, and $\widehat{\beta}$ is the solution to $U_z(\beta) = N_K^{-1} \sum_{g=1}^{N_K} \int_0^{\tau} \{Z_{(g)} - \overline{Z}(x; \beta)\} dN_{(g)}(x) = 0$. Under assumption

(A7), conditions (C1)-(C5), and by Theorem 3.2 of Pons and de Turckheim (1988), $N_K^{1/2}(\widehat{\beta} - \beta_0)$ converges weakly to a Gaussian variable with mean 0 and variance Ω_z^{-1} , where Ω_z is defined in condition (C2). Moreover, $N_K^{1/2}(\widehat{\beta} - \beta_0)$ can be asymptotically written as

$$N_K^{-1/2} \sum_{g=1}^{N_K} \Omega_z^{-1} \int_0^\tau \{Z_{(g)} - \bar{z}(x)\} dM_{(g)}(x) + o_p(1),$$

with $M_{(g)}(x) = \int_0^x \{dN_{(g)}(u) - I(X_{(g)} \geq u) \exp(\beta_0^\top Z_{(g)}) d\Lambda_0(u)\}$. Rewrite $M_{(g)}(x)$ as $M_{i,k+1}(x)$, and define $\phi_{i,k,z} = \int_0^\tau \{Z_{i,k} - \bar{z}(u)\} dM_{i,k+1}(u)$, then Lemma S1(i) follows directly.

Now we derive the asymptotic properties of $\widehat{\Lambda}_0(x)$. The convergence of $N_K^{1/2}\{\widehat{\Lambda}_0(x) - \Lambda_0(x)\}$ follows from Theorem 3.3 of Pons and de Turckheim (1988). To obtain the asymptotic expression in Lemma S1(ii), we first rewrite $N_K^{1/2}\{\widehat{\Lambda}_0(x) - \Lambda_0(x)\}$ as

$$\begin{aligned} & N_K^{1/2}\{\widehat{\Lambda}_0(x; \widehat{\beta}) - \widehat{\Lambda}_0(x; \beta_0)\} + N_K^{1/2}\{\widehat{\Lambda}_0(x; \beta_0) - \Lambda_0(x)\} \\ &= - \int_0^x z(u)^\top d\Lambda_0(u) N_K^{1/2}(\widehat{\beta} - \beta_0) + N_K^{1/2}\{\widehat{\Lambda}_0(x; \beta_0) - \Lambda_0(x)\} + o_p(1) \end{aligned}$$

where $N_K^{1/2}\{\widehat{\Lambda}_0(x; \beta_0) - \Lambda_0(x)\} = N_K^{-1} \sum_{g=1}^{N_K} \int_0^x s_{z,0}(u)^{-1} dM_{(g)}(u) + o_p(1)$.

Then together with Lemma S1(i), we have $N_K^{1/2}\{\widehat{\Lambda}_0(x) - \Lambda_0(x)\}$ is asymptotically equivalent with

$$N_K^{-1/2} \sum_{i,k} \left[\int_0^x s_{z,0}(u)^{-1} dM_{i,k+1}(u) - \int_0^x \bar{z}(u)^\top d\Lambda_0(u) \Omega_z^{-1} \phi_{i,k,z} \right] + o_p(1). \quad \blacksquare$$

Proof of Lemma S2. Similarly with Ramlau-Hansen (1983), we assume the kernel has support within $[-1, 1]$ and $0 \leq b_n \leq 1/2$ to simplify mathematics. Then by Taylor expansion,

$$\begin{aligned}
g^*(x) &= \int_0^\tau \frac{1}{b_n} K\left(\frac{u-x}{b_n}\right) g(u) du \\
&= \int_{-1}^1 K(t) g(x + tb_n) dt \\
&= \int_{-1}^1 K(t) \{g(x) + g'(x)tb_n + g''(x + v_tb_n)t^2b_n^2/2\} dt \\
&= g(x) + b_n^2/2 \int_{-1}^1 K(t) g''(x + v_tb_n)t^2 dt,
\end{aligned}$$

for $0 \leq v_t \leq t$. Therefore, $|g^*(x) - g(x)| \leq \int_{-1}^1 K(t)t^2 dt \sup_{|u-x| \leq 1/2} |g''(u)|b_n^2/2$ is $O(b_n^2)$ when $g(\cdot)$ has a bounded second-order derivative in the neighborhood of x . ■

Proof of Lemma S3. For the properties of $\xi_{i,k}$, by Lemma E.1 in Shi et al. (2022), there exists some constant c^* such that $\sup_{s,x} \|\Phi_L(s, x)\|_2 \leq c^*L$. Recall that $\xi_{i,k} = \Phi(S_{i,k}, X_{i,k}, A_{i,k})$ and $|\mathbb{A}| = m$, it can be seen that $E(\xi_{i,k}^\top \xi_{i,k}) \leq m \sup_{s,x} \|\Phi_L(s, x)\|_2^2 \leq c^*mL$ is $O(L)$. The other results follows directly from the proof of Lemma E.3 in Shi et al. (2022).

For the boundedness of $\phi_{i,k,1}$, by assumption (A3) and Lemma 1 in Shi et al. (2022), $Q_{\mathcal{I}}^\pi(\cdot, \cdot, a)$ is a p -smooth function that has bounded derivatives up to order $\lfloor p \rfloor$ with respect to $(s, x) \in \mathbb{S} \times \mathbb{R}^+$. Therefore, we have $\max_{s \in \mathbb{S}, x \in \mathbb{R}^+} |Q_{\mathcal{I}}^\pi(s, x, a)| \leq C_Q$ for some constant C_Q and for all $a \in \mathbb{A}$.

Then, the boundedness result $\max_{1 \leq i \leq n, 0 \leq k \leq K_i-1} |\phi_{i,k,1}| \leq c_R + 2C_Q$ follows from the condition $P(\max_k |R(T_k)| \leq c_R) = 1$ given in Theorem 1. ■

Proof of Lemma S4. To prove Lemma S4, we first rewrite

$$\begin{aligned} \widehat{D}_\pi &= N_K^{-1} \sum_{i,k} \xi_{i,k} (\xi_{i,k} - \widehat{\mathcal{U}}_{\pi,i,k+1})^\top \\ &= \underbrace{N_K^{-1} \sum_{i,k} \xi_{i,k} (\xi_{i,k} - \mathcal{U}_{\pi,i,k+1})^\top}_{\widehat{D}_{\pi,1}} + \underbrace{N_K^{-1} \sum_{i,k} \xi_{i,k} (\mathcal{U}_{\pi,i,k+1} - \widehat{\mathcal{U}}_{\pi,i,k+1})^\top}_{\widehat{D}_{\pi,2}}. \end{aligned}$$

Note that $E(\widehat{D}_{\pi,1}) = D_\pi$, then similar with Lemma E.2 in Shi et al. (2022), we have $\|D_\pi\|_2 = O(1)$, $\|D_\pi^{-1}\|_2 \leq 3\bar{c}^{-1}$, $\|\widehat{D}_{\pi,1}\| = O(1)$, and $\|\widehat{D}_{\pi,1} - D_\pi\| = O_p\{L^{1/2}N_K^{-1/2}\log(N_K)\}$.

On the other hand, rewrite $\widehat{D}_{\pi,2}$ as

$$N_K^{-1} \sum_{i,k} \xi_{i,k} \int_0^\tau \gamma^x \xi(S_{i,k+1}, x, \pi)^\top d\{\widehat{\mathcal{P}}_X(x; Z_{i,k}) - \mathcal{P}_X(x; Z_{i,k})\}.$$

Then following the asymptotic properties of $\{\widehat{\mathcal{P}}_X(x; Z_{i,k}) - \mathcal{P}_X(x; Z_{i,k})\}$ given in (S3.1), and using arguments similar with the proof of (S3.7), we can obtain $\|\widehat{D}_{\pi,2}\|_2 = O_p(N_K^{-1/2})$ and therefore, $\|\widehat{D}_\pi - D_\pi\|_2 \leq \|\widehat{D}_{\pi,1} - D_\pi\|_2 + \|\widehat{D}_{\pi,2}\|_2 = O_p\{L^{1/2}N_K^{-1/2}\log(N_K)\}$. When $L = o\{N_K^{1/2}/\log(N_K)\}$, $\|\widehat{D}_\pi - D_\pi\|_2 = o_p(1)$, and thus $\|\widehat{D}_\pi - D_\pi\|_2 \leq \bar{c}/6$ wpa 1. Together with $\|D_\pi^{-1}\|_2 \leq 3\bar{c}^{-1}$, we have

$$v^\top \widehat{D}_\pi v = v^\top D_\pi v + v^\top (\widehat{D}_\pi - D_\pi) v \geq \frac{\bar{c}}{3} \|v\|^2 - \frac{\bar{c}}{6} \|v\|^2 = \frac{\bar{c}}{6} \|v\|^2,$$

for any $v \in \mathbb{R}^{mL}$. This implies $\|\widehat{D}_\pi^{-1}\|_2 \leq 6\bar{c}^{-1}$ wpa 1. Therefore,

$$\|\widehat{D}_\pi^{-1} - D_\pi^{-1}\|_2 \leq \|\widehat{D}_\pi^{-1}\|_2 \|\widehat{D}_\pi - D_\pi\|_2 \|D_\pi^{-1}\|_2 \leq 18\bar{c}^{-2} \|\widehat{D}_\pi - D_\pi\|_2,$$

which is also $O_p\{L^{1/2}N_K^{-1/2}\log(N_K)\}$. ■

Bibliography

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models. *The Annals of Statistics* 26, 242–272.

Lin, F. and J. P. Fine (2009). Pseudomartingale estimating equations for modulated renewal process models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 71, 3–23.

Lin, F.-C., Y. K. Truong, and J. P. Fine (2013). Robust analysis of semi-parametric renewal process models. *Biometrika* 100(3), 709–726.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *The Annals of Probability* 2(4), 620–628.

Pons, O. and E. de Turckheim (1988). Cox’s periodic regression model. *The Annals of Statistics* 16, 678–693.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics* 11, 453–466.

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.