Statistical inference of the value function for reinforcement learning in infinite-horizon settings

Chengchun Shi¹ | Sheng Zhang² | Wenbin Lu² | Rui Song²

¹London School of Economics and Political Science, London, UK
²North Carolina State University, Raleigh, North Carolina, USA

Abstract

Reinforcement learning is a general technique that allows an agent to learn an optimal policy and interact with an environment in sequential decision-making problems. The goodness of a policy is measured by its value function starting from some initial state. The focus of this paper was to construct confidence intervals (CIs) for a policy’s value in infinite horizon settings where the number of decision points diverges to infinity. We propose to model the action-value state function (Q-function) associated with a policy based on series/sieve method to derive its confidence interval. When the target policy depends on the observed data as well, we propose a ScequentiAl Value Estimation (SAVE) method to recursively update the estimated policy and its value estimator. As long as either the number of trajectories or the number of decision points diverges to infinity, we show that the proposed CI achieves nominal coverage even in cases where the optimal policy is not unique. Simulation studies are conducted to back up our theoretical findings. We apply the proposed method to a dataset from mobile health studies and find that reinforcement learning algorithms could help improve patient’s health status. A Python implementation of the proposed procedure is available at https://github.com/shengzhang37/SAVE.
1 | INTRODUCTION

Reinforcement learning (RL) is a general technique that allows an agent to learn and interact with an environment. A policy defines the agent’s way of behaving. It maps the states of environments to a set of actions to be chosen from. RL algorithms have made tremendous achievements and found extensive applications in video games (Silver et al., 2016), robotics (Kormushev et al., 2013), bidding (Jin et al., 2018), ridesharing (Xu et al., 2018), etc. In particular, a number of RL methods have been proposed in precision medicine, to derive an optimal policy as a set of sequential treatment decision rules that optimize patients’ clinical outcomes over a fixed period of time (finite horizon). References include Murphy (2003); Zhang et al. (2013); Zhao et al. (2015); Shi et al. (2018a, b) and Zhang et al. (2018), to name a few.

Mobile health (or mHealth) technology has recently emerged due to the use of mobile devices such as mobile phones, tablet computers or wearable devices in health care. It allows health-care providers to communicate with patients and manage their illness in real time. It also collects rich longitudinal data (e.g. through mobile health apps) that can be used to estimate the optimal policy. Data from mHealth applications differ from those in finite horizon settings in that the number of treatment decision points for each patient is not necessarily fixed (infinite horizon) while the total number of patients could be limited. Take the OhioT1DM dataset (Marling & Bunescu, 2018) as an example. It contains data for six patients with type 1 diabetes. For all patients, their continuous glucose monitoring (CGM) blood glucose levels, insulin doses including bolus and basal rates, self-reported times of meals and exercises are continually measured and recorded for eight weeks. Developing an optimal policy as functions of these time-varying covariates could potentially assist these patients in improving their health status.

In this paper, we focus on the infinite horizon setting where the data generating process is modeled by a Markov decision process (MDP, Puterman, 1994). Specifically, at each time point, the agent selects an action based on the observed state. The system responds by giving the decision maker a corresponding outcome and moving into a new state in the next time step. This model is generally applicable to sequential decision making, including applications from mHealth, games, robotics, ridesharing, etc. After a policy is being proposed, it is important to examine its benefit prior to recommending it for practical use. The goodness of a policy is quantified by its (state) value function, corresponding to the discounted cumulative reward that the agent receives on average, starting from some initial state. The inference of the value function helps a decision maker to evaluate the impact of implementing a policy when the environment is in a certain state. In some applications, it is also important to evaluate the integrated value of a policy aggregated over different initial states. For example, in medical studies, one might wish to know the mean outcome of patients in the population. The integrated value could thus be used as a criterion for comparing different policies.

In statistics literature, a few methods have been proposed to estimate the optimal policy in infinite horizons. Ertefaie and Strawderman (2018) proposed a variant of the gradient Q-learning method. Luckett et al. (2019) proposed a V-learning to directly search the optimal policy among a restricted class of policies. Inference of the value function under a generic (data-dependent) policy has not been studied in these papers. In the computer science literature, Thomas et al.
SHI et al. (2015) and Jiang and Li (2016) proposed (augmented) inverse propensity-score weighted ((A) IPW) estimators for the value function in infinite horizons and derived their associated CIs. However, these methods are not suitable for settings where only a limited number of trajectories (e.g., plays of a game or patients in medical studies) are available, since (A)IPW estimators become increasingly unstable as the number of decision points diverges to infinity. Recently, Kallus and Uehara (2019) proposed a double reinforcement learning (DRL) method that achieves consistent estimation of the value under a fixed policy even with a limited number of trajectories. Their method computes a Q-function and a marginalized density ratio. Learning the density ratio is challenging in general and it remains difficult to investigate the goodness-of-fit of the estimated density ratio in practice.

The focus of this paper was to construct confidence intervals (CIs) for a (possibly data-dependent) policy’s value function at a given state as well as its integrated value with respect to a given reference distribution. Our proposed CI is derived by estimating the state-action value function (Q-function) under the target policy. Similar to the value, the Q-function measures the discounted cumulative reward that the agent receives on average, starting from some initial state-action pair. We use series/sieve method to approximate the Q-function based on $L$ basis functions, where $L$ grows with the total number of observations. The advances of our proposed method are summarized as follows. First, the proposed inference method is generally applicable. Specifically, it can be applied to any fixed policy (either deterministic or random) and any data-dependent policy whose value converges at a certain rate. The latter includes policies estimated by general Q-learning type algorithms that learns an optimal Q-function from the observed data, such as gradient Q-learning (Ertefaie & Strawderman, 2018; Maei et al., 2010), fitted Q-iteration (see e.g. Ernst et al., 2005; Riedmiller, 2005), etc. See Section 3.2.4 for detailed illustrations.

Second, when applied to data-dependent policies, our method is valid in nonregular cases where the optimal policy is not uniquely defined. Inference without requiring the uniqueness of the optimal policy is extremely challenging even in the simpler finite-horizon settings (see the related discussions in Luedtke & van der Laan, 2016). The major challenge lies in that the estimated policy may not stabilize as sample size grows, making the variance of the value estimator difficult to estimate (see Section 3.2.1 for details). We achieve valid inference by proposing a Sequential Value Evaluation (SAVE) method that splits the data into several blocks and recursively update the estimated policy and its value estimator. It is worth mentioning that the data-splitting rule cannot be arbitrarily determined since the observations are time-dependent in infinite horizon settings (see Section 3.2.2 for details).

Third, our CI is valid as long as either the number of trajectories $n$ in the data, or the number of decision points $T$ per trajectory diverges to infinity. It can thus be applied to a wide variety of real applications in infinite horizons ranging from the Framingham heart study (Tsao & Vasan, 2015) with over two thousand patients to the OhioT1DM dataset that contains eight weeks’ worth of data for six people. We also allow both $n$ and $T$ to approach infinity, which is the case in applications from video games. In contrast, CIs proposed by Thomas et al. (2015) and Jiang and Li (2016) require $n$ to grow to infinity to achieve nominal coverage.

Lastly, we consider both off-policy and on-policy learning methods. In off-policy settings, CIs are derived based on historical data collected by a potentially different behavior policy. Off-policy evaluation is critical in situations where running the target policy could be expensive, risky or unethical. In on-policy settings, the estimated policy is recursively updated as batches of new observations arrive. To the best of our knowledge, this is the first work on statistical inference of a data-dependent policy in on-policy settings in sequential decision making with infinite horizons.
To study the asymptotic properties of our proposed CI, we focus on tensor-product spline and wavelet series estimators. Our technical contributions are described as follows. First, we introduce a bidirectional-asymptotic framework that allows either \( n \) or \( T \) to approach infinity. Our major technical contribution is to derive a nonasymptotic error bound for the spectral norm of sums of mean zero random matrices formed by the data transactions from MDP as a function of \( n, T \) and \( L \) (see e.g. Lemma E.2). This result is important in studying the limiting distribution of series estimators under such a theoretical framework.

Second, for policies that are estimated by Q-learning type algorithms such as the greedy gradient Q-learning, fitted Q-iteration and deep Q-network (Mnih et al., 2015), we relate the convergence rate of their values to the prediction error of the corresponding estimated Q-functions. We show in Theorems 3 and 4 that the values can converge at faster rates than the estimated Q-functions under certain margin type conditions on the optimal Q-function. To the best of our knowledge, these findings have not been discovered in the reinforcement-learning literature. Our theorems form a basis for researchers to study the value properties of Q-learning type algorithms. Moreover, our theoretical results are consistent with findings in point treatment studies where there is only one single decision point (see e.g. Luedtke & van der Laan, 2016; Qian & Murphy, 2011). However, the derivation of Theorems 3 and 4 is more involved since the value function in our settings is an infinite series involving both immediate and future rewards.

Third, when these basis functions are used, we mathematically characterize the approximation error of the Q-function as a function of \( L \), the dimension of the state variables, and the smoothness of the Markov transition function and the conditional mean of the immediate reward as a function of the state-action pair. This offers some guidance to practitioners on the choice of the number of basis functions \( L \) when some prior knowledge on the degree of smoothness of the aforementioned functions are available.

The rest of the paper is organized as follows. We introduce the model setup in Section 2. In Sections 3 and 4, we present the proposed off-policy and on-policy evaluation methods respectively. Simulation studies are conducted to evaluate the empirical performance of the proposed inference methods in Section 5. We apply the proposed inference method to the OhioT1DM data-set in Section 6. Finally, we conclude our paper by a discussion section.

### 2 | OPTIMAL POLICY IN INFINITE-HORIZON SETTINGS

We begin by introducing the notion of the optimal policy, the Q-function and the value function in infinite-horizon settings. Let \( X_{0,t} \in \mathcal{X} \) be the time-varying covariates collected at time point \( t \), \( A_{0,t} \in \mathcal{A} \) denote the action taken at time \( t \), and \( Y_{0,t} \) stand for the immediate reward observed. Here, \( \mathcal{X} \) and \( \mathcal{A} \) denote the state and action space respectively. We assume \( \mathcal{X} \) is a subspace of \( \mathbb{R}^d \) where \( d \) is the number of state vectors and \( \mathcal{A} \) is a discrete space \( \{0, 1, \ldots, m - 1\} \) where \( m \) denotes the number of actions. Suppose the system satisfies the following Markov assumption (MA),

\[
\Pr(X_{0,t+1} \in B | X_{0,t} = x, A_{0,t} = a, \{Y_{0,j}\}_{0 \leq j < t}, \{X_{0,j}\}_{0 \leq j < t}, \{A_{0,j}\}_{0 \leq j < t}) = \mathcal{P}(B | x, a),
\]

for some transition function \( \mathcal{P} \). Here, \( \mathcal{P} \) defines the next state distribution conditional on the current state-action pair. Moreover, suppose the following conditional mean independence assumption (CMIA) holds
\[
\mathbb{E}(Y_{0,t}|X_{0,t} = x, A_{0,t} = a, \{Y_{0,j}\}_{0 \leq j < t}, \{X_{0,j}\}_{0 \leq j < t}, \{A_{0,j}\}_{0 \leq j < t}) \\
= \mathbb{E}(Y_{0,t}|X_{0,t} = x, A_{0,t} = a) = r(x, a),
\]
for some reward function \(r\). By MA, CMIA automatically holds when \(Y_{0,t}\) is a deterministic function of \(X_{0,t}, A_{0,t}\) and \(X_{0,t+1}\) that measures the system's status at time \(t + 1\). The latter is satisfied in our real data application (see Section 6 for details) and is commonly assumed in the reinforcement learning literature. CMIA is thus weaker than this condition. MA and CMIA are important to guarantee the existence of an optimal policy (see Equation 1) and derive the bidirectional-asymptotic theory of the proposed CI (see the discussions below Theorem 1). We assume both assumptions hold throughout this paper.

In the following, we focus on the class of stationary policies that map the covariate space \(X\) to probability mass functions on \(A\). Let \(\pi(\cdot|\cdot)\) denote such a policy. It satisfies \(\pi(a|x) \geq 0\), for any \(a \in A\), \(x \in X\) and \(\sum_{a \in A} \pi(a|x) = 1\), for any \(x \in X\). For a deterministic policy, we have \(\pi(a|x) \in [0, 1]\), for any \(a \in A\), \(x \in X\). Under \(\pi\), a decision maker will set \(A_{0,t} = a\) with probability \(\pi(a|X_{0,t})\) at time \(t\). For such a policy and a given discounted factor \(0 \leq \gamma < 1\), let \(V(\pi; x)\) denote the value function

\[
V(\pi; x) = \sum_{t \geq 0} \gamma^t \mathbb{E}(Y_{0,t} | X_{0,0} = x),
\]

where the expectation \(\mathbb{E}(\cdot)\) is taken by assuming that the system follows the policy \(\pi\). The rate \(\gamma\) reflects a trade-off between immediate and future rewards. If \(\gamma = 0\), the agent tends to choose actions that maximize the immediate reward. As \(\gamma\) increases, the agent will consider future rewards more seriously. Under CMIA, we have

\[
V(\pi; x) = \sum_{t \geq 0} \gamma^t \mathbb{E}\{\mathbb{E}(Y_{0,t} | X_{0,t}, A_{0,t})|X_{0,0} = x\} = \sum_{t \geq 0} \gamma^t \sum_{a \in A} \mathbb{E}(\pi(a|X_{0,t})r(X_{0,t}, a)|X_{0,0} = x). \]

Similar to Theorem 6.2.12 of Puterman (1994), we can show under the given conditions that there exists at least one optimal policy \(\pi^{opt}\) that satisfies

\[
V(\pi^{opt}; x) \geq V(\pi; x), \quad \forall \pi, x. \quad (1)
\]

To better understand \(\pi^{opt}\), we introduce the state-action function (Q-function) under a policy \(\pi\) as

\[
Q(\pi; x, a) = \sum_{t \geq 0} \gamma^t \mathbb{E}(Y_{0,t} | X_{0,0} = x, A_{0,0} = a).
\]

Let \(Q^{opt}\) denote the optimal Q-function, that is, \(Q^{opt}(\cdot, \cdot) = \sup_{\pi} Q(\pi; \cdot, \cdot)\). It can be shown that \(\pi^{opt}\) satisfies

\[
\pi^{opt}(a|x) = 0 \text{ if } a \notin \arg \max_{a'} Q^{opt}(x, a'), \quad \forall x, a.
\]

There exist infinitely many optimal policies when \(\arg \max_a Q^{opt}(x, a)\) is not unique for some \(x \in X\). Let \(\Pi^{opt}\) denote the set consisting of all these optimal policies. Define
\[ \pi_0^{\text{opt}}(a|x) = \begin{cases} 1, & \text{if } a = \text{sarg max}_{a'} Q^{\text{opt}}(x, a'), \\ 0, & \text{otherwise,} \end{cases} \]  

(2)

where sarg max denotes the smallest maximizer when the argmax is not unique. Such a deterministic optimal policy may be appealing in medical studies. For example, in optimal dose studies, it is preferred to assign each patient the smallest optimal dose level to avoid toxicity.

### 3 | OFF-POLICY EVALUATION

#### 3.1 | Inference of the value under a fixed policy

Let \( n \) denote the number of trajectories in the dataset. For the \( i \)th trajectory, let \( \{A_{i,t}\}_{t \geq 0}, \{X_{i,t}\}_{t \geq 0} \) and \( \{Y_{i,t}\}_{t \geq 0} \) denote the sequence of actions, states and rewards respectively. It is worth mentioning that the time points are not necessarily homogeneous across different trajectories. Suppose the data are generated according to a fixed policy \( b(\cdot|\cdot) \), better known as the behavior policy such that \( \{X_0, t, A_0, t, Y_0, t\} \) are i.i.d copies of \( \{(X_{0,t}, A_{0,t}, Y_{0,t})\}_{t \geq 0} \). The observed data can thus be summarized as \( \{(X_{i,t}, A_{i,t}, Y_{i,t}, X_{i,t+1})\}_{0 \leq t < T_i, 1 \leq i \leq n} \), where \( T_i \) is the termination time of the \( i \)th trajectory. The goal of off-policy evaluation is to learn the value under a target policy \( \pi(\cdot|\cdot) \), possibly different from \( b(\cdot|\cdot) \).

#### 3.1.1 | Modelling value or Q-function?

Luckett et al. (2019) showed that the value function satisfies

\[ \mathbb{E} \left[ \frac{\pi(A_{i,t}, X_{i,t})}{b(A_{i,t}, X_{i,t})} \left[ Y_{i,t} + \gamma V(\pi; X_{i,t+1}) - V(\pi; X_{i,t}) \right] \right] = 0. \]  

(3)

Based on Equation (3), they directly modelled the value function, constructed an estimator for the integrated value under their estimated optimal policy and proved that it is asymptotically normal (see Theorem 4.3, Luckett et al., 2019).

Following their procedure, for a fixed policy \( \pi \), one might estimate \( V(\pi; \cdot) \) nonparametrically and construct the CI using the resulting estimate. However, such an approach might not be appropriate for polices that are discontinuous functions of the covariates. To better illustrate this, notice that \( V(\pi, \cdot) \) satisfies the following Bellman equation

\[ V(\pi; x) = \sum_{a \in A} \pi(a|x) \left\{ r(x, a) + \gamma \int_{x'} V(\pi; x') P(dx'|x, a) \right\} C(\pi, x, a). \]  

(4)

When \( P \) satisfies certain smoothness conditions (see Condition A1 below), we have
\[
\left| \int x' V(\pi; x') P(dx' | x_1, a) - \int x' V(\pi; x') P(dx' | x_2, a) \right| \\
\leq \int x' |V(\pi; x')||P(dx' | x_1, a) - P(dx' | x_2, a)| \to 0, \quad \text{as } \|x_1 - x_2\|_2 \to 0, \tag{5}
\]

for any \(\pi\). Suppose \(r(\cdot, a)\) is continuous for any \(a \in A\). Then \(C(\pi, x, a)\) is continuous in \(x\) for any \(\pi\) and \(a\). When \(\pi\) is a non-continuous function of \(x\), it follows from Equation (4) that \(V(\pi, \cdot)\) is not continuous either. However, many nonparametric methods, such as kernel smoothers and series estimation, require the underlying function to possess certain degree of smoothness in order to achieve estimation consistency. Notice that any non-constant deterministic policy has jumps and is not continuous at certain points (such as the optimal policy \(\pi_0\) given in Equation 2). This poses significant challenges in performing inference to these policies.

To allow valid inference for both deterministic and random policies, we consider modelling the Q-function. Under CMIA, we have

\[
Q(\pi; x, a) = \sum_{t \geq 0} \gamma^t E_x \{ r(X_{0,t}, A_{0,t}) | X_{0,0} = x, A_{0,0} = a \}.
\]

This together with MA yields

\[
Q(\pi; x, a) = r(x, a) + \gamma E_x \left[ \sum_{t \geq 0} \gamma^t E_x r(X_{0,t}, A_{0,t}) | X_{0,1}, A_{0,1} \right] \left| X_{0,0} = x, A_{0,0} = a \right.
\]

\[= r(x, a) + \gamma E_x \{ Q(X_{0,1}, A_{0,1}) | X_{0,0} = x, A_{0,0} = a \}. \]

As a result, the Q-function satisfies the following Bellman equation

\[
Q(\pi; x, a) = r(x, a) + \gamma \sum_{a' \in A} \int x' q(\pi; x', a') P(dx' | x, a).
\tag{6}
\]

Similar to Equation (5), we can show the second term on the right-hand-side (RHS) of Equation (5) is a smooth function of \(x\) for any \(\pi\) and \(a\). When \(r(\cdot, a)\) is smooth, so is \(Q(\pi, \cdot, a)\). To formally establish these results, we introduce the notion of \(p\)-smoothness (also known as Hölder smoothness with exponent \(p\)) below.

Let \(h(\cdot)\) be an arbitrary function on \(\mathbb{R}\). For a \(d\)-tuple \(a = (a_1, \ldots, a_d)^T\) of nonnegative integers, let \(D^a\) denote the differential operator:

\[
D^a h(x) = \frac{\partial^{|a|}|h(x)}{\partial x_1^{a_1} \cdots \partial x_d^{a_d}}.
\]

Here, \(x_j\) denotes the \(j\)th element of \(x\). For any \(p > 0\), let \(|p|\) denote the largest integer that is smaller than \(p\). Define the class of \(p\)-smooth functions as follows:

\[
\Lambda(p, c) = \left\{ h: \sup_{||a|| \leq |p|} \sup_{x \in \mathbb{R}} |D^a h(x)| \leq c, \sup_{||a|| = |p|} \sup_{x, y \in \mathbb{R}, x \neq y} \frac{|D^a h(x) - D^a h(y)|}{\|x - y\|_2^{p-|p|}} \leq c \right\}.
\]
When $0 < p \leq 1$, we have $|p| = 0$. It is equivalent to require $h$ to satisfy $\sup_{x,y} |h(x) - h(y)|/\|x - y\|_2^p \leq c$. The notion of $p$-smoothness is thus reduced to the Hölder continuity.

For any $x \in \mathbb{X}$, $a \in \mathcal{A}$, suppose the transition kernel $P(\cdot| x, a)$ is absolutely continuous with respect to the Lebesgue measure. Then there exists some transition density function $q$ such that $P(dx'| x, a) = q(x'| x, a)dx'$. We impose the following condition.

(A1) There exist some $p, c > 0$ such that $r(\cdot, a), q(x'| \cdot, a) \in \Lambda(p, c)$ for any $a \in \mathcal{A}$, $x' \in \mathbb{X}$.

Lemma 1 Under A1, there exists some constant $c' > 0$ such that $Q(\pi; \cdot, a) \in \Lambda(p, c')$ for any policy $\pi$ and $a \in \mathcal{A}$.

Lemma 1 implies the Q-function has bounded derivatives up to order $|p|$. This motivates us to first estimate the Q-function and then derive the corresponding value estimators based on the relation $V(\pi; x) = \sum_{a \in \mathcal{A}} \pi(a| x)Q(\pi; x, a)$. By the Bellman equation (6), we can show the Q-function satisfies

$$E \left[ \left\{ Y_{i,t} + \gamma \sum_{a \in \mathcal{A}} Q(\pi; X_{i,t+1}, a)\pi(a| X_{i,t+1}) - Q(\pi; X_{i,t}, A_{i,t}) \right\} | X_{i,t}, A_{i,t} \right] = 0. \quad (7)$$

The above equation forms a basis of our methods to learn $Q(\pi; \cdot, \cdot)$ (see details in the next section). In contrast to Equation (3), the sampling ratio $\pi(\alpha|x)/b(\alpha|x)$ does not appear in Equation (7). This is because $A_{i,t}$ is the only sampling action and no further actions are involved in Equation (7). As a result, our method does not require correct specification of the behavior policy. Nor do we need to estimate it from the observed dataset. This is another advantage of modelling the Q-function over the value.

3.1.2 Method

We describe our procedure in this section. We propose to approximate $Q(\pi; \cdot, \cdot)$ based on linear sieves, which takes the form

$$Q(\pi; x, a) \approx \Phi_L^T(x)\beta_{\pi,a}, \quad \forall x \in \mathbb{X}, \ a \in \mathcal{A},$$

where $\Phi_L(\cdot) = \{\phi_{L,1}(\cdot), \ldots, \phi_{L,L}(\cdot)\}^T$ is a vector consisting of $L$ sieve basis functions, such as splines or wavelet bases (see e.g. Huang, 1998, for choices of basis functions). We allow $L$ to grow with the sample size to reduce the bias of the resulting estimates. Under certain mild conditions, there exist some $\{\beta_{\pi,a}^*\}_{a \in \mathcal{A}}$ that satisfy

$$E \left\{ Y_{i,t} + \gamma \sum_{a \in \mathcal{A}} \Phi_L^T(X_{i,t+1})\beta_{\pi,a}^* \pi(a| X_{i,t+1}) - \Phi_L^T(X_{i,t})\beta_{\pi,a'}^* \right\} | X_{i,t}, (A_{i,t} = a') = 0,$$

for any $a' \in \mathcal{A}$. Recall that $\mathcal{A} = \{0, 1, \ldots, m-1\}$. Define $\beta_{\pi}^* = (\beta_{\pi,1}^*, \ldots, \beta_{\pi,m}^*)^T$, $\xi(x,a) = (\Phi_L^T(x)\pi(0|x), \Phi_L^T(x)\pi(1|x), \ldots, \Phi_L^T(x)\pi(m-1|x))^T$, $U_\pi(x) = (\Phi_L^T(x)\pi(0|x), \Phi_L^T(x)\pi(1|x), \ldots, \Phi_L^T(x)\pi(m-1|x))^T$,.
\( \xi_{i,t} = \xi(X_{i,t}, A_{i,t}) \), \( U_{x,i,t} = U_x(X_{i,t}) \). The above equation can be rewritten as \( \mathbb{E}_{i,t}(\xi_{i,t} - \gamma U_{x,i,t+1})^\top \beta^*_x = \mathbb{E}_{i,t}Y_{i,t} \). Based on the observed data, we propose to estimate \( \beta^*_x \) by solving

\[
\hat{\beta}_x = \left( \frac{1}{\sum_i T_i} \sum_{t=0}^{T_i-1} \xi_{i,t}(\xi_{i,t} - \gamma U_{x,i,t+1})^\top \right)^{-1} \left( \frac{1}{\sum_i T_i} \sum_{t=0}^{T_i-1} \xi_{i,t}Y_{i,t} \right).
\]

Let \( \hat{\beta}_x = (\hat{\beta}_{x,1}^\top, \ldots, \hat{\beta}_{x,m}^\top)^\top \), we propose to estimate \( V(\pi; x) \) by

\[
\hat{V}(\pi; x) = \sum_{a \in \mathcal{A}} \Phi_L^T(x)\hat{\beta}_{x,a} \pi(a|x) = U_x^T(x)\hat{\beta}_x.
\]

A two-side CI is given by

\[
\begin{bmatrix}
\hat{V}(\pi; x) - z_{a/2} \left( \sum_i T_i \right)^{-1/2} \hat{\sigma}(\pi; x), \\
\hat{V}(\pi; x) + z_{a/2} \left( \sum_i T_i \right)^{-1/2} \hat{\sigma}(\pi; x)
\end{bmatrix}, \tag{8}
\]

where \( z_a \) denotes the upper \( a \)th quantile of a standard normal distribution, and

\[
\hat{\sigma}^2(\pi; x) = U_x^T(x)\hat{\Sigma}^{-1}_x \hat{\Delta}_x (\hat{\Sigma}_x^\top)^{-1} U_x(x),
\]

where

\[
\hat{\Delta}_x = \frac{1}{\sum_i T_i} \sum_{t=0}^{T_i-1} \sum_{i,t} \xi_{i,t}^\top \xi_{i,t} \left\{ Y_{i,t} + \gamma \sum_{a \in \mathcal{A}} \Phi_L^T(X_{i,t+1})\hat{\beta}_{x,a} \pi(a|X_{i,t+1}) - \Phi_L^T(X_{i,t})\hat{\beta}_{x,a} \right\}^2.
\]

Let \( \mathcal{G} \) be a reference distribution on the covariate space \( \mathcal{X} \). Define the following integrated value function

\[
V(\pi; \mathcal{G}) = \int_{x \in \mathcal{X}} V(\pi; x)\mathcal{G}(dx).
\]

By setting \( \mathcal{G}(\cdot) \) to be a Dirac measure \( \delta_x(\cdot) \), that is, \( \mathcal{G}(\mathcal{X}) = l(x \in \mathcal{X}) \), \( \forall \mathcal{X} \subseteq \mathcal{X}, V(\pi; \mathcal{G}) \) is reduced to \( V(\pi; x) \). Let \( v_0(\cdot) \) be the probability density function of \( X_{0,0} \). By setting \( \mathcal{G}(dx) = v_0(x)dx \), we obtain

\[
V(\pi; \mathcal{G}) = \int_{x \in \mathcal{X}} V(\pi; x)v_0(x)dx.
\]

Based on \( \hat{\beta}_x \), a two-side CI for \( V(\pi; \mathcal{G}) \) is given by

\[
\begin{bmatrix}
\hat{V}(\pi; \mathcal{G}) - z_{a/2} \left( \sum_i T_i \right)^{-1/2} \hat{\sigma}(\pi; \mathcal{G}), \\
\hat{V}(\pi; \mathcal{G}) + z_{a/2} \left( \sum_i T_i \right)^{-1/2} \hat{\sigma}(\pi; \mathcal{G})
\end{bmatrix}, \tag{9}
\]
where
\[ \hat{V}(\pi; G) = \int_{x \in \mathbb{X}} \hat{V}(\pi; x) G(dx), \]  
(10)
\[ \hat{\sigma}^2(\pi; G) = \left\{ \int_{x \in \mathbb{X}} U_\pi(x) G(dx) \right\}^T \hat{\Sigma}_\pi^{-1} \hat{\Omega}_\pi \left( \hat{\Sigma}_\pi^{-1} \right)^{-1} \left\{ \int_{x \in \mathbb{X}} U_\pi(x) G(dx) \right\}. \]
(11)

3.1.3 | Theory

In this section, we focus on proving the validity of the proposed CIs in Equation (9). By setting \( G(\cdot) = \delta_x(\cdot) \), it implies that the CI in Equation (8) achieves nominal coverage as well. To simplify the presentation, we assume \( T_1 = T_2 = \cdots = T_n = T \), all the covariates are continuous and \( \mathbb{X} = [0, 1]^d \). Our theory is valid regardless of whether \( T \) is bounded or diverges to infinity. We remark that the boundedness of \( T \) does not mean we work on a finite-horizon setting, since \( T \) is the termination time of the study, not the final time step of each trajectory.

In addition, we restrict our attentions to two particular types of sieve basis functions, corresponding to tensor product of B-splines with degree \( r \) and dimension \( L \) or Wavelets with regularity \( r \) and dimension \( L \). See section 6 of Chen and Christensen (2015) for a brief review of these sieve bases. This together with A1 implies that there exists a set of vectors \( \{ \beta^*_{x,a} \}_{a \in A} \) that satisfy \( \sup_{x \in \mathbb{X}, a \in A} |Q(\pi; x, a) - \Phi_L(\pi)\beta^*_{x,a}| = O(L^{-p/d}) \). See section 2.2 of Huang (1998) for detailed discussions on the approximation power of these sieve bases.

Following the behavior policy \( b(\cdot|\cdot) \), the set of variables \( \{X_{0,t}\}_{t \geq 0} \) forms a time-homogeneous Markov chain. Its transition kernel \( P_X \) is given by

\[ P_X(B|x) = \Pr(X_{0,1} \in B|X_{0,0} = x) = \sum_{a \in A} P(B|x, a) b(a|x), \quad \forall B \in \mathbb{X}. \]

We impose the following assumptions.

(A2) The Markov chain \( \{X_{0,t}\}_{t \geq 0} \) has an unique invariant distribution with some density function \( \mu(\cdot) \). The density functions \( \mu \) and \( \nu_0 \) are uniformly bounded away from 0 and \( \infty \).

(A3) Suppose (i) and (ii) hold when \( T \to \infty \) and (i) holds when \( T \) is bounded.

(i) \( \lambda_{\min} \left[ \sum_{t=0}^{T-1} \mathbb{E} \{ \xi_t, \xi_t^T - \gamma^2 u_x(X_{0,t}, A_{0,t}) u_x^T(X_{0,t}, A_{0,t}) \} \right] \geq T \bar{c} \) for some constant \( \bar{c} > 0 \), where \( u_x(\cdot, a) = \mathbb{E} \{ U_x(X_{0,1})|X_{0,0} = x, A_{0,0} = a \} \) and \( \lambda_{\min}(K) \) denotes the minimum eigenvalue of a matrix \( K \).

(ii) The Markov chain \( \{X_{0,t}\}_{t \geq 0} \) is geometrically ergodic.

We make a few remarks. First, we do not require the limiting density function \( \mu \) to be equal to the initial state density \( \nu_0 \).

Second, Condition A3(i) guarantees the matrix \( \mathbb{E} \hat{\Sigma}_\pi \) is invertible. In Section C.1 of the supplementary article, we show A3(i) is automatically satisfied when \( \mu = \nu_0 \), the target policy \( \pi \) is deterministic and \( b \) is the \( \epsilon \)-greedy policy with respect to \( \pi \) that satisfies \( \epsilon \leq 1 - \gamma^2 \).

Third, we present the detailed definition of geometric ergodicity in Appendix A to save space. Suppose the Markov chain \( \{X_{0,t}\}_{t \geq 0} \) has a finite state space. Assume \( P_X \) is diagonalizable. Then A3(ii) holds when the second largest eigenvalue of \( P_X \) is strictly smaller than 1. When \( X_{0,t} \)'s are generated by the vector autoregressive process \( \mathbb{E} \{ X_{0,t}|X_{0,t-1} \} \) = \( f(X_{0,t-1}) \) for some function \( f \), Saikkonen (2001) provided sufficient conditions that ensure the geometric ergodicity of the Markov chain.
Finally, when \( v_0 = \mu, \{X_{0,t}\}_{t \geq 0} \) is stationary. Under Condition A3(ii), it follows from Theorem 3.7 of Bradley (2005) that \( \{X_{0,t}\}_{t \geq 0} \) is exponentially \( \beta \)-mixing (see the proof of Lemma E.2 for details). When \( T \to \infty \), A3(ii) enables us to derive matrix concentration inequalities for \( \hat{\Sigma}_\pi \). This together with A3(i) implies that \( \hat{\Sigma}_\pi \) is invertible, with probability approaching 1 (wpa1). We remark that A3(ii) is not needed when \( T \) is bounded.

For any \( x \in \mathcal{X}, a \in \mathcal{A} \), define

\[
\omega_\pi(x, a) = \mathbb{E}\left\{ Y_{0,0} + \gamma \sum_{a \in \mathcal{A}} \pi(a|X_{0,1})Q(x; X_{0,1}, a) - Q(x; X_{0,0}, A_{0,0}) \right\}^2 | X_{0,0} = x, A_{0,0} = a.
\]

**Theorem 1** (Bidirectional asymptotics) Assume A1–A3 hold. Suppose \( L \) satisfies \( L = o(\sqrt{nT/\log(nT)}) \), \( L^{2p/d} \gg nT \{ 1 + \| \int_{\mathcal{X}} \Phi_L(x)G(dx) \|_2^{-2} \} \), and there exists some constant \( c_0 \geq 1 \) such that \( \omega_\pi(x, a) \geq c_0^{-1} \) for any \( x \in \mathcal{X}, a \in \mathcal{A} \) and \( \Pr(\max_{0 \leq t \leq T-1} |Y_{0,t}| \leq c_0) = 1 \). Then as either \( n \to \infty \) or \( T \to \infty \), we have

\[
\sqrt{nT} \hat{\pi}^{-1}(\pi; \mathcal{G}) (\hat{V}(\pi; \mathcal{G}) - V(\pi; \mathcal{G})) \xrightarrow{d} N(0, 1).
\]

A sketch for the proof of Theorem 1 is given in Appendix E.1. Under the conditions in Theorem 1, we can show that \( \hat{\pi}(\pi; \mathcal{G}) \) converges almost surely to some \( \pi(\pi; \mathcal{G}) \). The form of \( \pi(\pi; \mathcal{G}) \) is given in Section E.5. In addition, we have

\[
\frac{\hat{V}(\pi; \mathcal{G}) - V(\pi; \mathcal{G})}{(nT)^{-1/2} \hat{\sigma}(\pi; \mathcal{G})} = \frac{(nT)^{-1/2}}{\sigma(\pi; \mathcal{G})} \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ \int_{x \in \mathcal{X}} U_\pi(x)G(dx) \right\} \Sigma_\pi^{-1} \varepsilon_{i,t} \varepsilon_{i,t} + o_p(1),
\]

where \( \Sigma_\pi = \mathbb{E} \hat{\Sigma}_\pi \) and

\[
\varepsilon_{i,t} = Y_{i,t} + \gamma \sum_{a \in \mathcal{A}} Q(\pi; X_{i,t+1}, a) \pi(a|X_{i,t+1}) - Q(\pi; X_{i,t}, A_{i,t}).
\]

By MA, CMIA and Equation (7), the leading term on the RHS of Equation (12) forms a mean-zero martingale (details can be found in Section E.5). As either \( n \) or \( T \) grows to infinity, the asymptotic normality follows from the martingale central limit theorem.

When \( \sigma(\pi; \mathcal{G}) \) is bounded away from zero, it can be seen from Equation (12) that \( \hat{V}(\pi; \mathcal{G}) - V(\pi; \mathcal{G}) = O_p(n^{-1/2}T^{-1/2}) \). That is, the proposed value estimator converges at a rate of \( (nT)^{-1/2} \). In contrast, AIPW-type estimators typically converge at a rate of \( n^{-1/2} \) and are thus not suitable for settings with only a few trajectories.

### 3.2 Inference of the value under an (estimated) optimal policy

For simplicity, we assume \( T_1 = T_2 = \cdots = T_n = T \) throughout this section. Consider an estimated policy \( \hat{\pi} \), computed based on the data \( \{(X_{i,t}, A_{i,t}, Y_{i,t}, X_{i,t+1})\}_{0 \leq t < T, 1 \leq i \leq n} \). The integrated value under \( \hat{\pi} \) is given by

\[
V(\hat{\pi}; \mathcal{G}) = \int_{x \in \mathcal{X}} V(\hat{\pi}; x)G(dx).
\]
We will require the value of $\hat{\pi}$ to converge to some fixed policy $\pi^*$ (possibly different from $\pi^*_{\text{opt}}$), that is,

$$V(\hat{\pi}; G) - V(\pi^*; G) \xrightarrow{P} 0.$$  

(13)

In this section, we focus on constructing CIs for $V(\hat{\pi}; G)$ and $V(\pi^*; G)$.

### 3.2.1 The challenge

We begin by outlining the challenge of obtaining inference in the nonregular cases. Suppose $\pi^* \in \Pi^*_{\text{opt}}$. When the optimal policy is not unique, $\hat{\pi}$ might not converge to a fixed policy, despite that its value converges (see Equation 13). To better illustrate this, suppose $\hat{\pi}$ is computed by some $Q$-learning type algorithms, that is,

$$\hat{\pi}(a|x) = \begin{cases} 1, & \text{if } a = \text{sarg max}_{a'} Q(x, a'), \\ 0, & \text{otherwise}, \end{cases}$$  

(14)

where $Q(\cdot, \cdot)$ denotes some consistent estimator for $Q^*_{\text{opt}}(\cdot, \cdot)$. Assume there exists a subset $X_0$ of $X$ with positive Lebesgue measure such that the argmax of $Q^*_{\text{opt}}(x, \cdot)$ is not uniquely defined for any $x \in X_0$. Then $\hat{\pi}(\cdot|x)$ might not converge to a fixed quantity for any $x \in X_0$.

Consider the plug-in estimator $\hat{V}(\hat{\pi}; G)$ for $V(\hat{\pi}; G)$. Similar to Equation (12), we can show

$$\frac{\hat{V}(\hat{\pi}; G) - V(\hat{\pi}; G)}{(nT)^{-1/2} \hat{\sigma}(\hat{\pi}; G)} = \left( \sum_{i,t} \left\{ \left( \int_{x \in X} U_{\hat{\pi}}(x)G(dx) \right)^{T} \right\} \right) \xrightarrow{\text{a.s.}} \xi_{i,t} + o_p(1),$$  

(15)

where

$$\xi_{i,t} = Y_{i,t} + \gamma \sum_{a \in A} Q(\hat{\pi}; X_{i,t+1}, a)\hat{\pi}(a|X_{i,t+1}) - Q(\hat{\pi}; X_{i,t}, A_{i,t}).$$

When $\hat{\pi}$ does not converge, $\sigma(\hat{\pi}; G), U_{\hat{\pi}}(x), \Sigma_{\hat{\pi}}$, and $\xi_{i,t}$ will fluctuate randomly and might not stabilize. Since these quantities depend on the data as well, the martingale structure is violated. As a result, the leading term on the RHS of Equation (15) does not have a well tabulated limiting distribution. Thus, CIs based on $\hat{V}(\hat{\pi}; G)$ will fail to maintain the nominal coverage probability.

To allow for valid inference, we use a sequential value evaluation procedure to construct the CI. That is, we propose sequentially estimating the optimal policy and evaluating its value using different data subsets. This allows us to treat the estimated optimal policy as known conditional on past observations (see Equation E.6 in Appendix E.2). The martingale CLT can thus be applied to obtain the limiting distribution for our estimator (see Equation E.8 and the related discussions). We detail our procedure in the next section.

### 3.2.2 SAVE for the value under an (estimated) optimal policy

We begin by dividing $I_0 = \{(i, t): 1 \leq i \leq n, 0 \leq t < T\}$ into $K$ non-overlapping subsets, denoted by $I_1, I_2, \ldots, I_K$. At the $k$th step, we use the sub-dataset
to compute an estimated optimal policy (denoted by $\hat{\pi}_{T_k}$). Then we apply the proposed procedure in Section 3.1 to dataset in the $(k+1)$th block $O_{k+1} = \{ (X_{i,t}, A_{i,t}, Y_{i,t}, X_{i,t+1}) : (i,t) \in I_{k+1} \}$ to compute its value estimator $\hat{V}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathcal{G})$ and the associated standard error $\hat{\sigma}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathcal{G})$ (Details are given in Appendix B.1).

As commented in the introduction, the data-splitting rule cannot be arbitrary. For any of the two tuples $(i_1, t_1)$ and $(i_2, t_2)$, define an order $(i_2, t_2) > (i_1, t_1)$ if either $t_2 > t_1$ or $i_2 > i_1$. For any $(i_2, t_2) \in I_{k+1}$, we require the following:

$$(i_2, t_2) > (i_1, t_1), \quad \forall (i_1, t_1) \in I_k.$$  \hspace{1cm} (16)

Then $\hat{\pi}_{T_k}$ depends on the $i_2$th patent’s trajectory only through \{(X_{i_2,t'}, A_{i_2,t'}, Y_{i_2,t'})\}_{t'<t_2}$ and $X_{i_2,t'} A_{i_2,t'}$. Under MA and CMIA, Equation (7) still holds with $\pi = \hat{\pi}_{T_k}$ any $(i, t) = (i_2, t_2)$. Similar to Equation (12), we can show conditional on the observations in $\bar{I}_k$, $\sqrt{nT/K} \hat{V}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathcal{G})$ is asymptotically normal with variance consistently estimated by $\hat{\sigma}_{I_{k+1}}^2(\hat{\pi}_{T_k}; \mathcal{G})$.

Our final estimator $\hat{V}(\mathcal{G})$ is defined as a weighted average of these $\hat{V}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathcal{G})$’s. Specifically, we set

$$\hat{V}(\mathcal{G}) = \left\{ \sum_{k=1}^{K-1} \frac{1}{\hat{\sigma}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathcal{G})} \right\}^{-1} \left\{ \sum_{k=1}^{K-1} \hat{V}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathcal{G}) \right\}.$$  \hspace{1cm}

The inclusion of the inverse weight $1/\hat{\sigma}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathcal{G})$ is necessary for the theoretical development of asymptotic normality of $\hat{V}(\mathcal{G})$ (see Equation E.8). Our CI is given by

$$[\hat{V}(\mathcal{G}) - z_{a/2}\{nT(K-1)/K\}^{-1/2}\bar{\sigma}(\mathcal{G}), \hat{V}(\mathcal{G}) + z_{a/2}\{nT(K-1)/K\}^{-1/2}\bar{\sigma}(\mathcal{G})]. \hspace{1cm} (17)$$

where $\bar{\sigma}(\mathcal{G}) = (K-1)/\sum_{k=1}^{K-1} \frac{1}{\hat{\sigma}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathcal{G})}$. It remains to specify $I_1, I_2, \ldots, I_K$ that satisfy Equation (16). Consider some positive integers $n_{\min} \leq n$, $T_{\min} \leq T$. Assume $n$ and $T$ are divisible by $n_{\min}$ and $T_{\min}$ respectively. Let $K_n = n/n_{\min}$ and $K_T = T/T_{\min}$. We set $K = K_n K_T$. For any $1 \leq k_n \leq K_n$, $1 \leq k_T \leq K_T$, define a set $I(k_n, k_T)$ by

$$\{(i,t) : (k_n-1)n_{\min} < i \leq k_n n_{\min}, (k_T-1)T_{\min} < t < k_T T_{\min}\}.$$  

Thus, each block $I(k_n, k_T)$ contains data from $n_{\min}$ trajectories with $T_{\min}$ decision time points. Below, we introduce two special examples.

1. When only a few trajectories are available, we may set $n_{\min} = n$. Then, the blocks are constructed according to the times that decisions were being made.
2. When each trajectory contains a very short time period, we may set $T_{\min} = T$. Then, the observations are divided according to the trajectories they belong to.

We order these blocks by

$I(1, 1), I(2, 1), \ldots, I(K_n, 1), I(1, 2), I(2, 2), \ldots, I(K_n, 2), \ldots, I(1, k_T), \ldots, I(K_n, K_T)$.  \hspace{1cm}
Based on this order, we set $I_k = I(n(k), T(k))$ where $n(k)$ and $T(k)$ are the unique positive integers that satisfy $k = n(k) + (T(k) - 1)K_n$. For any $k_2 > k_1$, we have either $n(k_2) > n(k_1)$ or $T(k_2) > T(k_1)$. Thus, the proposed data-splitting rule guarantees (16) holds for any $k$.

In Theorem 2 below, we establish the validity of our CI in Equation (17). It relies on Condition A3* and A4. A3* is very similar to A3 and we present the detailed definition in Appendix A to save space.

\[(A4) \|V(\hat{\pi}_T; \mathcal{G}) - V(\pi^*; \mathcal{G})\| = O(\|I_k\|^{-b_0})\] for some $b_0 > 1/2$ such that $(nT)^{b_0-1/2} \gg \|\int \Phi(x)(\mathcal{G})dx\|^{-1}$, where the big-O term is uniform in $k$.

Set $I = I_0$. By Markov’s inequality, it is immediate to see A4 implies that Condition (13) holds. When the tensor-product B-splines are used, we have $\lim \inf_{I} \|\int \Phi(x)(\mathcal{G})dx\| > 0$. Thus, it is equivalent to require $E[V(\hat{\pi}_T; \mathcal{G}) - V(\pi^*; \mathcal{G})] = O(\|I_k\|^{-b_0})$ for some $b_0 > 1/2$. In Section 3.2.3, we discuss the rate $b_0$ in detail when $\pi^* = \pi^\text{opt}$ and $\hat{\pi}$ is a greedy policy derived based on some Q-learning algorithms.

\[\text{Theorem 2 (Bidirectional asymptotics) Assume A1–A2, A3* and A4 hold. Suppose $K = O(1)$ and $L$ satisfies $L = o\{\sqrt{nT}/\log(nT)\}$, $L^{p/d} \gg nT(1 + \|\int \Phi(x)(\mathcal{G})dx\|^{-1})$. Suppose $T_{\min} = T$ if $T$ is bounded. Assume there exists some constant $c_0 \geq 1$ such that $\omega_\pi(x, a) \geq c_0^{-1}$ for any $x, a, \pi$ and $Pr(\max_{0 \leq t \leq T} |V_{0t}| \leq c_0) = 1$. Then as either $n \to \infty$ or $T \to \infty$, we have}\]

\[
\sqrt{nT(K-1)/K\tilde{\sigma}^{-1}(\mathcal{G})[\tilde{V}(\mathcal{G}) - V(\hat{\pi}; \mathcal{G})]} \overset{d}{\to} N(0, 1),
\]

\[
\sqrt{nT(K-1)/K\tilde{\sigma}^{-1}(\mathcal{G})[\tilde{V}(\mathcal{G}) - V(\pi^*; \mathcal{G})]} \overset{d}{\to} N(0, 1).
\]

We provide a sketch for the proof of Theorem 2 in Appendix E.2.

### 3.2.3 Convergence of the value under an estimated optimal policy

For any $I \subseteq I_0$, we use $\hat{\pi}_I$ to denote an estimated optimal policy based on observations in $I$. Let $\hat{\pi}_I(\cdot, \cdot)$ denote some consistent estimator for $Q^{\text{opt}}(\cdot, \cdot)$ and $\hat{\pi}_I$ denote the greedy policy with respect to $\hat{\pi}_I(\cdot, \cdot)$ (see Equation 14).

In the following, we focus on relating $|V(\pi^{\text{opt}}; \mathcal{G}) - V(\hat{\pi}_I; \mathcal{G})|$ to the prediction loss $\hat{\pi}_I - Q^{\text{opt}}$. By definition, $V(\pi^{\text{opt}}; x) \geq V(\hat{\pi}_I; x)$, $\forall x \in \mathcal{X}$. Hence, $V(\pi^{\text{opt}}; \mathcal{G}) \geq V(\hat{\pi}_I; \mathcal{G})$. It suffices to provide an upper bound for $V(\pi^{\text{opt}}; \mathcal{G}) - V(\hat{\pi}_I; \mathcal{G})$. We introduce a margin-type condition A5 below.

\[(A5) \text{Assume there exist some constants } \alpha, \delta_0 > 0 \text{ such that}\]

\[
\lambda \left\{ x \in \mathcal{X}: \max_a Q^{\text{opt}}(x, a) - \max_{a' \in A - \arg \max_a Q^{\text{opt}}(x, a)} Q^{\text{opt}}(x, a') \leq \varepsilon \right\} = O(\varepsilon^\alpha),
\]

\[
\|\mathbb{G} \{ x \in \mathcal{X}: \max_a Q^{\text{opt}}(x, a) - \max_{a' \in A - \arg \max_a Q^{\text{opt}}(x, a)} Q^{\text{opt}}(x, a') \leq \varepsilon \} \| = O(\varepsilon^\alpha),
\]

where $\lambda$ denotes the Lebesgue measure, the big-O terms are uniform in $0 < \varepsilon \leq \delta_0$, and $\max_{a' \in A - \arg \max_a Q^{\text{opt}}(x, a)} Q^{\text{opt}}(x, a') = -\infty$ if the set $A - \arg \max_a Q^{\text{opt}}(x, a) = \emptyset$.

For each $x$, the quantity $\max_a Q^{\text{opt}}(x, a) - \max_{a' \in A - \arg \max_a Q^{\text{opt}}(x, a)} Q^{\text{opt}}(x, a')$ measures the difference in value between $\pi^{\text{opt}}$ and the policy that assigns the best suboptimal treatment(s) at the first decision point and follows $\pi^{\text{opt}}$ subsequently. In point treatment studies, Qian and
Murphy (2011) imposed a similar condition (see Equation 3, Qian & Murphy, 2011) to derive sharp convergence rate for the value under an estimated optimal individualized treatment regime. Here, we generalize their condition in infinite-horizon settings. A5 is also closely related to the margin condition commonly used to bound the excess misclassification error (Audibert & Tsybakov, 2007; Tsybakov, 2004).

The margin-type condition is mild. In Appendix A.3, we present detailed examples and show the condition holds under these examples. The following theorems summarize our results.

**Theorem 3**  Assume A1, Equations (18) and (19) hold. Suppose the following event occurs with probability at least $1 - O(|I|^{-\kappa})$ for any finite $\kappa > 0$,

$$\sup_{x \in \mathcal{X}, a \in \mathcal{A}} |\hat{Q}_I(x, a) - Q^{\text{opt}}(x, a)| = O(|I|^{-b_*}),$$

for some $b_* > 0$. Then $\mathbb{E}|V(\pi^{\text{opt}}; \mathcal{G}) - V(\hat{\pi}_I; \mathcal{G})| = O(|I|^{-b_*(1+a)})$.

In Theorem 3, we require the estimated Q-function to satisfy certain uniform convergence rate. In Theorem 4 below, we relax this condition by assuming that the integrated loss converges to zero at certain rate.

**Theorem 4**  Assume A1 and A5 hold. Suppose

$$\left( \mathbb{E}\int_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} |\hat{Q}_I(x, a) - Q^{\text{opt}}(x, a)|^2 dx \right)^{1/2} = O(|I|^{-b_*}),$$

for some $b_* > 0$. Then $\mathbb{E}|V(\pi^{\text{opt}}; \mathcal{G}) - V(\hat{\pi}_I; \mathcal{G})| = O(|I|^{-b_*(2+2a)/(2+a)})$.

It can be seen from Theorems 3 and 4 that the integrated value converges faster compared to the Q-function. We provide a sketch for the proofs of both theorems in Appendix E.3.

### 3.2.4 Applications

In this section, we provide several examples to illustrate the convergence rate of $\hat{Q}_I$. The proposed methods can be applied to evaluating the values under these estimated policies. The algorithm in Example 1 requires to impose a linear model assumption for the optimal Q-function. The algorithm in Example 2 allows more general nonlinear and nonparametric models for the optimal Q-function.

**Example 1** (Greedy gradient Q-learning) The optimal Q-function satisfies

$$Q^{\text{opt}}(x, a) = r(x, a) + \gamma \max_{x', a' \in \mathcal{A}} Q^{\text{opt}}(x', a') P(dx' | x, a),$$

for any $a$ and $x$, and hence

$$\mathbb{E} \left\{ Y_{t,t} + \gamma \max_{a' \in \mathcal{A}} Q^{\text{opt}}(X_{t,t+1}, a') - Q^{\text{opt}}(X_{t,t}, A_{t,t}) \right\} | X_{t,t}, A_{t,t} = 0.$$
Suppose we model $Q^{\text{opt}}(x, a)$ by linear sieves $\Phi^T(x)\theta_a$. Then we can compute $\{\hat{\theta}_{a,I}\}_{a \in A}$ by minimizing the following projected Bellman error:

$$\arg \min_{\{\theta_a\}_{a \in A}} \left( \sum_{(i,t) \in I} \delta_{i,t}(\{\theta_a\}_{a \in A}) \xi_{i,t} \right)^T \left( \sum_{(i,t) \in I} \xi_{i,t} \xi_{i,t}^T \right)^{-1} \left( \sum_{(i,t) \in I} \delta_{i,t}(\{\theta_a\}_{a \in A}) \xi_{i,t} \right),$$

where $\delta_{i,t}(\{\theta_a\}_{a \in A}) = Y_{i,t} + \gamma \max_{a' \in A} \Phi^T(x_{i,t+1}) \theta_{a'} - \Phi^T(x_{i,t}) \theta_{A_{i,t}}$. The above loss is non-smooth and non-convex as a function of $\{\theta_a\}_{a \in A}$. The estimator $\{\hat{\theta}_{a,I}\}_{a \in A}$ can be computed based on the greedy gradient Q-learning algorithm.

Assuming the optimal Q-function is correctly specified, Ertefaie and Strawderman (2018) established the consistency and asymptotic normality of the parameter estimates under the scenario where both $L$ and $T$ are fixed. Set $\hat{Q}_I(x, a) = \Phi^T(x)\hat{\theta}_{a,I}$. Using similar arguments in proving Theorem 1, we can show that with proper choice of $L$, super-$\frac{p}{(2p+d)}$ coverages at a rate of $O(|I|^{-p/(2p+d)})$ up to some logarithmic factors, with probability at least $1 - O(n^{-2}T^{-2})$. The condition in Theorem 3 thus holds for any $b_\epsilon < p/(2p + d)$.

**Example 2** (Fitted Q-iteration) In fitted Q-iteration (FQI), the optimal Q-function is approximated by some nonparametric models $Q(\cdot, \cdot, \theta)$ indexed by $\theta$. The parameter $\theta$ is iteratively updated by

$$\hat{\theta}_{k+1} = \arg \min_{\theta} \sum_{(i,t) \in I_k} \left\{ Y_{i,t} + \gamma \max_{a} Q(X_{i,t+1}, a; \hat{\theta}_k) - Q(X_{i,t}, A_{i,t}; \theta) \right\}^2,$$

for $k = 0, 1, 2, \ldots, K - 1$, where $I_k$'s are some subsets of $I$. When $I_1 = \cdots = I_K = I$ and $Q(\cdot, \cdot, \theta)$ is the family of neural networks, this algorithm is the neural FQI proposed by Riedmiller (2005). Fan et al. (2020) studied a variant of neural FQI by assuming $I_k$'s are disjoint and the training samples in $\cup_{k=1}^K I_k$ are independent. Using similar arguments in the proof of Theorem 4.4 in Fan et al. (2020), we can show $E \int x^2 dP_{x \in X, \theta \in A} |\hat{Q}_I(x, a) - Q^{\text{opt}}(x, a)|^2$ coverages at a rate of $O(|I|^{-2p/(2p+d)})$ up to some logarithmic factors. The conditions in Theorem 4 thus hold for any $b_\epsilon < p/(2p + d)$.

4 | EXTENSIONS TO ON-POLICY EVALUATION

We now extend our methodology in Section 3 to on-policy settings. The proposed CI is similar to that presented in Section 3.2.2 and applies to any reinforcement learning algorithms that iteratively update the estimated policy based on batches of observations. Let $\{T(k)\}_{k \geq 1}$ be a monotonically increasing sequence that diverges to infinity. At the $k$-th iteration, define $\bar{T}_k = \{(i, t) : 1 \leq i \leq n, 0 \leq t < \sum_{j=1}^k T(j)\}$. The data observed so far can be summarized as $\{(X_{i,t}, A_{i,t}, Y_{i,t}, X_{i,t+1})\}_{1 \leq i \leq n, 0 \leq t < \sum_{j=1}^k T(j)}$. We compute the estimated policy $\hat{\pi}_{\bar{T}_k}$ based on these data. Then we determine the behavior policy $\hat{b}_{\bar{T}_k}$ as a function of $\hat{\pi}_{\bar{T}_k}$ and generate new observations

$$\{(A_{i,t}, Y_{i,t}, X_{i,t+1})\}_{1 \leq i \leq n, \sum_{j=1}^k T(j) \leq t < \sum_{j=1}^{k+1} T(j)},$$

(20)

according to $\hat{b}_{\bar{T}_k}$. To balance the exploration-exploitation trade-off, a common choice of $\hat{b}_{\bar{T}_k}$ is the $\epsilon$-greedy policy with respect to $\hat{\pi}_{\bar{T}_k}$. 
Let $I_{k+1} = \{1 \leq i \leq n, \sum_{j=1}^{k} T(j) \leq t < \sum_{j=1}^{k+1} T(j)\}$. The new observations in Equation (20) are conditionally independent of $\hat{\pi}_{T_k}$ given those in $\overline{T}_k$. So the Bellman equation in (7) is valid with $\pi = \hat{\pi}_{T_k}$ for any $(i, t) \in \overline{T}_{k+1}$. We compute $\hat{V}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathbb{G})$ and $\hat{\sigma}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathbb{G})$ as in Appendix B.1 of the supplementary article, where the number of basis $L(k + 1)$ depends on both $n$ and $T(k + 1)$. We iterate this procedure for $k = 1, 2, \ldots, K - 1$. The estimated value and CI for $V(\hat{\pi}_{T_k}; \mathbb{G})$ are given by

$$
\hat{V}(\mathbb{G}) = \left\{ \sum_{k=1}^{K-1} \frac{\sqrt{T(k)}}{\hat{\sigma}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathbb{G})} \right\}^{-1} \left\{ \sum_{k=1}^{K-1} \frac{\sqrt{T(k)} \hat{V}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathbb{G})}{\hat{\sigma}_{I_{k+1}}(\hat{\pi}_{T_k}; \mathbb{G})} \right\},
$$

and

$$
\left[ \hat{V}(\mathbb{G}) - z_{a/2} \left\{ \sum_{k=2}^{K} \frac{n T(k)}{K - 1} \right\}^{-1/2} \hat{\sigma}(\mathbb{G}), \hat{V}(\mathbb{G}) + z_{a/2} \left\{ \sum_{k=2}^{K} \frac{n T(k)}{K - 1} \right\}^{-1/2} \hat{\sigma}(\mathbb{G}) \right],
$$

where $\hat{\sigma}(\mathbb{G}) = \{ \sum_{k=2}^{K} \sqrt{T(k)} \} \{ \sum_{k=2}^{K} \sqrt{T(k)} \hat{\sigma}_{k}^{-1}(\hat{\pi}_{T_{k-1}}; \mathbb{G}) \}^{-1}$. Similar to Theorem 2, we can show such a CI achieves nominal coverage under certain conditions. To save space, we provide our technical results in Section C.3 of the supplementary article.

5 | SIMULATIONS

In this section, we conduct Monte Carlo simulations to examine the finite sample performance of the proposed CI. We consider off-policy settings in Sections 5.1 and 5.2, where CIs for values under both fixed and optimal policies are reported. In Section 5.3, we report CIs computed in on-policy settings. The state vector $X_{0,t}$ in our settings might not have bounded supports. For $j = 1, \ldots, d$, we define $X_{0,t}^{(j)} = \Phi(X_{0,t}^{(j)})$ where $X_{0,t}^{(j)}$ stands for the $j$th element of $X_{0,t}$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. This gives us a transformed state vector with bounded support. The basis functions are constructed from the tensor product of $d$ one-dimensional cubic B-spline sets where knots are placed at equally spaced sample quantiles of the transformed state variables. For discrete state space $\mathbb{X} = \{x_1, \ldots, x_M\}$, we set $L = M$ and $\Phi_M(\cdot) = \{ \mathbb{I}(\cdot = x_1), \ldots, \mathbb{I}(\cdot = x_M) \}^\top$. We set the discount factor $\gamma = 0.5$ in all settings, and set $L = \lceil (nT)^\eta \rceil$ with $\eta = 3/7$. Here, for any $z \in \mathbb{R}$, $[z]$ denotes the largest integer that is smaller or equal to $z$. We tried several other values of the parameter $\eta$, and the resulting CIs are very similar and not sensitive to the choice of $\eta$. We also tried several other values of $\gamma$. Overall, the proposed CI achieves nominal coverage and performs better than other baseline methods. More details can be found in Appendix D.2 of the supplementary article.

5.1 | Off-policy evaluation with a fixed target policy

We consider three scenarios. In Scenarios (A) and (B), the system dynamics are given by
for \( t \geq 0 \), where \( \{z_t\}_{t \geq 0} \overset{iid}{\sim} N(0, I_2/4) \) and \( X_{0,0} \sim N(0, I_2) \). In Scenario (A), we consider a completely randomized study and set \( \{A_{0,t}\}_{t \geq 0} \) to i.i.d Bernoulli random variables with expectation 0.5. In Scenario (B), we allow the treatment assignment mechanism to depend on the observed state. Specifically, we set \( m = 2 \) and \( \Pr(A_{0,t} = 1|X_{0,t}) = 0.5 \text{sigmoid}(X_{0,t,1}) + 0.5 \text{sigmoid}(X_{0,t,2}) \) where \( X_{0,t,i} \) denotes the \( i \)th element in \( X_{0,t} \). The target policy we consider is designed as follows,

\[
\pi(a|x) = \begin{cases} 
0, & x_1 > 0 \text{ and } x_2 > 0; \\
1, & \text{otherwise}, 
\end{cases}
\]

where \( x_i \) denotes the \( i \)th element of \( x \). The reference distribution \( \mathcal{G} \) is set to \( N(0_2, I_2) \).

In Scenario (C), we consider a standard RL setting included in OpenAI Gym (Brockman et al., 2016): Cliff Walking. This RL example is detailed in Example 6.6 in Sutton and Barto (2018). The objective is to identify the optimal path from the starting point \( S \) to the destination point \( G \) without falling off the cliff (see Figure 1). This scenario corresponds to an episodic task where the agent will be sent instantly to the starting point wherever it steps into the cliff or arrives at the destination. We manually add some noises to the immediate rewards simulated by the OpenAI Gym to ensure that the system dynamics are not deterministic. We remark that this task is considered in Kallus and Uehara (2020) as well. The target policy is the optimal policy and the behavior policy is a 50-50 mixture of the optimal and uniform random policies.

The true value function \( V(\pi, \mathcal{G}) \) is computed by Monte Carlo approximations. Specifically, we simulate \( N = 10^5 \) independent trajectories with initial state variable distributed according to \( \mathcal{G} \). The action at each decision point is chosen according to \( \pi \). Then we approximate \( V(\pi, \mathcal{G}) \) by

\[
\sum_{j=1}^N \sum_{t=0}^{T_j-1} r_t Y_{j,t}/N \text{ where } T_j \text{ is set to 500 in Scenarios (A), (B) and the termination time of each episode in Scenario (C). The integrals in Equations (10) and (11) are computed via Monte Carlo methods. For Scenarios (A) and (B), we further consider 9 cases by setting } n = 25, 50, 100 \text{ and } T = 30, 50, 70. \text{ For Scenario (C), we consider 3 cases by setting } n = 500, 1000, 1500. \text{ Each trajectory have 13 time points on average, under the behavior policy.}
\]

\[ X_{0,t+1} = \begin{bmatrix} \frac{3}{4}(2A_{0,t} - 1) & 0 \\ 0 & \frac{3}{4}(1 - 2A_{0,t}) \end{bmatrix} X_{0,t} + z_t, \]

\[ Y_{0,t} = X_{0,t+1}^T \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \frac{1}{4}(2A_{0,t} - 1), \]

\[ X_{0,t} + z_t, \]

\[ Y_{0,t} = X_{0,t+1}^T \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \frac{1}{4}(2A_{0,t} - 1), \]

\[ \text{FIGURE 1 Illustration of Cliff Walking [Colour figure can be viewed at wileyonlinelibrary.com]} \]
The DRL estimator has been shown to be much more efficient than AIPW or IPW estimators (Jiang & Li, 2016; Thomas et al., 2015). So we focus on comparing our approach with DRL. DRL requires the calculation of the Q-function, the marginalized density ratio and the behavior policy. Here, we treat the behavior policy as known and estimate the Q-function and the density ratio based on nonparametric sieve regression.

In Figure 2 and Table 1, we report the empirical coverage probabilities (ECPs) and average lengths (ALs) of CIs constructed by the proposed method, with different choices of $n$ and $T$. It can be seen that our CI achieves nominal coverage in all cases. Its length decreases as $nT$ increases. This is consistent with our theoretical findings where we show the proposed value estimator converges at a rate of $n^{-1/2}T^{-1/2}$ under certain conditions (see the discussions below Theorem 1).

**FIGURE 2** Empirical coverage probabilities and average lengths of confidence intervals constructed by the proposed method (colored in blue) and the double reinforcement learning method (colored in red) as well as the mean-squared errors of the corresponding value estimators, with different choices of $n$ and $T$. Settings correspond to Scenario (A) and Scenario (B), from top plots to bottom plots [Colour figure can be viewed at wileyonlinelibrary.com]

**TABLE 1** Empirical coverage probabilities (ECP) and average lengths (AL) of confidence intervals (CIs) constructed by the proposed method and double reinforcement learning (DRL) as well as the mean-squared errors (MSE) of the corresponding value estimators under Scenario (C)

| $n$  | SAVE  | DRL   |
|------|-------|-------|
|      | ECP   | AL    | log(MSE) | ECP | AL    | log(MSE) |
| 500  | 0.96  | 0.11  | −8.11    | 0.89 | 0.13  | −6.50    |
| 1000 | 0.93  | 0.08  | −7.60    | 0.85 | 0.09  | −7.26    |
| 1500 | 0.96  | 0.06  | −8.11    | 0.87 | 0.07  | −7.82    |
Comparing our method with DRL, it is clear that our CIs are in general narrower than those constructed by DRL. In addition, MSEs of the proposed value estimates are smaller than those based on DRL. This is consistent with our theoretical analysis in Appendix C.2 where we show the variance of our value estimator is strictly smaller than that based on DRL under certain conditions. In addition, it can be seen from Table 1 that ECPs of DRL are below 90% in Scenario (C).

In Appendix D.3, we conduct some additional simulation studies under Scenario (A) by setting the reference distribution $\mathcal{G}$ to a Dirac measure. The proposed CI achieves nominal coverage under these settings as well.

5.2 Off-policy evaluation with an (estimated) optimal policy

In this section, we focus on constructing the CI for value under an optimal policy. Specifically, we use a version of fitted Q-iteration (double FQI) to compute the estimated optimal policy. Detailed algorithm can be found in Section B.3 of the supplementary article. To implement the proposed CI in Section 3.2, we set $K_n = 2$, $K_T = 2$ in Scenarios (A), (B) and $K_n = 3$, $K_T = 1$ in Scenario (C). To evaluate our CI, we generate a very large dataset to compute an estimated optimal policy $\hat{\pi}^*$ based on double FQI and use the Monte Carlo methods described in Section 5.1 to evaluate its value $V(\hat{\pi}^*; \mathcal{G})$. Then we treat $V(\hat{\pi}^*; \mathcal{G})$ as the true optimal value $V(\pi^*; \mathcal{G})$. We consider the same three scenarios detailed in Section 5.1. For Scenarios (A) and (B), we fix $\gamma = 0.5$ and consider 6 cases by setting $n = 100, 200$ and $T = 60, 100, 140$. For Scenario (C), we consider 4 cases by setting $n = 3000, 4500$ and $\gamma = 0.5, 0.7$. The ECP, AL and MSE of our CI are reported in the top two

![FIGURE 3](wileyonlinelibrary.com)
panels of Figure 3 in Table 2. It can be seen that these ECPs are close to the nominal level in most cases. ALs and MSEs decay as either $n$ or $T$ increases.

In addition, we design a non-regular setting Scenario (D) where the actions do not have effects on the transition dynamics or the immediate rewards. Specifically, for any $t \geq 0$, we set

$$X_{0,t+1} = \begin{bmatrix} \frac{3}{4} & 0 \\ 0 & \frac{3}{4} \end{bmatrix} X_{0,t} + z_t,$$

and

$$Y_{0,t} = X_{0,t+1}^T \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$

where \(\{z_t\}_{t \geq 0} \sim \text{iid } N(0, I_2/4)\) and \(\{A_{0,t}\}_{t \geq 0}\) are i.i.d Bernoulli random variables with expectation 0.5. Under this setup, any policy will achieve the same value function. As a result, the optimal policy is not unique. We consider the same reference distribution $G$, and the same combinations of $n$ and $T$ as in the regular setting. ECPs and ALs of the proposed CIs are plotted in the bottom panels of Figure 3. It can be seen that our CIs achieve nominal coverage in the non-regular setting as well.

In Appendix D.3, we conduct some additional simulation studies under Scenario (A) by setting the reference distribution $G$ to a Dirac measure. Findings are very similar to those in cases where $G = N(0, I_2)$.

### 5.3 On-policy evaluation with an (estimated) optimal policy

We consider a setting where the transition dynamics and immediate rewards are defined by Equation (21). In the first block of data \(\{(X_{i,t}, A_{i,t}, Y_{i,t}, X_{i,t+1})\}_{0 \leq t < T(1), 1 \leq i \leq n}\) the actions are generated according to i.i.d Bernoulli random variables with expectation 0.5. For $k = 2, \ldots, K$, we use double FQI to estimate the optimal policy based on the data observed so far \(\{(X_{i,t}, A_{i,t}, Y_{i,t}, X_{i,t+1})\}_{0 \leq t < T(k-1), 1 \leq i \leq n}\) and use an $\epsilon$-greedy method to generate actions in the next block of data. In our experiments, we set $\epsilon = 0.2$, $K = 4$ and $T(1) = T(2) = \cdots = T(K) = T$. We fix $n = 25$ and consider three choices of $T$, corresponding to $T = 120$, 200 and 280. We consider three choices of $G$, corresponding to $G = \delta_{(0.5, 0.5)}$, $\delta_{(-0.5, -0.5)}$ and $N(0, I_2)$. The true optimal value function is approximated by Monte Carlo methods, as in off-policy settings. ECPs and ALs of the

| $\gamma$ | $n = 3000$ | $n = 4500$ |
|---------|------------|------------|
|         | ECP | AL | log(MSE) | ECP | AL | log(MSE) |
| 0.5     | 0.94 | 0.12 | $-6.91$ | 0.94 | 0.10 | $-7.51$ |
| 0.7     | 0.95 | 0.23 | $-5.81$ | 0.96 | 0.19 | $-6.35$ |
proposed CIs are reported in Table 3. It can be seen that ECPs are close to the nominal level in almost all cases and ALs decrease as $T$ increases.

### Table 3  Empirical coverage probabilities and average lengths of confidence intervals (CIs) constructed in on-policy settings with different choices of $T$ and $G$

|           | ECPs   | ALs   |
|-----------|--------|-------|
|           | $T = 120$ | $200$ | $280$ | $T = 120$ | $200$ | $280$ |
| $G = \delta_{(0.5,0.5)}$ | 0.947 | 0.914 | 0.931 | 0.51 | 0.46 | 0.38 |
| $G = \delta_{(0.5,0.5)}$ | 0.925 | 0.942 | 0.940 | 0.72 | 0.59 | 0.48 |
| $G = N(0_2, I_2)$ | 0.914 | 0.926 | 0.948 | 0.29 | 0.21 | 0.17 |

### 6  Application to the OhioT1DM Dataset

As commented in the introduction, this dataset contains eight weeks’ records of CGM blood glucose levels, insulin doses and self-reported life-event data for each of six patients with type 1 diabetes. To analyse this data, we divide these eight weeks into 3-h intervals. The state variable $X_{i,t}$ is set to be a three-dimensional vector. Specifically, its first element $X_{i,t}^{(1)}$ is the average CGM blood glucose levels during the 3-h interval $[t-1, t)$. The second covariate $X_{i,t}^{(2)}$ is constructed based on the $i$-patient’s self-reported time and the carbohydrate estimate for the meal. Suppose the patient has meals at time $t_1$, $t_2$, ..., $t_N \in [t - 1, t)$ with the carbohydrate estimates $CE_1, CE_2, ..., CE_N$. Define

$$X_{i,t}^{(2)} = \sum_{j=1}^{N} CE_j \gamma_c^{36(t_j - t + 1)},$$

where $\gamma_c$ corresponds to the decay rate every 5 min. Here, we set $\gamma_c = 0.5$. The third covariate $X_{i,t}^{(3)}$ is defined as an average of the basal rate during the 3-h interval.

We discretize the action according to the amount of insulin injected in the 3-h interval. Specifically, $A_{i,t} = 1$ when the total amount of insulin delivered to the $i$th patient is greater than one unit. Otherwise, we set $A_{i,t} = 0$. The immediate reward $Y_{i,t}$ is defined according to the Index of Glycemic Control (IGC, Rodbard, 2009), which is a non-linear function of the blood glucose levels. Specifically, we set

$$Y_{i,t} = \begin{cases} 
  -\frac{1}{30} (80 - X_{i,t+1}^{(1)})^2, & X_{i,t+1}^{(1)} < 80; \\
  0, & 80 \leq X_{i,t+1}^{(1)} < 140; \\
  -\frac{1}{30} (X_{i,t+1}^{(1)} - 140)^{1.35}, & 140 \leq X_{i,t+1}^{(1)}. 
\end{cases}$$

A large IGC indicates the patient is in good health status. We set the discount factor $\gamma = 0.5$, as in simulations.

For the $i$th patient, we apply the double FQI algorithm to the data

$$\{(X_{i,t}, A_{i,t}, Y_{i,t}, X_{i,t+1})\}_{0 \leq t < T},$$
to estimate a patient-specific optimal policy. Then we compute the estimator for the value function $V(\pi^{\text{opt}}; X_{i,0})$ starting from the initial state variable $X_{i,0}$. In addition, we extend our methodology in Section 5.2 to construct the confidence interval for the value difference $V(\pi^{\text{opt}}; X_{i,0}) - V(b; X_{i,0})$ where $V(b; X_{i,0})$ corresponds to the value under the behavior policy. See Appendix B.2 for details.

In Figure 4, we plot our proposed CI for the value difference, for each of the six patients, when the initial starting time is either 8:00 am or 2:00 pm in Day 1. It can be seen that the estimated value differences are strictly positive in all cases. This implies that the optimal value is strictly larger than the observed discounted cumulative reward. In some cases, the lower bound of our CI is larger than zero. The difference is thus significant. In Figure 5, we fix the starting time to 8:00 am and plot the

**Figure 4** Confidence interval of the value difference between an estimated optimal policy and the behavior policy for each of the six patients, with $\gamma = 0.5$ [Colour figure can be viewed at wileyonlinelibrary.com]

**Figure 5** Confidence intervals of the value difference between an estimated optimal policy and the behavior policy for each of the six patients, with different $\gamma$ [Colour figure can be viewed at wileyonlinelibrary.com]
CI of the value difference with different $\gamma$. Results show a similar qualitative pattern. This suggests applying reinforcement learning algorithms could potentially improve some patients’ health status.

7 | DISCUSSION

7.1 | Comparison between DRL

We discuss the advantages and limitations among the proposed method and DRL when inferring the value under a fixed decision rule. Generally speaking, the proposed method results in narrower CIs and would be preferred in cases where $(nT)^{-1/2}$-consistent estimation of the Q-function is feasible. This includes settings where the dimension of the state-vector is not large, as in our real data applications. In contrast, DRL would be preferred in ergodic environments with high-dimensional covariates where $(nT)^{-1/2}$-consistent estimation of the Q-function is infeasible.

Specifically, in Appendix C.2, we consider settings where both the behavior policy and the target policy are nondynamic. Under certain conditions, we prove that the variance of the DRL estimator is strictly larger than that of the proposed estimator. This in turn implies that our method yields a narrower CI in general.

In addition, we remark that the CI constructed by DRL requires the data to be generated from an ergodic environment. In the Cliff Walking example, the data are generated by a mixture of the optimal and random policy. Since the agent will be sent instantly to the starting point wherever it steps into the cliff or arrives at the destination, the Markov chain formed by the state-action pair is no longer ergodic. It can be seen from Table 1 where ECP of the CI is well below the nominal level in the Cliff Walking example. Although our procedure also requires the ergodicity assumption (see Condition (A3)(ii)), this assumption is not necessary. It can be seen from the proof of Theorem 1 that our CI is valid as long as the random matrix $\hat{\Sigma}_\pi$ stabilizes. This is consistent with the findings in Table 1 where our CI achieves nominal coverage in all cases.

However, to ensure the proposed CI is valid, we require the bias of our Q-estimator to decay at a rate of $o((nT)^{-1/2})$. Consequently, our estimator converges at a rate of $(nT)^{-1/2}$. This rate might not be achievable in high dimensions. In contrast, DRL requires a weaker condition. The CI based on DRL is valid when both the Q-estimator and the estimated marginalized density ratio converge at a rate of $o((nT)^{-1/4})$.

Another potential limitation of our method is that in cases where $\hat{\Sigma}_\pi$ is close to a singular matrix, the resulting Q-estimator might suffer from over-fitting, leading to an unbounded outcome. In practice, we could add a ridge penalty to reduce over-fitting. We discuss in detail in Appendix D.4.

7.2 | Number of basis functions

We outline a procedure to choose the number of basis function $L$ in this section. The idea is to simulate the model dynamics and select $L$ such that the resulting confidence interval achieves nominal coverage under the simulated model. Specifically, given the observed data \{$(X_{i,t}, A_{i,t}, Y_{i,t}, X_{i,t+1})\}_{0 \leq t < T_i, 1 \leq i \leq n}$ we propose to learn the conditional density function of $(Y_{i,t}, X_{i,t+1})$ given $(A_{i,t}, X_{i,t})$. Following Janner et al. (2019), we recommend to use a Gaussian distribution to model the conditional density function in practice,

$$(Y_{i,t}, X_{i,t+1})|(A_{i,t}, X_{i,t}) = (a, x) \sim N(\mu(a, x), \Sigma(a, x)). \quad (22)$$
The conditional mean $\mu$ can be estimated via nonparametric regression (e.g. random forest). Let $\hat{\mu}$ denote the corresponding estimator. Given the set of estimated residuals $(Y_{i,t}, X_{i,t+1})^T - \hat{\mu}(A_{i,t}, X_{i,t})$, the conditional covariance function $\Sigma$ can be estimated via regression as well. We remark that in addition to the Gaussian function, other density functions could be used to model the system dynamics as well.

The behavior policy $b$ can be similarly estimated via regression. Given an estimated behavior policy $\hat{b}$ and $\hat{\Sigma}$, we generate simulated trajectories to investigate the performance of the proposed CI with different choices of $L$.

Finally, we choose $L$ such that the resulting CI is the shortest among all CIs whose coverage probabilities are above certain level (e.g. 93%) under the simulated environment.

In Appendix D.1, we investigate the finite sample performance of such a method and find that it performs reasonably well. We remark that alternative to the aforementioned method, cross-validation could be applied to select $L$.

### 7.3 Sensitivity to the ordering of trajectories

The proposed sequential value evaluation procedure in Section 3.2 divides the data into blocks defined both by trajectories and by time. While there is a natural order in time, there does not appear to be a natural order in the trajectories. In Appendix D.2.3, we conduct additional simulation studies to investigate the sensitivity of our CI to the ordering of trajectories under Scenario (B). Results suggest that our CI is not overly sensitive under our simulation setting.

As suggested by one of the referees, we may aggregate CIs over multiple orderings in cases where the results depend strongly on the ordering of the trajectories. Dezeure et al. (2015) derived a CI for the regression coefficients in high-dimensional models by aggregating results over multiple sample splits using a quantile function. We can adopt their method to aggregate our CIs over multiple orderings. Alternatively, one may average the value estimates over sufficiently many orderings and apply similar methods developed in Wang et al. (2020); Shi et al. (2020) to derive the CI. However, these algorithms are much more time-consuming.

### 7.4 More on value-based method

In Section 3.1.1, we discuss a potential drawback of using nonparametric methods to directly model the value function. We remark that a kernel-type importance sampling estimator for the $V(\pi; x)$ will not suffer from this issue, since it does not directly model the value function, but uses an inverse propensity-score weighted method instead. Both IPW and regression type estimators have their own merits. In general, IPWEs might suffer from a large variance whereas regression-based estimators might suffer from a large bias. There exist methods that combine both for more robust off-policy evaluation (see e.g. Kallus & Uehara, 2019; Shi et al., 2021; Tang et al., 2020; Uehara et al., 2020). However, as commented in Section 7.1, they might yield larger CIs compared to our method.
7.5  Rate of convergence of Q-learning type algorithms

Through authors’ communication, we found a recent independent work by Hu et al. (2021) that derived a nonasymptotic error bound on the value of the estimated optimal policy computed by Q-learning type algorithms under the margin condition. Their results are consistent with our theoretical findings in Theorems 3 and 4 that show the value of the estimated optimal policy converges to the optimal value at a faster rate than the estimated Q-function.

ACKNOWLEDGEMENTS

Shi’s research was partially supported by the LSE New Research Support Fund. Song’s research was partially supported by the National Science Foundation (DMS-1555244 and DMS-2113637). The authors thank the AE, and the reviewers for their constructive comments, which have led to a significant improvement of the earlier version of this article.

ORCID

Rui Song  https://orcid.org/0000-0003-1875-2115

REFERENCES

Audibert, J.-Y. & Tsybakov, A.B. (2007) Fast learning rates for plug-in classifiers. The Annals of Statistics, 35, 608–633.
Bradley, R.C. (2005) Basic properties of strong mixing conditions. A survey and some open questions. Probability Surveys, 2, 107–144. Update of, and a supplement to, the 1986 original.
Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J. et al. (2016) Openai gym. arXiv preprint arXiv:1606.01540.
Chen, X. & Christensen, T.M. (2015) Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. Journal of Econometrics, 188, 447–465.
Dezeure, R., Bühlmann, P., Meier, L. & Meinshausen, N. (2015) High-dimensional inference: confidence intervals, p-values and R-software hdi. Statistical Science, 30(4), 533–558.
Ernst, D., Geurts, P. & Wehenkel, L. (2005) Tree-based batch mode reinforcement learning. Journal of Machine Learning Research, 6, 503–556.
Ertefaie, A. & Strawderman, R.L. (2018) Constructing dynamic treatment regimes over indefinite time horizons. Biometrika, 105, 963–977.
Fan, J., Wang, Z., Xie, Y. & Yang, Z. PMLR workshop paper. http://proceedings.mlr.press/v120/yang20a.html
Hu, Y., Kallus, N. & Uehara, M. (2021) Fast rates for the regret of offline reinforcement learning. arXiv preprint arXiv:2102.00479.
Huang, J.Z. (1998) Projection estimation in multiple regression with application to functional ANOVA models. The Annals of Statistics, 26, 242–272.
Janner, M., Fu, J., Zhang, M. & Levine, S. (2019) When to trust your model: model-based policy optimization. In: Advances in neural information processing systems, pp. 12519–12530.
Jiang, N. & Li, L. (2016) Doubly robust off-policy value evaluation for reinforcement learning. In: International conference on machine learning, pp. 652–661.
Jin, J., Song, C., Li, H., Gai, K., Wang, J. & Zhang, W. (2018) Real-time bidding with multi-agent reinforcement learning in display advertising. In: Proceedings of the 27th ACM international conference on information and knowledge management, pp. 2193–2201. ACM.
Kallus, N. & Uehara, M. (2019) Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. arXiv preprint arXiv:1909.05850.
Kallus, N. & Uehara, M. (2020) Double reinforcement learning for efficient off-policy evaluation in markov decision processes. Journal of Machine Learning Research, 21, 1–63.
Kormushev, P., Calinon, S. & Caldwell, D. (2013) Reinforcement learning in robotics: applications and real-world challenges. Robotics, 2, 122–148.
Luckett, D.J., Laber, E.B., Kahkoska, A.R., Maahs, D.M., Mayer-Davis, E. & Kosorok, M.R. (2019) Estimating dynamic treatment regimes in mobile health using V-learning. *Journal of the American Statistical Association*, 115(530), 692–706. https://doi.org/10.1080/01621459.2018.1537919

Luedtke, A.R. & van der Laan, M.J. (2016) Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44, 713–742.

Maei, H.R., Szepesvári, C., Bhatnagar, S. & Sutton, R.S. (2010) Toward off-policy learning control with function approximation. In: ICML, pp. 719–726.

Marling, C. & Bunescu, R.C. (2018) The OhioT1DM dataset for blood glucose level prediction. In: *KHD@ IJCAI*, pp. 60–63.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G. et al. (2015) Human-level control through deep reinforcement learning. *Nature*, 518, 529.

Murphy, S.A. (2003) Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 331–366.

Puterman, M.L. (1994) *Markov decision processes: discrete stochastic dynamic programming*. Wiley series in probability and mathematical statistics: applied probability and statistics. New York: John Wiley & Sons, Inc. A Wiley-Interscience Publication.

Qian, M. & Murphy, S.A. (2011) Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39, 1180–1210.

Riedmiller, M. (2005) Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In: *European conference on machine learning*, Springer, pp. 317–328.

Rodbard, D. (2009) Interpretation of continuous glucose monitoring data: glycemic variability and quality of glycemic control. *Diabetes Technology & Therapeutics*, 11, S-55.

Saikkonen, P. (2001) Stability results for nonlinear vector autoregressions with an application to a nonlinear error correction model. *Technical report*, Discussion Papers, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes.

Shi, C., Fan, A., Song, R. & Lu, W. (2018a) High-dimensional A-learning for optimal dynamic treatment regimes. *The Annals of Statistics*, 46, 925–957.

Shi, C., Song, R., Lu, W. & Fu, B. (2018b) Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 681–702.

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

Shi, C., Wan, R., Chernozhukov, V. & Song, R. (2021) Deeply-debiased off-policy interval estimation. In: *International conference on machine learning*, accepted.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G. et al. (2016) Mastering the game of go with deep neural networks and tree search. *Nature*, 529, 484.

Sutton, R.S. & Barto, A.G. (2018) *Reinforcement learning: an introduction*, 2nd edn. Adaptive computation and machine learning. Cambridge, MA: MIT Press.

Tang, Z., Feng, Y., Li, L., Zhou, D. & Liu, Q. (2020) Doubly robust bias reduction in infinite horizon off-policy estimation. In: *International conference on learning representations*.

Thomas, P.S., Theocharous, G. & Ghavamzadeh, M. (2015) High-confidence off-policy evaluation. In: *Twenty-Ninth AAAI conference on artificial intelligence*.

Tsao, C.W. & Vasan, R.S. (2015) Cohort profile: the Famingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *International Journal of Epidemiology*, 44, 1800–1813.

Tsybakov, A.B. (2004) Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32, 135–166.

Uehara, M., Huang, J. & Jiang, N. (2020) Minimax weight and Q-function learning for off-policy evaluation. In: *International conference on machine learning*.

Wang, J., He, X. & Xu, G. (2020) Debiased inference on treatment effect in a high-dimensional model. *Journal of the American Statistical Association*, 115, 442–454.

Xu, Z., Li, Z., Guan, Q., Zhang, D., Li, Q., Nan, J. et al. (2018) Large-scale order dispatch in on-demand ride-hailing platforms: a learning and planning approach. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 905–913. ACM.
A. SOME TECHNICAL CONDITIONS

A.1. More on Conditions A3

Define $P^t_X$ as the $t$-step transition kernel, that is, $P^t_X(B|x) = \Pr(X_{0:t} \in B | X_{0,0} = x)$. Geometric ergodicity implies that there exists some function $M(\cdot)$ on $\mathcal{X}$ and some constant $\rho < 1$ such that
\[
\int_{x \in \mathcal{X}} M(x) \mu(x) \, dx < +\infty
\]
and
\[
\| P^t_X(\cdot|x) - \mu(\cdot) \|_{TV} \leq M(x) \rho^t, \quad \forall t \geq 0,
\]
where $\| \cdot \|_{TV}$ denotes the total variation norm.

A.2. Conditions A3*

We present the technical condition (A3*) below. We assume the estimated policy satisfies $\hat{\pi}_I \in \Pi$ with probability 1, for any $I$. For example, if Q-learning type algorithms are used and we approximate the optimal Q-function based on a linear model $\Phi^\top(x, a) \beta$ with some basis function $\Phi(x, a) \in \mathbb{R}^M$. Then for any $\beta$, we can define a policy $\pi(\beta)$ as follows:

\[
\pi_\beta(a|x) = \begin{cases} 
1, & \text{if } a = \text{sarg max}_{a' \in A} \Phi^\top(x, a') \beta, \\
0, & \text{otherwise}.
\end{cases}
\]

Then we have $\Pi = \{ \pi_\beta : \beta \in \mathbb{R}^M \}$.

(A3*.) Assume (i) and (ii) hold if $T \to \infty$ and (i) holds if $T$ is bounded.
(i) \( \inf_{\pi \in \Pi} \lambda_{\min} \left[ \sum_{t=0}^{T-1} \mathbb{E} \left\{ \xi_{0,t}^T g_{0,t}^T - \gamma^2 u_\pi (X_{0,t}, A_{0,t}) u_\pi^T (X_{0,t}, A_{0,t}) \right\} \right] \geq \bar{c}T \) for some constant \( \bar{c} > 0 \).

(ii) The Markov chain \( \{X_{0,t}\}_{t \geq 0} \) is geometrically ergodic.

We remark that Condition A3*(ii) is the same as A3(ii).

A.3. More on the margin condition

To better understand Condition A5, we consider a simple scenario where \( \mathcal{A} = \{0, 1\} \). Define \( \tau(x) = Q^\text{opt}(x, 1) - Q^\text{opt}(x, 0) \). It follows that

\[
\max_a Q^\text{opt}(x, a) - \max_{a' \in \mathcal{A} - \arg \max_a Q^\text{opt}(x, a)} Q^\text{opt}(x, a') = \begin{cases} 
|\tau(x)|, & \text{if } \tau(x) \neq 0, \\
+\infty, & \text{otherwise}.
\end{cases}
\]

As a result, Equations (18) and (19) are equivalent to the followings:

\[
\lambda \left\{ x \in \mathbb{X} : 0 < |\tau(x)| \leq \varepsilon \right\} = O(\varepsilon^\alpha), \quad (A1)
\]
\[
G \left\{ x \in \mathbb{X} : 0 < |\tau(x)| \leq \varepsilon \right\} = O(\varepsilon^\alpha). \quad (A2)
\]

Apparently, these two conditions hold when \( \inf_{x \in \mathbb{X}} |\tau(x)| > 0 \). They are satisfied in many other cases. For example, let \( d = 1 \). Consider

\[
\tau(x) = \begin{cases} 
x^{1/\alpha}, & \text{if } x > 0, \\
0, & \text{otherwise},
\end{cases}
\]

for some \( \alpha > 0 \). Then, with some calculations, we can show

\[
\lambda \left\{ x \in \mathbb{X} : 0 < |\tau(x)| \leq \varepsilon \right\} \leq \lambda \{ x : 0 < x < \varepsilon^\alpha \} = \varepsilon^\alpha.
\]

This verifies Equation (A1). When \( G \) has a bounded density function on \( \mathbb{X} \), Equation (A2) is reduced to (A1). If \( G(\cdot) \) equals the Dirac measure \( \delta_{\varepsilon}(\cdot) \), then Equation (A2) automatically holds for any \( \alpha > 0 \).