Stat

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Generalized Fitted Q-Iteration With Clustered Data

Liyuan Hu[1] | Jitao Wang[2] | Zhenke Wu[2,3] | Chengchun Shi[1]

[1]Department of Statistics, London School of Economics and Political Science, London, UK | [2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA | [3]Michigan Institute for Data and AI in Society (MIDAS), University of Michigan, Ann Arbor, Michigan, USA

Correspondence: Chengchun Shi (c.shi7@lse.ac.uk)

Received: 31 May 2025 | **Revised:** 17 September 2025 | **Accepted:** 2 October 2025

Keywords: clustered data | generalized estimating equations | reinforcement learning

ABSTRACT

This paper focuses on reinforcement learning (RL) with clustered data, which is commonly encountered in healthcare applications. We propose a generalized fitted Q-iteration (FQI) algorithm that incorporates generalized estimating equations into policy learning to handle the intra-cluster correlations. Theoretically, we demonstrate (i) the optimalities of our Q-function and policy estimators when the correlation structure is correctly specified and (ii) their consistencies when the structure is mis-specified. Empirically, through simulations and analyses of a mobile health dataset, we find the proposed generalized FQI achieves, on average, a half reduction in regret compared to the standard FQI.

1 | Introduction

Reinforcement learning (RL) has emerged as a powerful machine learning tool for sequential decision-making across various fields. This paper studies RL with cluster-structured data, where subjects are grouped into clusters, with possible correlations among their observations. Our primary motivating application is the Intern Health Study (IHS, NeCamp et al. 2020), which studied the time-varying causal effect of health-related mobile app prompts on the well-being of first-year medical interns across various institutions in the United States, paving the way for developing more effective and personalized messages tailored for each intern at every decision point. These interns naturally form clusters by their memberships to training institutions. Even after controlling for individual-level observed information such as baseline surveys, subjects in the same institution may share unobserved or imperfectly measured factors, for example, curricula and training environment, and social and peer interactions, resulting in observed intra-institution correlations.

To illustrate these correlations, we regress the cubic root of interns' daily step counts on the message interventions they received, along with their time-varying observations. We then calculate the resulting residuals and visualize the random intercepts fitted from linear mixed effect models for institutions with at least five individuals in the left panel of Figure 1, where we observe significantly different random effects across a large number of clusters (institutions), indicating the presence of correlations within these clusters.

The primary challenge of applying RL to clustered data lies in the presence of intra-cluster correlations. Existing RL algorithms typically assume that the observations are independent and identically distributed (i.i.d., see, e.g., J. Chen and Jiang 2019; Fan et al. 2020) and ignore these dependencies, resulting in suboptimal policy learning. To elaborate the challenges posed by the intra-cluster correlations in policy learning, we provide a numerical example using highly correlated data and visualize the cumulative reward of the estimated optimal policies computed by the traditional fitted Q-iteration (FQI, see, e.g., Ernst et al. 2005; Riedmiller 2005) algorithm and our proposed algorithm (denoted by GFQI) in the right panel of Figure 1. In this example, the 'oracle' optimal policy achieves a cumulative reward of 0.275. It can be seen that there is a constant gap between this optimal value and the cumulative reward achieved by the standard FQI. In contrast, the cumulative reward obtained by GFQI, which effectively handles intra-cluster correlations, closely approximates the optimal value.

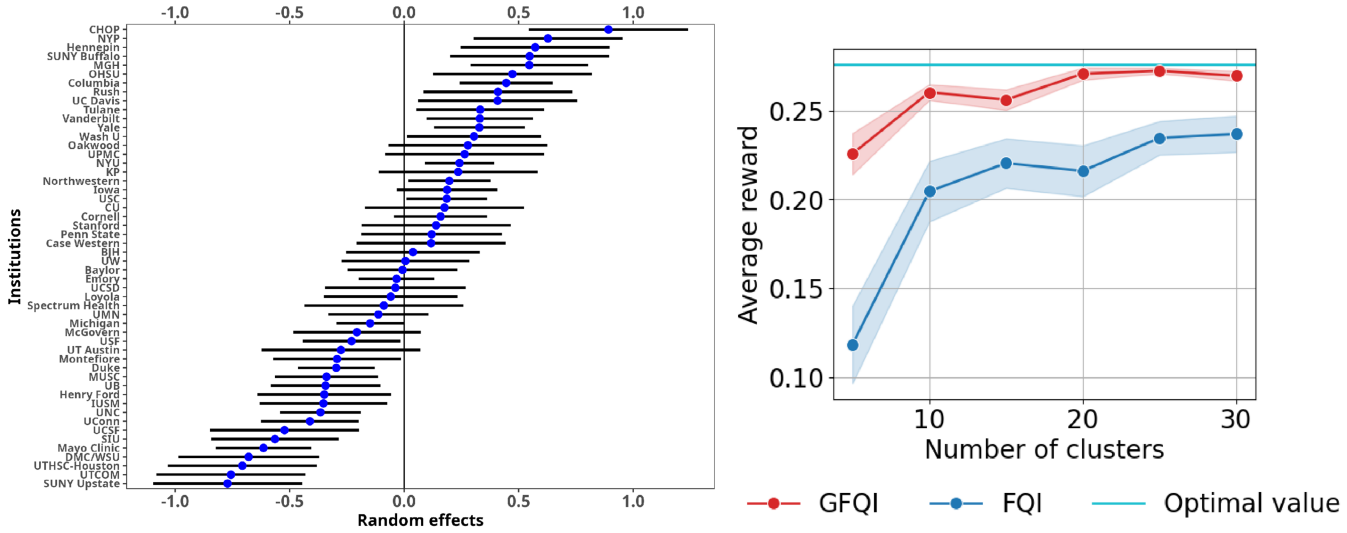

FIGURE 1 | Left panel: caterpillar plot of random effects for institutions with at least five interns. The error bars indicate 95% confidence intervals. Right panel: average reward of policies computed by standard FQI (coloured in blue) and the proposed generalized FQI (coloured in red) with increasing number of clusters. The horizontal line (coloured in cyan) depicts the optimal value, computed by an online deep Q-network (DQN) agent with sufficiently many data. Both the number of subjects per cluster and the time horizon are fixed to 5.

This work makes several contributions:

1. **Methodologically**, we propose a novel RL algorithm that incorporates generalized estimating equations (GEEs)—a classical statistical tool commonly used in analysing longitudinal and clustered data (K. Liang and Zeger 1986; Zeger and Liang 1986)—into the FQI for more effective policy learning in the presence of intra-cluster correlations. To our knowledge, this is the first study to explore RL with clustered data that frequently arises in fields such as healthcare, education and social sciences where subjects are commonly organized in clusters.
2. **Theoretically**, we establish the robustness and efficiency of the proposed generalized FQI (referred to as GFQI) algorithm. In particular, we demonstrate that (i) when the modelled correlation structure is correctly specified, the proposed estimator achieves certain optimality properties; (ii) even when this correlation structure is mis-specified, the proposed estimator maintains consistency, ensuring its robustness (Overall and Tonidandel 2004). These properties apply to both the Q-function estimator and the estimated optimal policy.
3. **Empirically**, we demonstrate the usefulness of our proposal via simulations and real-world data-based semi-synthetic analyses. Our results show that the generalized FQI significantly outperforms the standard correlation-unaware FQI, achieving an approximately 10% reduction in regret under weak intra-cluster correlations and a considerable 80% reduction under strong correlations.

2 | Related Works

We review GEE and RL in this section, as they are closely related to our proposal.

2.1 | GEEs

GEE is a classical statistical tool used for analysing longitudinal and clustered data. It estimates the parameters of interest by solving a set of estimating equations that specify the conditional mean and variance of a response, while accounting for intra-cluster correlations via a 'working' correlation structure that is allowed to be mis-specified.

More specifically, suppose we have n clusters, each containing multiple subjects. Let $\left(X_i^{(j)}, Y_i^{(j)}\right)$ denote the predictor-response pair from the jth subject within the ith cluster. We employ a generalized linear model (GLM, McCullagh 2019) to model the relationship between $Y_i^{(j)}$ and $X_i^{(j)}$. Specifically, the conditional mean of $Y_i^{(j)}$, denoted by $\mu_i^{(j)} = \mathbb{E}\left(Y_i^{(j)} \mid X_i^{(j)}\right)$, is modelled through a link function g, such that $g\left(\mu_i^{(j)}\right) = \beta^\top X_i^{(j)}$, where β is a vector of regression coefficients.

Different from standard GLMs, the responses $Y_i^{(j)}$'s within each cluster are allowed to be correlated, with their correlation structure modelled by a working correlation matrix $\boldsymbol{C}_i$. Let $\boldsymbol{B}_i$ denote a diagonal matrix whose jth diagonal element is given by the conditional variance $\mathrm{Var}\left(Y_i^{(j)} \mid X_i^{(j)}\right)$; GEE solves the following estimating equations to estimate β:

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \beta} \boldsymbol{B}_i^{-1/2} \boldsymbol{C}_i^{-1} \boldsymbol{B}_i^{-1/2} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i\right) = 0, \tag{1}$$

where we use boldface letters $\boldsymbol{\mu}_i$ and $\boldsymbol{Y}_i$ to denote vectors $\left\{\mu_i^{(j)}\right\}_j$ and $\left\{Y_i^{(j)}\right\}_j$ within the ith cluster, respectively. When setting $\boldsymbol{C}_i$ to the identity matrix, (1) is reduced to the estimating equations under the i.i.d. assumption.

GEE has several defining features (K. Liang and Zeger 1986):

1. Unlike maximum likelihood estimators (MLEs) that require to specify the joint data distribution, GEE only specifies the first two moments.
2. When the working correlation structure is correctly specified, the resulting estimator achieves a smaller asymptotic variance compared to MLEs obtained under the i.i.d. assumption.
3. Even if the correlation structure is mis-specified, GEE still produces consistent and asymptotically normal estimators.

These features make GEE particularly appealing for handling correlated data that frequently arises in biostatistics (P. Diggle et al. 2002; Fitzmaurice et al. 2008; Hedeker Donald 2006). Over time, GEE has been extended to accommodate missing data (Copas and Seaman 2010; P. Diggle et al. 2007; Roderick Little 2019; Rubin 1976; C. Shen and Chen 2012, 2013), high-dimensional data (L. Wang et al. 2011), and other complex data structures (Friedel et al. 2019; Lipsitz et al. 1994; J. Liu et al. 2009; Paradis and Claude 2002; Y. Wang et al. 2024).

2.2 | RL

RL has its roots in experimental psychology but has evolved into a powerful AI tool for optimal policy learning through interactions with a given environment (Sutton and Barto 2018). Over the past decade, it has become one of the most popular frontiers in machine learning, with a number of successful applications across diverse fields including game playing (Mnih et al. 2015; Silver et al. 2016), ride-sharing (Qin et al. 2025; Tang et al. 2019; Xu et al. 2018; Shi, Wang, et al. 2023), real-time bidding (Cai et al. 2017; J. Jin et al. 2018), finance (Carta et al. 2021; X. Liu et al. 2020; Wu et al. 2020; H. Yang et al. 2020) and large language modelling (Ouyang et al. 2022; Shao et al. 2024). Numerous algorithms have been proposed in the machine learning literature, ranging from the classical Q-learning (Watkins and Dayan 1992) and actor-critic (Konda and Tsitsiklis 1999) to more advanced trust region policy optimization (Schulman et al. 2015), deep Q-network (DQN, Mnih et al. 2015), quantile DQN (Dabney et al. 2018) and more recent offline RL algorithms (see, e.g., Levine et al. 2020; Uehara, Shi, et al. 2022, for reviews).

In the statistics literature, RL is closely related to a large body of works on learning dynamic treatment regimes (DTRs, see Chakraborty and Moodie 2013; Kosorok and Laber 2019; Kosorok and Moodie 2015; Z. Li et al. 2023; Tsiatis et al. 2019, for reviews). More recent works have studied Markov decision processes (MDPs, Puterman 2014) and their variations (e.g., partially observable MPDs, Krishnamurthy 2016), which can broadly be categorized into three types: (i) The first type of methods focuses on the estimation of the optimal policy (X. Chen et al. 2023; Ertefaie and Strawderman 2018; G. Li et al. 2024; Liao et al. 2022; Luckett et al. 2020; Miao et al. 2025; G. Shen et al. 2025; Shi, Luo, et al. 2024; Shi, Qi, et al. 2024; J. Wang et al. 2025; W. Yang et al. 2022; Zhong et al. 2020; Zhou et al. 2024; Zhu et al. 2025); (ii) the second type of methods aims to evaluate the return of a target policy (Bian et al. 2025; Duan et al. 2024; Y. Hu and Wager 2023; Kallus and Uehara 2024; T. Liang and Recht 2025; Liao et al. 2021; W. Liu et al. 2023; Luo et al. 2024; Qi et al. 2025; Ramprasad et al. 2023; Shi, Wan, et al. 2023; Shi et al. 2022; Shi, Zhu, et al. 2024; Sun et al. 2024; J. Wang et al. 2023; Zhang et al. 2023; Zhou et al. 2023); (iii) the last type of methods considers other related tasks such as variable selection, clustering, order detection or stationarity testing to enable more sample efficient policy learning and/or evaluation (Andersen and Zhao 2025; E. Chen et al. 2024; L. Hu et al. 2022; M. Li et al. 2025; Ma et al. 2023).

Among all the aforementioned works, our proposal is particularly related to the following two branches of RL algorithms:

1. ***FQI***. FQI is perhaps one of the most popular Q-learning type algorithms. It casts the estimation of the optimal Q-function into a sequential regression problem, enabling the use of existing supervised learning algorithms to estimate the optimal Q-function (Ernst et al. 2005; Riedmiller 2005). Theoretically, the regret of the resulting estimated optimal policy has been widely studied in the literature (see, e.g., J. Chen and Jiang 2019; Fan et al. 2020; Y. Jin et al. 2021; Munos and Szepesvári 2008). Our proposal differs from the aforementioned studies both methodologically and theoretically. Methodologically, the proposed GFQI accounts for intra-cluster correlations, unlike standard FQI that treats observations as i.i.d. Theoretically, we conduct a refined analysis to show that our estimated optimal policy incurs a smaller regret than the standard FQI. Such a difference cannot be captured in traditional regret analyses, which typically focus on the order of magnitude of the regret bound, as the two algorithms' regret bounds are of the same order.
2. ***Generalized temporal difference (GTD) learning***. Our proposal is also closely related to the GTD learning algorithm developed by Ueno et al. (2011), which proposes to utilize a set of estimating equations to solve temporal difference (TD) learning for policy evaluation and derives an 'optimal' estimating equation that minimizes the variance of the estimated parameters within that set. A similar work by Min et al. (2021) develops a variance-aware weighted regression method that assigns weights inversely proportional to the variance of the value function for parameter estimation and the subsequent policy evaluation. However, utilizing the variance alone without modifying the estimating equation is insufficient to achieve semi-parametric efficiency (Ueno et al. 2011). Our proposal differs from GTD in that (i) while both our proposal and GTD identify optimal estimating equations to minimize the variance of the parameter estimates, we employ GEE to account for the intra-cluster correlations, whereas GTD still treats observations as i.i.d. (ii) We consider policy learning, a more complicated problem than policy evaluation which often serves as an intermediate step in policy learning (Sutton and Barto 2018).

3 | Clustered MDPs

We propose a clustered MDP to model the intra-cluster correlations within the clustered data. To elaborate on this DGP,

we discuss its connections to and differences from the standard MDP model in this section. Notably, the standard MDP requires the error residuals to be independent. In contrast, the proposed clustered MDP allows residuals to be dependent over the population.

We first introduce the standard MDP commonly employed in RL to formulate numerous sequential decision-making problems. An MDP is typically denoted by $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho, \gamma \rangle$ where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively; $\mathcal{T}$ and $\mathcal{R}$ are the state transition and reward functions; ρ denotes the initial state distribution and γ denotes the discount factor, bounded between 0 and 1.

Let $\left(S_t, A_t, R_t\right)_{t\geq 0}$ denote the sequence of state-action-reward triplets generated under this MDP. Its data generating process can be described as follows:

1. At the initial time $t = 0$, the environment's state $S_t \in \mathcal{S}$ is generated according to ρ, that is, $S_0 \sim \rho(\bullet)$;
2. Subsequently, at each time t, the agent selects an action $A_t \in \mathcal{A}$ after observing S_t;
3. Next, the environment provides an immediate reward R_t with an expected value of $\mathcal{R}\left(A_t, S_t\right)$ and transits into a new state $S_{t+1} \sim \mathcal{T}\left(\bullet \mid A_t, S_t\right)$ at time $t + 1$;
4. Steps 2 and 3 repeat until the environment reaches a terminal state.

An MDP contains two basic assumptions (see the left panel of Figure 2 for illustrations):

i. The first one is a Markov assumption, which requires both immediate reward and future state to depend on the past data history only through the current state-action pair. This assumption precludes any directed path from the past state to future state that does not traverse current state-action pair.

ii. The second one is an independence assumption, which requires the state-action-reward triplets to be independent across different trajectories. This assumption precludes any directed path among trajectories.

The goal is to learn an optimal policy π^* that maximizes the expected cumulative reward $\sum_{t\geq 0}\gamma^t \mathbb{E}^{\pi}\left(R_t\right)$ where $\mathbb{E}^{\pi}\left(R_t\right)$ denotes the expected reward the agent receives at time t following a given policy π.

This paper studies MDPs under the clustered data setting. We use $\left\langle \mathcal{S}, \mathcal{A}, \mathscr{M}, \{\mathcal{T}_m\}_m, \mathcal{R}, \{\rho_m\}_m, \gamma \right\rangle$ to denote such a clustered MDP. In this notation:

- $\mathcal{S}, \mathcal{A}, \mathcal{R}$ and γ denote the same state space, action space, reward function and discount factor as those in a standard MDP;
- $\mathscr{M}$ denotes the probability mass function of M—the number of subjects within each cluster;
- ρ_M and $\mathcal{T}_M$ denote the joint distribution function of all M initial states, and the transition function of all the M states, assuming there are M subjects in the cluster. When restricting to a single state, its marginal initial state distribution and marginal transition function are reduced to ρ and $\mathcal{T}$, respectively.

Such a clustered MDP generates M trajectories, each containing a sequence of state-action-reward triplets $\left(S_t^{(m)}, A_t^{(m)}, R_t^{(m)}\right)_{t\geq 0}$ over time, for $1 \leq m \leq M$. Similar to the MDP model, each trajectory marginally follows an MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho, \gamma \rangle$. However, different from the MDP, these trajectories can be dependent.

To elaborate, we visualize a clustered MDP in the right panel of Figure 2. In this example, a common residual process $\{e_t\}_{t\geq 1}$ affects the rewards and/or states of two subjects simultaneously. For instance, the reward residuals of two subjects could be identical,

$$R_t^{(1)} - \mathcal{R}\left(A_t^{(1)}, S_t^{(1)}\right) = R_t^{(2)} - \mathcal{R}\left(A_t^{(2)}, S_t^{(2)}\right), \forall t.$$

This invalidates the independence assumption. However, we do require $\{e_t\}_{t\geq 1}$ to be independent over time to meet the Markov assumption. Otherwise, each trajectory follows a partially observable MDP, invalidating most RL algorithms that rely on the MDP assumption. As such, each trajectory in a clustered MDP follows an MDP marginally, and its expected cumulative

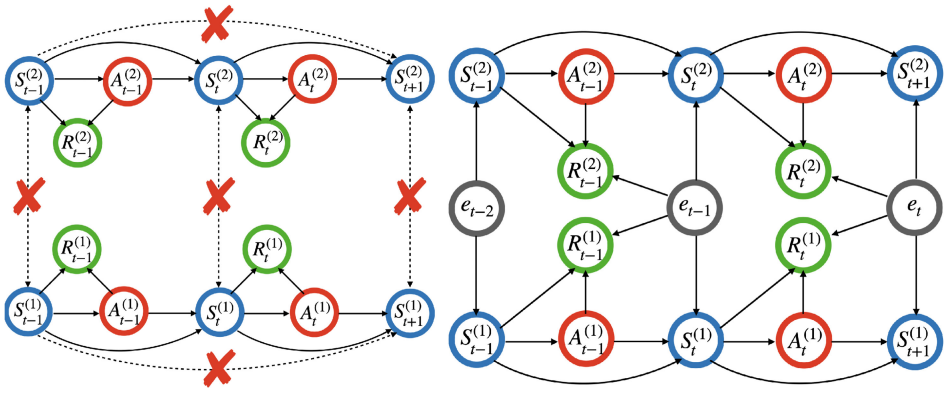

FIGURE 2 | Left panel: visualization of MDPs. $\left(S_t^{(1)}, A_t^{(1)}, R_t^{(1)}\right)_{t\geq 1}$ and $\left(S_t^{(2)}, A_t^{(2)}, R_t^{(2)}\right)_{t\geq 1}$ denote two data trajectories. The Markov assumption precludes paths $S_{t-1}^{(1)} \to S_{t+1}^{(1)}$ and $S_{t-1}^{(2)} \to S_{t+1}^{(2)}$. The independence assumption precludes paths across the two trajectories. Right panel: a clustered MDP example.

reward $\sum_{t\geq 0}\gamma^t\mathbb{E}^{\pi}(R_t)$ under any policy π[1] remains the same as if it were generated independently from an MDP. Consequently, π^* continues to maximize the expected cumulative reward. Moreover, most existing RL algorithms remain consistent to identify π^*. However, they are no longer as sample efficient as under Model I, as demonstrated in the right panel of Figure 1.

4 | Generalized FQI

In this section, we proposed a generalized FQI algorithm for more effective policy learning in the presence of intra-cluster correlations within a clustered MDP. For illustration purposes, we begin by introducing the standard FQI algorithm and adapting GTD for policy learning (denoted as AGTD). However, neither FQI nor AGTD addresses data correlations. Our proposal, presented later, is built upon these algorithms and specifically designed to manage such correlations. All three algorithms are Q-learning type methods, which focus on estimating the following optimal Q-function, defined by

$$Q^*(a,s)=\sum_{t\geq 0}\gamma^t\mathbb{E}^{\pi^*}\left(R_t|A_0=a,S_0=s\right),$$

and derive the estimated optimal policy as the greedy policy with respect to the estimated Q-function. See Algorithms 1–3 for their pseudocodes.

4.1 | Algorithm 1 (FQI)

As commented in Section 3, standard FQI remains consistent in identifying the optimal policy at a population level when applied to clustered MDPs, although it lacks sample efficiency in finite samples. It organizes all data into a collection of state-action-reward-next-state (S,A,R,S') tuples $\mathcal{D}$ and iteratively estimates Q^* via supervised learning. More specifically, the initial Q-function estimator $Q^{(0)}$ is set to a zero function. During the kth iteration, FQI solves the following optimization,

$$\arg\min_{Q\in\mathcal{Q}}\sum_{(S,A,R,S')\in\mathcal{D}}\left[R+\max_a\gamma Q^{(k-1)}(a,S')-Q(A,S)\right]^2, \tag{2}$$

to compute $Q^{(k)}$.

Here, $\mathcal{Q}$ denotes the function class used to model Q^*. It can be any parametric class, such as the class of linear models where $\mathcal{Q}=\{Q:Q(a,s)=\beta^{\top}\phi(a,s)\}$ for a set of state-action features ϕ with dimension d, or other more complex models such as neural networks.

ALGORITHM 1 | Fitted Q-iteration (FQI).

Initialize $Q^{(0)}\leftarrow 0$ and $k\leftarrow 0$.
repeat
  $k\leftarrow k+1$
  Update $Q^{(k)}$ by solving (2).
until Convergence criteria is met.
return $Q^{(k)}$

4.2 | Algorithm 2 (Adaptation of GTD for Policy Learning)

GTD is originally designed for the purpose of Q-function evaluation. Here, we extend it to learn the optimal policy π^*. Following GTD, we assume the optimal Q-function is linear, that is, $Q^*(a,s)=\phi^{\top}(a,s)\beta^*$ for some $\beta^*\in\mathbb{R}^d$. According to the Bellman optimality equation (see, e.g., Agarwal et al. 2019), for any d-dimensional state-action feature Φ, the following estimating equation is unbiased for estimating β^*,

$$\sum_{(S,A,R,S')\in\mathcal{D}}\Phi(A,S)\left[R+\gamma\max_a\phi^{\top}(a,S')\widehat{\beta}-\phi^{\top}(A,S)\widehat{\beta}\right]=0, \tag{3}$$

that is, the left-hand-side (LHS) of (3) equals zero when replacing $\widehat{\beta}$ with β^*. This motivates us to solve the above equation to compute the estimator $\widehat{\beta}$ for β^*.

The choice of the state-action feature Φ in (3) is crucial. Each Φ leads to a different estimator $\widehat{\beta}$. Motivated by GTD, we aim to identify the optimal state-action feature that minimizes the mean squared error (MSE) of $\widehat{\beta}$. Assuming the data comes from a standard MDP with independent trajectories, it follows from Theorem 1 (see Section 5) that the resulting optimal Φ is specified by $\phi^*(a,s)\sigma^{-2}(a,s)$ where

$$\phi^*(A,S)=\phi(A,S)-\gamma\mathbb{E}\left[\phi\left(\pi^*(S'),S'\right)|A,S\right], \tag{4}$$

and $\sigma^2(A,S)$ denotes the conditional variance of the TD error $R+\gamma\ \max_a Q^*(\pi^*(S'),S')-Q^*(A,S)$ given A and S.

To solve this estimating equation, we estimate β^* iteratively in a similar manner to FQI. At each iteration, we first use the currently estimated optimal policy—greedy with respect to the current Q-function estimator—and supervised learning to approximate π^*, the expectation in (4) and σ^2, in order to construct the optimal Φ. We next solve the resulting estimating equation in (3). The aforementioned procedure is repeated until convergence.

ALGORITHM 2 | Adapted generalized TD (AGTD).

Initialize $\beta^{(0)}\leftarrow 0$ and $k\leftarrow 0$.
repeat
  $k\leftarrow k+1$
  Estimate π^*, the expectation in (4) and σ^2 by the currently estimated optimal policy and supervised learning to estimate ϕ^* and σ^2. Let $\widehat{\phi}^*$ and $\widehat{\sigma}^2$ denote their estimators.
  Update $\beta^{(k+1)}$ by solving
$$\sum_{(S,A,R,S')\in\mathcal{D}}\frac{\widehat{\phi}^*(A,S)}{\widehat{\sigma}^2(A,S)}\left[R+\gamma\times\max_a\phi^{\top}(a,S')\beta^{(k)}-\phi^{\top}(A,S)\beta^{(k+1)}\right]=0.$$
until Convergence criteria is met.
return $\beta^{(k)}$

4.3 | Algorithm 3 (GFQI)

Algorithm 2 derives the optimal basis function ϕ^* under the independence assumption, which is clearly violated in clustered MDPs. To address this limitation, we incorporate GEE into FQI to effectively accommodate data correlations.

For a clustered MDP, we use boldface letters $\mathbf{S}$, $\mathbf{A}$, $\mathbf{R}$ and $\mathbf{S}'$ to denote the collections of state-action-reward-next-state tuples across all M subjects within a cluster at a given time, that is, $\mathbf{S} = \left(S^{(1)\top}, \cdots, S^{(M)\top}\right)^\top$ and $\mathbf{A}$, $\mathbf{R}$, $\mathbf{S}'$ can be defined similarly. Under the given model assumption, the parameter of interest β^* satisfies the Bellman optimality equation $\mathbb{E}\left[\boldsymbol{\delta}\left(\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{S}'; \beta^*\right)\right] = \mathbf{0}$ where $\boldsymbol{\delta}\left(\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{S}'; \beta^*\right)$ denotes an M-dimensional cluster-wise TD error whose jth entry is given by

$$R^{(j)} + \max_a \ \gamma\phi\left(a, S^{(j)'}\right)\beta^* - \phi\left(A^{(j)}, S^{(j)}\right)\beta^*.$$

This motivates us to construct the following estimating equation to estimate β^*:

$$\sum_{(\mathbf{S},\mathbf{A},\mathbf{R},\mathbf{S}')} \Phi(\mathbf{A}, \mathbf{S})\boldsymbol{\delta}\left(\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{S}'; \widehat{\beta}\right) = \mathbf{0}, \tag{5}$$

for any $d \times M$ matrix Φ that is a function of $\mathbf{A}$ and $\mathbf{S}$, where the summation in (5) is taken over all $\left(\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{S}'\right)$ tuples across different times and clusters.

According to Theorem 1, the optimal Φ is given by

$$\Phi^*(\mathbf{A}, \mathbf{S}) = \left[\phi^*\left(A^{(1)}, S^{(1)}\right), \cdots, \phi^*\left(A^{(M)}, S^{(M)}\right)\right]\mathbf{V}^{-1},$$

where ϕ^* is defined in (4) and $\mathbf{V}$ is the covariance matrix of the cluster-wise TD error.

Similar to GEE, we model $\mathbf{V}$ via $\boldsymbol{BCB}$ where $\boldsymbol{B}$ is a diagonal matrix with each jth diagonal element equal to the standard deviation of the jth TD error $\sigma\left(A^{(j)}, S^{(j)}\right)$. The matrix $\boldsymbol{C}$ in the middle serves as a working correlation matrix to model the intra-cluster correlations. Typically, $\boldsymbol{C}$ is set to an exchangeable correlation matrix, where the diagonal elements are 1, and the off-diagonal elements equal some correlation coefficient $-1 < \rho < 1$.

To solve the resulting GEE in (5), we employ the same iterative approach to AGTD. Specifically, at each iteration, we estimate the correlation coefficient ρ and ϕ^* and update the parameter $\beta^{(k)}$ by solving

$$\sum_{(\mathbf{S},\mathbf{A},\mathbf{R},\mathbf{S}')} \Phi^*(\mathbf{A}, \mathbf{S})\boldsymbol{\delta}\left(\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{S}'; \beta^{(k)}, \beta^{(k-1)}\right) = 0, \tag{6}$$

where $\boldsymbol{\delta}\left(\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{S}'; \beta^{(k)}, \beta^{(k-1)}\right)$ denotes the TD error vector with the jth entry given by

$$R^{(j)} + \max_a \ \gamma\phi\left(a, S^{(j)'}\right)\beta^{(k-1)} - \phi\left(A^{(j)}, S^{(j)}\right)\beta^{(k)}.$$

Since we iteratively solve the GEE to estimate the optimal Q-function, we refer to our algorithm as generalized FQI, or GFQI, for short. Notice that GFQI is reduced to AGTD when the correlation matrix $\boldsymbol{C}$ is set to the identity matrix.

5 | Statistical Properties

In this section, we study the theoretical properties of the proposed GFQI (Algorithm 3). We first obtain a bias-variance decomposition of the MSE of the estimated β^*. We next analyse the regret of the resulting estimated optimal policy. See Table 1 for a summary of our theoretical findings. Without loss of generality,

ALGORITHM 3 | Generalized fitted Q-iteration (GFQI).

Set the iteration counter $k \leftarrow 0$.
Initialize $\beta^{(0)} \leftarrow 0$ so that the Q-function $Q^{(k)}(a, s) = \phi^\top(a, s)\beta^{(k)}$ is initialized to zero.
repeat
  Estimate π^* by the greedy policy $\pi^{(k)} \leftarrow \arg\max_a \phi^\top(a, s)\beta^{(k)}$ with respect to $Q^{(k)}$.
  Estimate $\phi^*(A, S) = \mathbb{E}\left[\phi\left(\pi^*\left(S'\right), S'\right) \mid A, S\right]$ using supervised learning, with $\left\{\phi\left(\pi^{(k)}\left(S'\right), S'\right)\right\}$ as the response variables and $\{(A, S)\}$ as the predictors.
  Calculate the TD error $\delta\left(A_t^{(m)}, S_t^{(m)}, R_t^{(m)}, S_{t+1}^{(m)}\right) \leftarrow R_t^{(m)} + \gamma\max_a Q^{(k-1)}\left(a, S_{t+1}^{(m)}\right) - Q^{(k)}\left(A_t^{(m)}, S_t^{(m)}\right)$ for $m = 1, \cdots, M$ and $t = 1, \cdots, T$.
  Use these TD errors to estimate $\mathbf{V} = \left\{V_{m_1,m_2}\right\}_{m_1,m_2}$ where $V_{m_1,m_2} = \mathrm{Cov}\left(\delta\left(A_t^{(m_1)}, S_t^{(m_1)}, R_t^{(m_1)}, S_{t+1}^{(m_1)}\right), \delta\left(A_t^{(m_2)}, S_t^{(m_2)}, R_t^{(m_2)}, S_{t+1}^{(m_2)}\right)\right)$.
  Estimate Φ^* by plugging the estimated ϕ^* and $\boldsymbol{V}$ into
  $$\Phi^*(\mathbf{A}, \mathbf{S}) = \left[\phi^*\left(A^{(1)}, S^{(1)}\right), \cdots, \phi^*(A^{(M)}, S^{(M)})\right]\mathbf{V}^{-1},$$
  Update $\beta^{(k+1)}$ by solving
  $$\sum_{(\mathbf{S},\mathbf{A},\mathbf{R},\mathbf{S}')} \Phi^*(\mathbf{A}, \mathbf{S})\boldsymbol{\delta}\left(\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{S}'; \beta^{(k)}, \beta^{(k-1)}\right) = 0,$$
  with the jth element entry of $\boldsymbol{\delta}\left(\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{S}'; \beta^{(k)}, \beta^{(k-1)}\right)$ given by $\sigma\left(A^{(j)}, S^{(j)}\right)$.
  $k \leftarrow k + 1$
until Convergence criteria is met.
return $\beta^{(k)}$

TABLE 1 | A summary of our theoretical findings.

Correlation matrix	Theorem 1 (estimated β^*)	Theorem 2 (policy)
Correctly specified	Minimal asymptotic variance	Minimal asymptotic regret
Mis-specified	Consistent	Consistent

we assume that the numbers of subjects for all clusters are the same and denote this number as M. Let N denotes the number of state-action-reward-next-state tuples $(\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{S}')$ in the dataset.

We impose Assumptions 1–8 to derive our theories and relegate them to Appendix S1 to save space. We briefly discuss these assumptions below:

- Assumptions 1, 2 and 3 are standard in the RL literature (see, e.g., J. Chen and Jiang 2019; Fan et al. 2020; Uehara, Imaizumi, et al. 2022). They require that all immediate rewards be bounded by a constant $R_{\max} < \infty$, the linear function class be rich enough to cover the reward function $\mathscr{R}$ (realizability) and be closed under the Bellman operator (completeness), respectively.
- Assumption 4 requires the underlying clustered MDP to satisfy the geometric ergodicity condition (Bradley 2005), which relaxes the existing i.i.d. data assumption commonly imposed in RL (see, e.g., J. Chen and Jiang 2019; Dai et al. 2020; Fan et al. 2020). It is equivalent to the β-mixing condition (Luckett et al. 2020) and is strictly weaker than the uniform ergodicity assumption imposed in the literature (Zou et al. 2019).
- Assumption 5 is similar to the stability assumption used in the off-policy evaluation literature (see, e.g., Perdomo et al. 2022) but is specifically tailored for policy learning. This assumption is frequently employed in settings where linear function approximation is utilized for modelling the (optimal) Q-function (Bian et al. 2025; Ertefaie and Strawderman 2018; M. Li et al. 2025; Perdomo et al. 2023; Shi et al. 2022). Following the arguments in Shi et al. (2022), we can show that this assumption is likely to hold when data are collected using the ϵ-greedy algorithm; see Section S1.1 of the Supporting Information for details.
- Assumption 6 requires the number of GFQI iterations K to be much larger than $\log(N) / \log(\gamma^{-1})$. This assumption is mild as K is user-specified.
- Finally, Assumption 7 requires the uniqueness of the optimal policy whereas Assumption 8 requires the value function to be a smooth function of the model parameter β. These assumptions enable us to derive the leading terms of the MSE and regret, which are essential for comparing the performance of different algorithms; see our discussions below in Theorems 1 and 2.

5.1 | Error Analysis of the Estimated β^*

The following theorem establishes the efficiency and robustness of the estimated β^* obtained from Algorithm 3.

Theorem 1. *(MSE). Suppose Assumptions 1–7 are satisfied. For a sufficiently large sample size N, $\widehat{\beta}$ computed by Algorithm 3 attains the following properties:*

1. ***Robustness**: The MSE of $\widehat{\beta}$ is given by*
2. ***Efficiency**: $\widehat{\beta}$ reaches the minimal asymptotic MSE—obtained by removing errors that are high-order in the sample size N—among the class of solutions to (5) if $\Phi = \Phi^*$ with a correctly specified correlation matrix $\boldsymbol{C}$.*

$$\frac{\operatorname{tr}\left[\boldsymbol{W}^{-1}(\Phi)\Sigma(\Phi)\left(\boldsymbol{W}^{-1}(\Phi)\right)^{\top}\right]}{N} + O\left(\frac{R_{\max}\log(N)}{N^{\frac{3}{2}}(1-\gamma)^5}\right), \tag{7}$$

where $tr(\bullet)$ denotes the trace of a matrix, $\boldsymbol{W}(\Phi) = (1-\gamma)^{-1}M^{-1}\mathbb{E}\left[\Phi(\mathbf{A},\mathbf{S})\left\{\phi(\mathbf{A},\mathbf{S}) - \gamma\phi(\pi^*(\mathbf{S}'),\mathbf{S}')\right\}\right]$, $\phi(\mathbf{A},\mathbf{S}) = \left[\phi\left(A^{(1)},S^{(1)}\right), \cdots, \phi\left(A^{(M)},S^{(M)}\right)\right]$ and $\Sigma(\Phi) = M^{-1}\mathbb{E}\left[\Phi(\mathbf{A},\mathbf{S})\mathbf{V}\Phi^{\top}(\mathbf{A},\mathbf{S})\right]$.

The first part of Theorem 1 derives a rigorous bias-variance decomposition for the MSE of $\widehat{\beta}$. Specifically, the first term in the error bound (see (7)) represents the leading asymptotic variance, which scales as N^{-1}, whereas the second term arises from its finite-sample bias, which decays at a faster rate of $N^{-3/2}$. As a result, the bias term diminishes more rapidly than the asymptotic variance as the sample size N increases. Meanwhile, the bias term is proportional to the reward upper bound $R_{\max}$ and increases with the $(1-\gamma)^{-1}$ term, which can be interpreted as the 'horizon' in episodic tasks. Finally, regardless of the working correlation structure employed, both the bias term and asymptotic variance decay to zero as N approaches infinity, demonstrating the robustness of our proposed GFQI.

The second part of Theorem 1 demonstrates that, when setting Φ to Φ^* with a correctly specified correlation matrix, the proposed GFQI minimizes the leading asymptotic variance of the resulting estimator. This efficiency is achieved through the following three key steps: (i) using ϕ^* instead of ϕ itself to construct the state-action feature; (ii) inversely weighting each state-action feature by the conditional variance of the TD error σ^2; (iii) incorporating $\boldsymbol{C}$ to account for the within-cluster correlation. In comparison:

- ***Standard FQI*** adopts none of these steps, which often incurs a large MSE;
- ***Variance-aware regression*** (Min et al. 2021) functions similarly to the second step but remains inefficient as it omits the first and third steps;
- ***AGTD*** improves upon variance-aware regression by incorporating the first two steps but still falls short of efficiency because it does not exploit the correlation structure.

Our work builds on AGTD by accounting for this correlation, resulting in a more efficient estimator.

Finally, we emphasize that a major difference of our proposal from the existing literature lies in the refined theoretical analysis of the error terms. While existing results in the machine learning literature often focus on establishing the order of the MSE, our

analysis goes further by explicitly characterizing the leading term and developing a method to optimize it. Without such an analysis, all the aforementioned methods would achieve the same order of magnitude, making it impossible to distinguish their performance.

5.2 | Regret Analysis

The following theorem demonstrates that the estimation error of $\widehat{\beta}$ can be translated into the regret of the resulting estimated optimal policy.

Theorem 2. *Suppose Assumptions 1–8 are satisfied. For a sufficiently large N, we have the following results:*

1. ***Robustness**: The regret of the estimated optimal policy is given by*

$$-\frac{1}{2}\mathrm{tr}\left(\mathrm{Var}_A\left(\widehat{\beta}\right)H\right)+O\left(\frac{R_{\max}^{3/2}\log(N)}{(1-\gamma)^6 N^{\frac{3}{2}}}\right), \tag{8}$$

   where $\mathrm{Var}_A\left(\widehat{\beta}\right)$ denotes the asymptotic covariance matrix of $\widehat{\beta}$, given by $N^{-1}\boldsymbol{W}^{-1}(\Phi)\Sigma(\Phi)\left(\boldsymbol{W}^{-1}(\Phi)\right)^{\top}$, and H denotes the Hessian matrix:

$$H=\frac{\partial\mathcal{V}(\beta)}{\partial\beta\partial\beta^{\top}}\Big|_{\beta=\beta^*},$$

   where $\mathcal{V}(\beta)$ denotes the expected cumulative reward $\sum_{t\geq 0}\gamma^t\mathbb{E}^{\pi(\beta)}(R_t)$ under the greedy policy $\pi_\beta(s)=\arg\max_a\phi^{\top}(a,s)\beta$ parameterized by β.

2. ***Efficiency**: The regret bound in (8) is asymptotically minimized (i.e., its leading first term is minimized) when the correlation matrix is correctly specified.*

As commented earlier, existing theoretical analyses provide only the order of magnitude of the policy regret, which is insufficient for comparing different algorithms when their regrets share the same order. To address this limitation, similar to Theorem 1, we refine the existing analysis by decomposing the regret into a leading term, which is related to the asymptotic variance of $\widehat{\beta}$, and a higher-order residual term in (8). By definition, $\mathcal{V}(\beta)$ is maximized at β^*. As such, H is negative semidefinite, making the leading term nonnegative. Given that the leading term is of the order $O(N^{-1})$, it is indeed the dominating factor. Meanwhile, this dominating factor is minimized when the correlation matrix is correctly specified, which demonstrates the efficiency of the proposed GFQI. When mis-specified, both the leading term and the higher order term converge to zero as the sample size N approaches infinity, verifying the robustness property.

6 | Numerical Study

In this section, we employ the IHS dataset to create a simulation environment to investigate the empirical performance of the proposed GFQI in real-world applications. To save space, additional synthetic data analysis results are relegated to Section S2 of the Supporting Information. The code for reproducing these experiments is available at https://github.com/zaza0209/GEERL.

To mimic the data generation process in the IHS study, we first train a state transition model and a reward model. Here, the state variable S_t is set to be the cubic-root of daily step count of an intern at day $t-1$, and the reward variable R_t is set to be their daily mood score at day t. We remark that the original daily step count is heavy-tailed, and the cubic-root transformation brings its distribution closer to normality. See Figure C.5 of M. Li et al. (2025) for a visualization of the resulting transformed variable. The action A_t is binary, where $A_t=1$ indicates that a message was received on day t, and $A_t=0$ indicates no message was received. Further details regarding the generative model and correlation structure are provided in Section S2 of the Supporting Information.

The sample size is determined by three parameters: (i) the number of clusters n; (ii) the cluster size M and (iii) the number of days T, for which the study lasts. We consider a base simulation setting with $n=5$, $M=10$ and $T=14$ (2 weeks). We next vary each parameter independently to explore their effects: $n\in\{5,10,15,20,25,30\}$, $M\in\{10,20,30,40,50\}$, and $T\in\{7,14,21,28,35\}$, while keeping the other two parameters fixed. The decision of actions is made on each day of a week. We also vary (iv) an additional parameter $\psi\in\{1,3,5,7,9\}$ which determines the intra-cluster correlations. This yields a total of $16\times 5=80$ simulation settings. For each setting, we repeats the experiments 20 times with different seeds.

Comparison is made among the following algorithms/policies:

- **FQI**: The standard FQI (Algorithm 1).
- **AGTD**: The adapted GTD (Algorithm 2).
- **GFQI**: The proposed GFQI with exchangeable working correlation structure (Algorithm 3).
- **CQL** (Kumar et al. 2020): The conservative Q-learning algorithm that adds a regularization term to the loss function to prevent the learned Q-values from being overestimated.
- **DDQN** (van Hasselt et al. 2015): An offline version of the double deep Q-network algorithm, designed to reduce the maximization bias by using two separate deep neural network models for the optimal Q-function, one for greedy action selection and the other for Q-value evaluation.
- **Behaviour**: The uniform random policy used in the IHS study to collect the offline data.

Notice that the first five algorithms are all Q-learning-based. In particular, the first three algorithms employ linear polynomial basis functions to parameterize the optimal Q-function, with its degree selected from {1,2,3,4} using 5-fold cross-validation. For GFQI and AGTD, the conditional expectation in (4) at each iteration is estimated via linear regression. In practice, we recommend using supervised learning algorithms that are computationally efficient, as the proposed GFQI iteratively updates the Q-function, and each iteration requires to compute the conditional expectation in (4). More generally, kernel ridge regression can also be employed, as the kernel matrix only needs to be computed once and reused across iterations. The covariance matrix $\mathbf{V}$ is updated at each iteration by imposing the exchangeable correlation structure among TD errors. The next two algorithms, CQL and DDQN, use deep neural networks for Q-function estimation.

Estimated optimal policies are evaluated via the Monte Carlo method. Specifically, we generate 10 clusters' data that lasts for 35 days, where each cluster contains 100 trajectories. To minimize the evaluation error, we omit the reward residuals when generating the rewards. Finally, we average these generated rewards across all trajectories to evaluate the expected return.

Figure 3 visualizes the regrets of estimated policies computed by various RL algorithms with varying cluster sizes, numbers of clusters and horizons. We summarize our results as follows:

1. First, we find that GFQI considerably outperforms FQI and AGTD in scenarios with large intra-cluster correlations (i.e., $\psi \geq 5$) and achieves better performance in general with small correlations (i.e., $\psi \leq 3$). This result highlights the importance of accounting for the correlations in clustered data for policy learning.
2. Second, the proposed GFQI consistently achieves smaller or comparable regrets than CQL and DDQN. This comparison implies that the danger of ignoring data correlations in policy learning can overshadow the benefits provided by the flexibility of deep neural network used in CQL and DDQN.
3. Finally, all RL algorithms achieve significantly smaller regrets than the behaviour policy in most cases, highlighting the value of policy learning in improving the medical outcomes of interns.

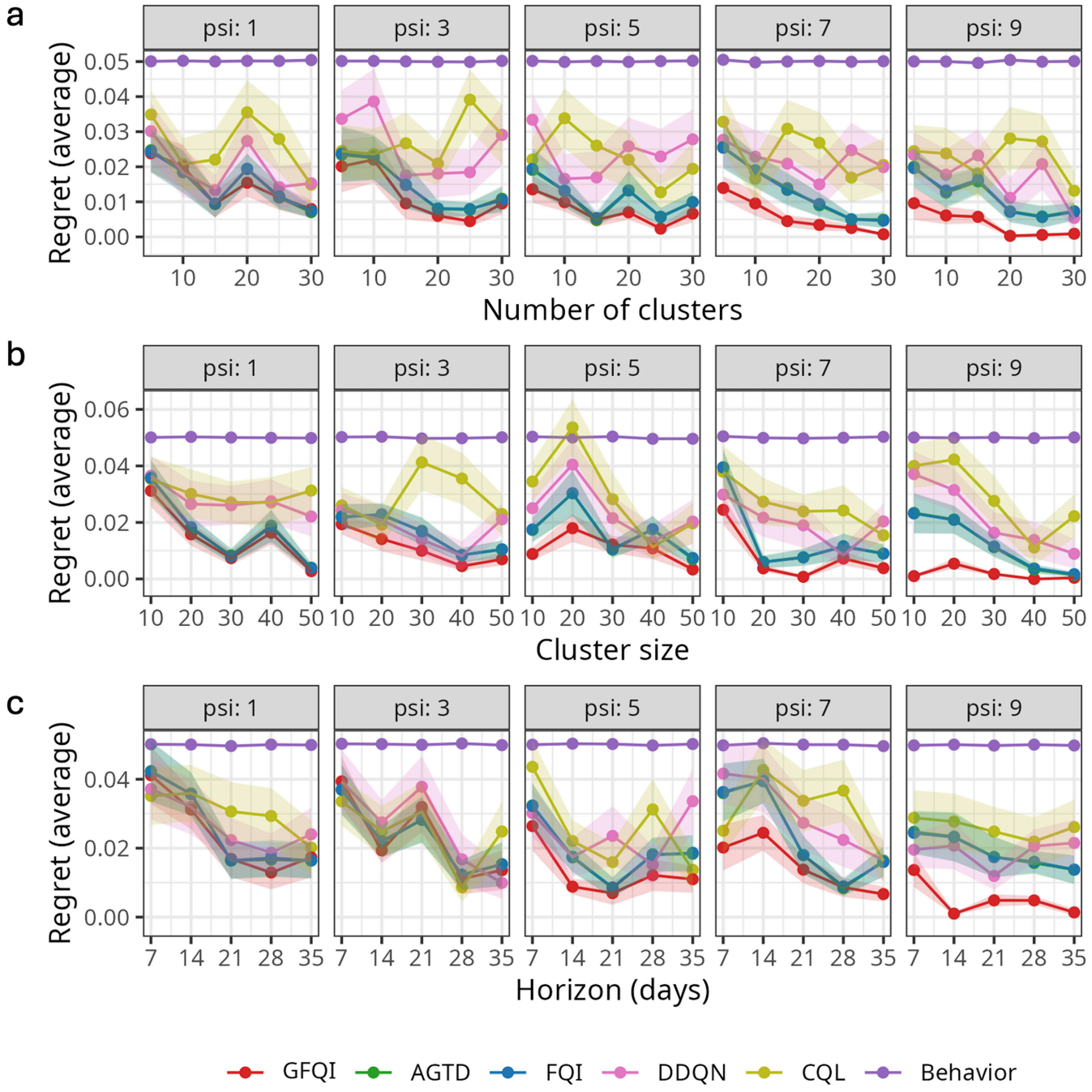

FIGURE 3 | Regret of the average reward with varying (i) numbers of clusters, (ii) cluster sizes, (iii) numbers of days and (iv) values of intra-cluster correlation parameter (ψ). The shaded band represents the standard error. The green line (AGTD) and blue line (FQI) are largely overlapped.

Author Contributions

Liyuan Hu: conceptualization, methodology, data analysis, writing – original draft preparation, theoretical proofs, simulation, writing – review and editing. **Jitao Wang:** data analysis, simulation, writing – review and editing. **Zhenke Wu:** conceptualization, methodology, writing – review and editing. **Chengchun Shi:** conceptualization, methodology, theoretical proofs, writing – review and editing.

Acknowledgements

The authors are grateful for the contributions of the researchers, administrators and participants involved in the IHS (https://clinicaltrials.gov/study/NCT03972293). This work was partly supported by National Institutes of Health Grants (R01MH101459 from National Institute of Mental Health to Z.W.; and R01NR013658 from National Institute of Nursing Research to Z.W. and J.W.) and an Engineering and Physical Sciences Research Council Grant (EP/W014971/1 to C.S.). The authors also thank the editor, associate editor and reviewers for their comments, which have led to substantial improvements in the manuscript.

Data Availability Statement

Data are available from the authors with the permission of the PI of IHS, Dr. Srijan Sen.

Endnotes

[1] In this paper, we focus on policies whose action assignment mechanisms depend only on their own trajectory's data history and not on other trajectories.

References

Agarwal, A., N. Jiang, S. M. Kakade, and W. Sun. 2019. "Reinforcement Learning: Theory and Algorithms." CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep, 32, 96.

Andersen, J. B., and Q. Zhao. 2025. "A Graphical Approach to State Variable Selection in Off-Policy Learning." arXiv Preprint arXiv:2501.00854.

Bian, Z., C. Shi, Z. Qi, and L. Wang. 2025. "Off-Policy Evaluation in Doubly Inhomogeneous Environments." *Journal of the American Statistical Association* 120, no. 550: 1102–1114.

Bradley, R. C. 2005. "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions." *Probability Surveys* 2: 107–144.

Cai, H., K. Ren, W. Zhang, et al. 2017. "Real-Time Bidding by Reinforcement Learning in Display Advertising." In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 661–670.

Carta, S., A. Corriga, A. Ferreira, A. S. Podda, and D. R. Recupero. 2021. "A Multi-Layer and Multi-Ensemble Stock Trader Using Deep Learning and Deep Reinforcement Learning." *Applied Intelligence* 51: 889–905.

Chakraborty, B., and E. E. Moodie. 2013. *Statistical Methods for Dynamic Treatment Regimes*. Springer-Verlag. https://doi.org/10.1007/978-1-4614-7428-9.

Chen, E. Y., R. Song, and M. I. Jordan. 2024. "Reinforcement Learning in Latent Heterogeneous Environments." *Journal of the American Statistical Association* 119: 1–32.

Chen, J., and N. Jiang. 2019. "Information-Theoretic Considerations in Batch Reinforcement Learning." In *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov, vol. 97, 1042–1051. PMLR. https://proceedings.mlr.press/v97/chen19e.html.

Chen, X., Z. Qi, and R. Wan. 2023. "Steel: Singularity-Aware Reinforcement Learning." arXiv preprint arXiv:2301.13152.

Copas, A. J., and S. R. Seaman. 2010. "Bias From the Use of Generalized Estimating Equations to Analyze Incomplete Longitudinal Binary Data." *Journal of Applied Statistics* 37, no. 6: 911–922. https://doi.org/10.1080/02664760902939604.

Dabney, W., M. Rowland, M. Bellemare, and R. Munos. 2018. "Distributional Reinforcement Learning With Quantile Regression." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Dai, B., O. Nachum, Y. Chow, L. Li, C. Szepesvári, and D. Schuurmans. 2020. "Coindice: Off-Policy Confidence Interval Estimation." *Advances in Neural Information Processing Systems* 33: 9398–9411.

Diggle, P., D. Farewell, and R. Henderson. 2007. "Analysis of Longitudinal Data With Drop-Out: Objectives, Assumptions and a Proposal." *Journal of the Royal Statistical Society: Series C: Applied Statistics* 56, no. 5: 499–550. https://doi.org/10.1111/j.1467-9876.2007.00590.x.

Diggle, P. J., P. J. Heagerty, K.-y. Liang, and S. L. Zeger. 2002. *Analysis of Longitudinal Data*. Oxford University Press. https://doi.org/10.1093/oso/9780198524847.001.0001.

Duan, Y., M. Wang, and M. J. Wainwright. 2024. "Optimal Policy Evaluation Using Kernel-Based Temporal Difference Methods." *Annals of Statistics* 52, no. 5: 1927–1952.

Ernst, D., P. Geurts, and L. Wehenkel. 2005. "Tree-Based Batch Mode Reinforcement Learning." *Journal of Machine Learning Research* 6, no. 18: 503–556. http://jmlr.org/papers/v6/ernst05a.html.

Ertefaie, A., and R. L. Strawderman. 2018. "Constructing Dynamic Treatment Regimes Over Indefinite Time Horizons." *Biometrika* 105, no. 4: 963–977.

Fan, J., Z. Wang, Y. Xie, and Z. Yang. 2020. "A Theoretical Analysis of Deep q-Learning." In *Proceedingsof the 2nd Conference on Learning for Dynamics and Control*, edited by A. M. Bayen, A. Jadbabaie, G. J. Pappas, et al., vol. 120, 486–489. PMLR. https://proceedings.mlr.press/v120/yang20a.html.

Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs, eds. 2008. *Longitudinal Data Analysis*, 1st ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781420011579.

Friedel, J. E., W. B. DeHart, A. M. Foreman, and M. E. Andrew. 2019. "A Monte Carlo Method for Comparing Generalized Estimating Equations to Conventional Statistical Techniques for Discounting Data." *Journal of the Experimental Analysis of Behavior* 111, no. 2: 207–224. https://api.semanticscholar.org/CorpusID:59249461.

Hedeker Donald, G. R. D. 2006. *Analysis of Longitudinal Data*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/0470036486#:~:text=Reflecting%20the%20growing%20importance%20and?msockid=2ef388ddba056a1339ae9b4bbbd76ba0.

Hu, L., M. Li, C. Shi, Z. Wu, and P. Fryzlewicz. 2022. "Doubly Inhomogeneous Reinforcement Learning." arXiv preprint arXiv:2211.03983.

Hu, Y., and S. Wager. 2023. "Off-Policy Evaluation in Partially Observed Markov Decision Processes Under Sequential Ignorability." *Annals of Statistics* 51, no. 4: 1561–1585.

Jin, J., C.-N. Song, H. Li, K. Gai, J. Wang, and W. Zhang. 2018. "Real-Time Bidding With Multi-Agent Reinforcement Learning in Display Advertising." *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. https://api.semanticscholar.org/CorpusID:3595955.

Jin, Y., Z. Yang, and Z. Wang. 2021. "Is Pessimism Provably Efficient for Offline RL?" In *International Conference on Machine Learning*, 5084–5096.

Kallus, N., and M. Uehara. 2024. "Efficient Evaluation of Natural Stochastic Policies in Off-Line Reinforcement Learning." *Biometrika* 111, no. 1: 51–69.

Konda, V., and J. Tsitsiklis. 1999. "Actor-Critic Algorithms." *Advances in Neural Information Processing Systems* 12. https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Kosorok, M. R., and E. B. Laber. 2019. "Precision Medicine." *Annual Review of Statistics and Its Application* 6, no. 1: 263–286.

Kosorok, M. R., and E. E. Moodie. 2015. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. SIAM.

Krishnamurthy, V. 2016. *Partially Observed Markov Decision Processes*. Cambridge University Press.

Kumar, A., A. Zhou, G. Tucker, and S. Levine. 2020. "Conservative q-Learning for Offline Reinforcement Learning." https://arxiv.org/abs/2006.04779.

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

Li, G., L. Shi, Y. Chen, Y. Chi, and Y. Wei. 2024. "Settling the Sample Complexity of Model-Based Offline Reinforcement Learning." *Annals of Statistics* 52, no. 1: 233–260.

Li, M., C. Shi, Z. Wu, and P. Fryzlewicz. 2025. "Testing Stationarity and Change Point Detection in Reinforcement Learning." *Annals of Statistics* 53, no. 3: 1230–1256.

Li, Z., J. Chen, E. Laber, F. Liu, and R. Baumgartner. 2023. "Optimal Treatment Regimes: A Review and Empirical Comparison." *International Statistical Review* 91, no. 3: 427–463.

Liang, K.-Y., and S. L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73, no. 1: 13–22. https://doi.org/10.1093/biomet/73.1.13.

Liang, T., and B. Recht. 2025. "Randomization Inference When n Equals One." *Biometrika* 112, no. 2: asaf013.

Liao, P., P. Klasnja, and S. Murphy. 2021. "Off-Policy Estimation of Long-Term Average Outcomes With Applications to Mobile Health." *Journal of the American Statistical Association* 116, no. 533: 382–391.

Liao, P., Z. Qi, R. Wan, P. Klasnja, and S. A. Murphy. 2022. "Batch Policy Learning in Average Reward Markov Decision Processes." *Annals of Statistics* 50, no. 6: 3364–3387.

Lipsitz, S. R., K. Kim, and L. Zhao. 1994. "Analysis of Repeated Categorical Data Using Generalized Estimating Equations." *Statistics in Medicine* 13, no. 11: 1149–1163. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780131106.

Liu, J., Y.-F. Pei, C. J. Papasian, and H.-W. Deng. 2009. "Bivariate Association Analyses for the Mixture of Continuous and Binary Traits With the Use of Extended Generalized Estimating Equations." *Genetic Epidemiology* 33: 217–227. https://api.semanticscholar.org/CorpusID:27437560.

Liu, W., J. Tu, Y. Zhang, and X. Chen. 2023. "Online Estimation and Inference for Robust Policy Evaluation in Reinforcement Learning." arXiv preprint arXiv:2310.02581.

Liu, X.-Y., H. Yang, Q. Chen, et al. 2020. "Finrl: A Deep Reinforcement Learning Library for Automated Stock Trading in Quantitative Finance." arXiv Preprint arXiv:2011.09607.

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

Luo, S., Y. Yang, C. Shi, F. Yao, J. Ye, and H. Zhu. 2024. "Policy Evaluation for Temporal and/or Spatial Dependent Experiments." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 86: 623–649.

Ma, T., J. Zhu, H. Cai, et al. 2023. "Sequential Knockoffs for Variable Selection in Reinforcement Learning." arXiv Preprint arXiv:2303.14281.

McCullagh, P. 2019. *Generalized Linear Models*. Routledge.

Miao, R., B. Shahbaba, and A. Qu. 2025. "Reinforcement Learning for Individual Optimal Policy From Heterogeneous Data." *Annals of Statistics* 53: 2505.09496.

Min, Y., T. Wang, D. Zhou, and Q. Gu. 2021. "Variance-Aware Off-Policy Evaluation With Linear Function Approximation." In *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, vol. 34, 7598–7610. Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/3e6260b81898beacda3d16db379ed329-Paper.pdf.

Mnih, V., K. Kavukcuoglu, D. Silver, et al. 2015. "Human-Level Control Through Deep Reinforcement Learning." *Nature* 518: 529–533. https://doi.org/10.1038/nature14236.

Munos, R., and C. Szepesvári. 2008. "Finite-Time Bounds for Fitted Value Iteration." *Journal of Machine Learning Research* 9, no. 27: 815–857. http://jmlr.org/papers/v9/munos08a.html.

NeCamp, T., S. Sen, E. Frank, et al. 2020. "Assessing Real-Time Moderation for Developing Adaptive Mobile Health Interventions for Medical Interns: Micro-Randomized Trial." *Journal of Medical Internet Research* 22, no. 3: e15033.

Ouyang, L., J. Wu, X. Jiang, et al. 2022. "Training Language Models to Follow Instructions With Human Feedback." *Advances in Neural Information Processing Systems* 35: 27730–27744.

Overall, J. E., and S. Tonidandel. 2004. "Robustness of Generalized Estimating Equation (GEE) Tests of Significance Against Misspecification of the Error Structure Model." *Biometrical Journal* 46, no. 2: 203–213.

Paradis, E., and J. Claude. 2002. "Analysis of Comparative Data Using Generalized Estimating Equations." *Journal of Theoretical Biology* 218, no. 2: 175–185. https://www.sciencedirect.com/science/article/pii/S0022519302930663.

Perdomo, J. C., A. Krishnamurthy, P. Bartlett, and S. Kakade. 2022. "A Complete Characterization of Linear Estimators for Offline Policy Evaluation." *Journal of Machine Learning Research* 24, no. 284: 1–50.

Perdomo, J. C., A. Krishnamurthy, P. Bartlett, and S. Kakade. 2023. "A Complete Characterization of Linear Estimators for Offline Policy Evaluation." *Journal of Machine Learning Research* 24, no. 284: 1–50.

Puterman, M. L. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Qi, Z., C. Bai, Z. Wang, and L. Wang. 2025. "Distributional Off-Policy Evaluation in Reinforcement Learning." *Journal of the American Statistical Association* 1–24. https://doi.org/10.1080/01621459.2025.2506197.

Qin, Z., X. Tang, Q. Li, H. Zhu, and J. Ye. 2025. *Reinforcement Learning in the Ridesharing Marketplace*. Springer.

Ramprasad, P., Y. Li, Z. Yang, Z. Wang, W. W. Sun, and G. Cheng. 2023. "Online Bootstrap Inference for Policy Evaluation in Reinforcement Learning." *Journal of the American Statistical Association* 118, no. 544: 2901–2914.

Riedmiller, M. 2005. "Neural Fitted q Iteration–First Experiences With a Data Efficient Neural Reinforcement Learning Method." In *Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3–7, 2005*. Proceedings 16, 317–328.

Roderick Little, D. R. 2019. *Statistical Analysis With Missing Data*. Wiley.

Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika* 63, no. 3: 581–592. https://doi.org/10.1093/biomet/63.3.581.

Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz. 2015. "Trust Region Policy Optimization." In *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, vol. 37, 1889–1897. PMLR. https://proceedings.mlr.press/v37/schulman15.html.

Shao, Z., P. Wang, Q. Zhu, et al. 2024. "Deepseekmath: Pushing the Limits of Mathematical Reasoning in Open Language Models." arXiv preprint arXiv:2402.03300.

Shen, C.-W., and Y.-H. Chen. 2012. "Model Selection for Generalized Estimating Equations Accommodating Dropout Missingness."

Biometrics 68, no. 4: 1046–1054. https://doi.org/10.1111/j.1541-0420.2012.01758.x.

Shen, C.-W., and Y.-H. Chen. 2013. "Model Selection of Generalized Estimating Equations With Multiply Imputed Longitudinal Data." *Biometrical Journal* 55. Retrieved from: 899–911. https://api.semanticscholar.org/CorpusID:260709490.

Shen, G., R. Dai, G. Wu, S. Luo, C. Shi, and H. Zhu. 2025. "Deep Distributional Learning With Non-Crossing Quantile Network." arXiv preprint arXiv:2504.08215.

Shi, C., S. Luo, Y. Le, H. Zhu, and R. Song. 2024. "Statistically Efficient Advantage Learning for Offline Reinforcement Learning in Infinite Horizons." *Journal of the American Statistical Association* 119, no. 545: 232–245.

Shi, C., Z. Qi, J. Wang, and F. Zhou. 2024. "Value Enhancement of Reinforcement Learning via Efficient and Robust Trust Region Optimization." *Journal of the American Statistical Association* 119, no. 547: 2011–2025.

Shi, C., R. Wan, G. Song, S. Luo, H. Zhu, and R. Song. 2023. "A Multiagent Reinforcement Learning Framework for Off-Policy Evaluation in Two-Sided Markets." *Annals of Applied Statistics* 17, no. 4: 2701–2722.

Shi, C., X. Wang, S. Luo, H. Zhu, J. Ye, and R. Song. 2023. "Dynamic Causal Effects Evaluation in a/b Testing With a Reinforcement Learning Framework." *Journal of the American Statistical Association* 118, no. 543: 2059–2071.

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, no. 3: 765–793.

Shi, C., J. Zhu, Y. Shen, S. Luo, H. Zhu, and R. Song. 2024. "Off-Policy Confidence Interval Estimation With Confounded Markov Decision Process." *Journal of the American Statistical Association* 119, no. 545: 273–284.

Silver, D., A. Huang, C. J. Maddison, et al. 2016. "Mastering the Game of Go With Deep Neural Networks and Tree Search." *Nature* 529, no. 7587: 484–489.

Sun, K., L. Kong, H. Zhu, and C. Shi. 2024. "Arma-Design: Optimal Treatment Allocation Strategies for a/B Testing in Partially Observable Time Series Experiments." arXiv preprint arXiv:2408.05342.

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. MIT Press.

Tang, X., Z. Qin, F. Zhang, et al. 2019. "A Deep Value-Network Based Approach for Multi-Driver Order Dispatching." In *Proceedings of the 25th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*, 1780–1790.

Tsiatis, A. A., M. Davidian, S. T. Holloway, and E. B. Laber. 2019. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC Press.

Uehara, M., M. Imaizumi, N. Jiang, N. Kallus, W. Sun, and T. Xie. 2022. "Finite Sample Analysis of Minimax Offline Reinforcement Learning: Completeness, Fast Rates and First-Order Efficiency." https://arxiv.org/abs/2102.02981.

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

Ueno, T., S. ichi Maeda, M. Kawanabe, and S. Ishii. 2011. "Generalized Td Learning." *Journal of Machine Learning Research* 12: 1977–2020. https://api.semanticscholar.org/CorpusID:2634909.

van Hasselt, H., A. Guez, and D. Silver. 2015. "Deep Reinforcement Learning With Double Q-Learning." CoRR, abs/1509.06461. http://arxiv.org/abs/1509.06461.

Wang, J., Z. Qi, and R. K. Wong. 2023. "Projected State-Action Balancing Weights for Offline Reinforcement Learning." *Annals of Statistics* 51, no. 4: 1639–1665.

Wang, J., C. Shi, J. D. Piette, J. R. Loftus, D. Zeng, and Z. Wu. 2025. "Counterfactually Fair Reinforcement Learning via Sequential Data Preprocessing." arXiv Preprint arXiv:2501.06366.

Wang, L., J. Zhou, and A. Qu. 2011. "Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis." *Biometrics* 68, no. 2: 353–360. https://doi.org/10.1111/j.1541-0420.2011.01678.x.

Wang, Y.-W., H.-C. Yang, Y.-H. Chen, and C.-Y. Guo. 2024. "Generalized Estimating Equations Boosting (Geeb) Machine for Correlated Data." *Journal of Big Data* 11: 1–19. https://api.semanticscholar.org/CorpusID:267061583.

Watkins, C. J., and P. Dayan. 1992. "Q-Learning." *Machine Learning* 8: 279–292.

Wu, X., H. Chen, J. Wang, L. Troiano, V. Loia, and H. Fujita. 2020. "Adaptive Stock Trading Strategies With Deep Reinforcement Learning Methods." *Information Sciences* 538: 142–158. https://api.semanticscholar.org/CorpusID:224967139.

Xu, Z., Z. Li, Q. Guan, 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*, 905–913.

Yang, H., X.-Y. Liu, S. Zhong, and A. Walid. 2020. "Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy." *Proceedings of the First ACM International Conference on AI in Finance*. https://api.semanticscholar.org/CorpusID:225206052.

Yang, W., L. Zhang, and Z. Zhang. 2022. "Toward Theoretical Understandings of Robust Markov Decision Processes: Sample Complexity and Asymptotics." *Annals of Statistics* 50, no. 6: 3223–3248.

Zeger, S. L., and K.-Y. Liang. 1986. "Longitudinal Data Analysis for Discrete and Continuous Outcomes." *Biometrics* 42, no. 1: 121–130. http://www.jstor.org/stable/2531248.

Zhang, L., Y. Peng, J. Liang, W. Yang, and Z. Zhang. 2023. "Estimation and Inference in Distributional Reinforcement Learning." arXiv preprint arXiv:2309.17262.

Zhong, H., X. Deng, E. X. Fang, Z. Yang, Z. Wang, and R. Li. 2020. "Risk-Sensitive Deep RL: Variance-Constrained Actor-Critic Provably Finds Globally Optimal Policy." arXiv Preprint arXiv:2012.14098.

Zhou, W., Y. Li, R. Zhu, and A. Qu. 2023. "Distributional Shift-Aware Off-Policy Interval Estimation: A Unified Error Quantification Framework." arXiv Preprint arXiv:2309.13278.

Zhou, W., R. Zhu, and A. Qu. 2024. "Estimating Optimal Infinite Horizon Dynamic Treatment Regimes via PT-Learning." *Journal of the American Statistical Association* 119, no. 545: 625–638.

Zhu, J., X. Zhou, J. Yao, et al. 2025. "Semi-Pessimistic Reinforcement Learning." arXiv preprint arXiv:2505.19002.

Zou, S., T. Xu, and Y. Liang. 2019. "Finite-Sample Analysis for Sarsa With Linear Function Approximation." *Advances in Neural Information Processing Systems* 32. https://papers.neurips.cc/paper_files/paper/2019/file/9f9e8cba3700df6a947a8cf91035ab84-Paper.pdf.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Appendix S1:** Supporting information.