Supplementary Materials for Generalized Fitted Q Iterations with Application in Cluster Data

A Proofs

A.1 Technical Conditions

To establish the asymptotic properties and the regret bound for estimators from Algorithm 3, we impose the following assumptions.

(A1) (Realizability) Assume the environment is a linear MDP (Xie et al., 2023) where both the reward function and the transition dynamics are linear in a known feature map $\phi(s, a)$, i.e., $\mathcal{T}(s'|a, s) = \phi(a, s)^\top \mu(s')$ and $\mathcal{R}(a, s) = \phi(a, s)^\top \omega$.

(A2) (Stability) The matrix

$$M^{-1}\lambda_{\min}\mathbb{E}\left[\Phi^\top(\mathbf{A}, \mathbf{S})\phi(\mathbf{A}, \mathbf{S}) - \gamma\Phi^\top(\mathbf{A}, \mathbf{S})\phi(\pi^*(\mathbf{S}'), \mathbf{S}')\right]$$

is uniformly bounded away from zero.

(A3) (Uniqueness) π^* is unique.

(A4) (FQI iterations) The maximum number of iterations K in GFQI satisfies $\log(N) \ll K = O(N^{c'})$ for any $c' > 0$.

(A5) (Behavior policy) The data is generated by a Markov policy.

(A6) (Value smoothness) Let $\pi(\beta)$ be the greedy policy derived by $\phi(a, s)^\top \beta$. Then the expected cumulative reward for $\pi(\beta)$ has third-order derivative w.r.t β.

Assumption **(A1)** indicates that $\forall \pi$, there exits a $\beta^\pi \in \mathcal{R}^p$, such that the Q function $Q^\pi(a, s) = \phi(a, s)^\top \beta^\pi$. Assumption **(A2)** is crucial for guaranteeing a unique solution to the optimal GEE and is commonly adopted in the literature on linear-sieve Q-function approximators (Ertefaie and Strawderman, 2018; Luckett et al., 2020; Li et al., 2022). The stability assumption is weaker than the Bellman completeness assumption and is sufficient for FQI under Assumption **(A1)** (Perdomo et al., 2022). Assumption **(A3)** is needed for establishing the limiting distribution of $\widehat{\beta}$ computed via FQI and has been similarly adopted in the statistical literature (Ertefaie and Strawderman, 2018; Luckett et al., 2020). Assumption **(A4)** requires that the maximum number of iterations in FQI to satisfy certain rate as sample size goes to infinity. Assumption **(A5)** is automatically satisfied in randomized studies where the behavior policy is usually a constant function of the state. Assumption **(A6)** allows us to analyze the change in regret as β varies near the optimal β^*.

A.2 Proof of Theorem 1

Without loss of generality, we assume that the numbers of subjects for all clusters are the same and denote this number as M.

Theorem 1. *(Formal statement) Suppose Assumption (**A1**)-(**A5**) are satisfied. $\widehat{\beta}$ computed by Algorithm 3 has the following properties:*

1. *The asymptotic distribution of $\sqrt{MN}(\widehat{\beta}-\beta^*)$ is $\mathcal{N}(\mathbf{0}, W^{-1}\Sigma W^{-1\top})$ where*

$$W(\boldsymbol{\Phi}) = \frac{1}{M}\mathbb{E}\left[\boldsymbol{\Phi}(\mathbf{A},\mathbf{S})\left\{\phi(\mathbf{A},\mathbf{S}) - \gamma\phi(\pi^*(\mathbf{S}'),\mathbf{S}')\right\}\right],$$

   and $\Sigma(\boldsymbol{\Phi}) = \frac{1}{M}\mathbb{E}(\boldsymbol{\Phi}\mathbf{V}^*\boldsymbol{\Phi}^\top)$.

2. *When the correlation structure of the TD errors is correctly specified, and the estimator $\widehat{\Phi}^*(\mathbf{A},\mathbf{S})$ converges to $\Phi^*(\mathbf{A},\mathbf{S})$ with a rate at least $O(N^{-b}\log^{-1}(N))$ for some $b>0$, $\widehat{\beta}$ reaches the minimal asymptotic variance $W(\Phi^*)^{-1}$ among the class of solutions to* (5)*.*

Proof of the asymptotic normality. We first show the asymptotic normality of the estimated Q-function. To simplify notation, we focus on a finite MPD, where $\mathcal{S}$ and $\mathcal{A}$ are discrete. In FQI, we iteratively update the Q-function according to the formula,

$$\mathbf{0} = \sum_{i,t}\boldsymbol{\Phi}(A_t^{(i)}, S_t^{(i)})\boldsymbol{\delta}(A_t^{(i)}, S_t^{(i)}, R_t^{(i)}, S_{t+1}^{(i)}; \beta^{(k)}, \beta^{(k-1)}).$$

At the k th iteration, we define the population-level Q-function

$$Q^{(k+1),*}(a,s) = \mathcal{R}(a,s) + \gamma\sum_{s'}\max_{a'} Q^{(k)}\left(a', s'\right)\mathcal{T}\left(s' \mid a, s\right). \tag{1}$$

According to the Bellman optimality equation,

$$Q^{*}(a,s) = \mathcal{R}(a,s) + \gamma\sum_{s'}\max_{a'} Q^{*}\left(a', s'\right)\mathcal{T}\left(s' \mid a, s\right).$$

It follows that

$$\begin{aligned}
&\sup_{a,s}\left|Q^*(a,s) - Q^{(k+1)}(a,s)\right| \\
&\le \sup_{a,s}\left|Q^{(k+1),*}(a,s) - Q^{(k+1)}(a,s)\right| + \sup_{a,s}\left|Q^*(a,s) - Q^{(k+1),*}(a,s)\right| \\
&\le \sup_{a,s}\left|Q^{(k+1),*}(a,s) - Q^{(k+1)}(a,s)\right| + \gamma\sup_{a,s}\left|Q^*(a,s) - Q^{(k)}(a,s)\right|.
\end{aligned}$$

Iteratively applying this inequality for $k = K, K-1, \cdots, 1$, we obtain that

$$\begin{aligned}
&\sup_{a,s}\left|Q^*(a,s) - \widehat{Q}(a,s)\right| \\
&\le \sum_k \gamma^{K-k}\sup_{a,s}\left|Q^{(k+1),*}(a,s) - Q^{(k+1)}(a,s)\right| + \gamma^{K+1}\sup_{a,s}\left|Q^*(a,s) - Q^{(0)}(a,s)\right|.
\end{aligned} \tag{2}$$

As K diverges to infinity, the second term on the RHS decays to zero. The first term is upper bounded by

$$\frac{1}{1-\gamma}\sup_{a,s,k}\left|Q^{(k+1),*}(a,s) - Q^{(k+1)}(a,s)\right|. \tag{3}$$

It remains to show this term decays to zero as the sample size approaches to infinity. Let $\beta^{(k)}$ denote the estimated regression coefficients such that $Q^{(k)}(a,s) = \phi_L^\top(a,s)\beta^{(k)}$. Define

$$\beta^{(k+1),*} = \left\{\frac{1}{M}\mathbb{E}\left[\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\phi\left(S_t^{(i)}, A_t^{(i)}\right)\right]\right\}^{-1}\left\{\frac{1}{M}\mathbb{E}\left[\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\mathbf{Q}_i^{(k+1),*}\left(S_t^{(i)}, A_t^{(i)}\right)\right]\right\}.$$

We claim that there exist some constants $C, \bar{C} > 0$ such that

$$\sup_{a,s}\max_{k\in\{0,\cdots,j\}}\left\{\left|Q^{(k)}(a,s) - Q^{(k),*}(a,s)\right|, \left|Q^{(k)}(a,s)\right|\right\} \leq C \text{ and } \max_{k\in\{0,\cdots,j\}}\left\{\left\|\beta^{(k)} - \beta^{(k),*}\right\|_2, \left\|\beta^{(k)}\right\|_2\right\} \leq \bar{C}, \quad (4)$$

with probability at least $1-(j+1)/N$, for sufficiently large N. The values of C and $\bar{C}$ will be specified later.

We will prove this assertion by induction. Since $\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\phi$'s are independent, we can show that

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{i,t}\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\phi\left(S_t^{(i)}, A_t^{(i)}\right) - \mathbb{E}\left(\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\phi\left(S_t^{(i)}, A_t^{(i)}\right)\right)\right\|_2 \\ &\leq c\sqrt{N^{-1}\log(N)},\end{aligned} \quad (5)$$

for some constant $c > 0$, with probability at least $1 - N^{-1}$. Under Assumption (**A2**), $M^{-1}\lambda_{\min}\left\{\mathbb{E}\left(\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\phi\left(S_t^{(i)}, A_t^{(i)}\right)\right)\right\}$ is uniformly bounded away from zero. On the event set defined by (5), for sufficiently large N, there exists some $\bar{c} > 0$ such that

$$\lambda_{\min}\left\{\frac{1}{MN}\sum_{i,t}\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\phi\left(S_t^{(i)}, A_t^{(i)}\right)\right\} \geq \bar{c}. \quad (6)$$

When $k = 0$, this posit assertion in (4) holds as long as $\sup_{a,s}\left|Q^{(0)}(a,s)\right| \leq C$ and $\left\|\beta^{(0)}\right\|_2 \leq \bar{C}$. Suppose the assertion holds for $k = 0, 1, \cdots, J$. We aim to prove this assertion holds for $k = J+1$. Since the reward is uniformly bounded, so is $\sup_{a,s}|\mathcal{R}(a,s)|$. We will choose C to be such that $C \geq 2(1-\gamma)^{-1}\sup_{a,s}|\mathcal{R}(a,s)|$. It follows from (1) that $\sup_{a,s}\left|Q^{(k+1),*}(a,s)\right| \leq (1+\gamma)C/2$. Under Assumption (**A1**), $Q^{(k+1),*} \in \mathcal{F}_L$. Therefore, it is bounded and the numerator in the RHS is of order $O(1)$. We have that there exist some constants $c, C > 0$ such that $\left\|\beta^{(k+1),*}\right\|_2 \leq c$ and $Q^{(k+1),*}(A,S) = \phi^\top(A,S)\beta^{(k+1),*}$. We choose $\bar{C}$ to be such that $\bar{C} \geq 2cC$. As such, it suffices to show that on the set defined in (6), the estimation errors $\sup_{S,A}\left|Q^{(k+1)}(S,A) - Q^{(k+1),*}(S,A)\right| \leq (1-\gamma)C/2$ and $\left\|\beta^{(k+1)} - \beta^{(k+1),*}\right\|_2 \leq cC$, with probability at least $1 - N^{-1}$.

By definition, we have

$$\begin{aligned}\beta^{(k+1)} - \beta^{(k+1),*} = &\left\{\frac{1}{MN}\sum_{i,t}\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\phi\left(A_{i,t}, S_{i,t}\right)\right\}^{-1} \\ &\times\left[\frac{1}{MN}\sum_{i,t}\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\left\{R_t^{(i)} + \gamma\max_a\phi\left(a, S_{t+1}^{(i)}\right)\beta^{(k)} - \phi\left(A_t^{(i)}, S_t^{(i)}\right)\beta^{(k+1),*}\right\}\right].\end{aligned}$$

On the set defined in (6), $\left\|\beta^{(k+1)} - \beta^{(k+1),*}\right\|_2$ is upper bounded by

$$\begin{aligned}&\sup_{\beta_0^{(k)}\in\mathcal{B}^{(k)}(2cC)}\left\|\frac{1}{MN\bar{c}}\sum_{i=1}^{M}\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})\left\{R_t^{(i)} + \gamma\max_a\phi\left(a, S_{t+1}^{(i)}\right)\beta_0^{(k)} - \phi\left(A_t^{(i)}, S_t^{(i)}\right)\beta^{(k+1),*}\right\}\right\|_2 \\ \leq &\sup_{\beta_0^{(k)}\in\mathcal{B}^{(k)}(2cC)}\frac{1}{MN}\sum_{i,t}\sum_{m}\left\|\mathbf{\Phi}_{i,t,(m)}\left\{R_t^{(i,m)} + \gamma\max_a\phi^\top(a, S_{t+1}^{(i,m)})\beta_0^{(k)} - \phi(A_t^{(i,m)}, S_t^{(i,m)})^\top\beta^{(k+1),*}\right\}\right\|_2\end{aligned} \quad (7)$$

where $\mathcal{B}^{(k)}(2cC) = \{\beta_0^{(k)} \in \mathbb{R}^L : \|\beta_0^{(k)}\|_2 \leq 2cC\}$ and $\mathbf{\Phi}_{i,t,(m)}$ is the corresponding row in $\mathbf{\Phi}(A_t^{(i)}, S_t^{(i)})$ for mth subject.

Notice that the suprema in (7) are taken with respect to infinity many β's. As such, standard concentration inequalityies are not applicable to bound (7). Towards that end, we first take an ϵ-net of $\mathcal{B}^{(k)}(2cC)$ for some small $\epsilon > 0$, denoted by $\mathcal{B}^{k,\diamond}(2cC)$, such that for any $\beta \in \mathcal{B}^{(k)}(2cC)$, there exist some $\beta^* \in \mathcal{B}^{(k),\diamond}(2cC)$ that satisfies $\|\beta - \beta^*\|_2 \leq \epsilon$. The purpose of introducing some an ϵ-net is to approximate these sets by collections of finitely many β so that concentration inequalities are applicable to establish the upper bound. Set $\epsilon = CN^{-2}$. It follows from Lemma 2.2 of Mendelson et al. (2008) that there exist some $\mathcal{B}^{k,\diamond}(2cC)$ with number of elements $O(N)$. By the Lemma 16 in Li et al. (2022), $\phi(a,s)$ has bounded norm of $O(1)$. Note that the quantity after the suprema symbol is a Lipschitz continous function of $\beta^{(k)}$ with the Lipschitz constant upper bounded by $O(1)$. As such (7) can be approximated by

$$\sup_{\beta_0^{(k)} \in \mathcal{B}^*(2cC)} \frac{1}{MN} \sum_{i,t} \sum_{m} \left\| \mathbf{\Phi}_{i,t,(m)} \left\{ R_t^{(i,m)} + \gamma \max_a \phi^\top(a, S_{t+1}^{(i,m)}) \beta_0^{(k)} - \phi(A_t^{(i,m)}, S_t^{(i,m)})^\top \beta^{(k+1),*} \right\} \right\|_2 \tag{8}$$

with the approximation error given by $O(N^{-2})$. Using similar arguments as the Step 3 of the proof of Theorem 2 in Li et al. (2022), the quantities after the suprema symbol is upper bounded by $O(N^{-1/2}\log^{1/2}(N))$ with probability at least $1 - O(N^{-2})$. By the Bonferroni inequality, we have shown that the upper bound of $\|\beta^{(k+1)} - \beta^{(k+1,*)}\|_2$ is of order $O(N^{-2}) + O(N^{-1/2}\log^{1/2}(N))$ with probability at least $1 - O(N^{-1})$. For sufficiently large N, the assertion $\left\|\beta^{(k+1)} - \beta^{(k+1),*}\right\|_2 \leq cC/2$ is automatically satisfied. Similarly, we obtain that $\sup_{a,s} \left|Q^{(k+1)}(a,s) - Q^{(k+1),*}(a,s)\right| \preceq N^{-\bar{C}} + \sqrt{N^{-1}\log(N)} \ll (1-\gamma)C/2$. The assertion is thus proven.

Under the given conditions on K, the maximum number of iterations, we obtain that both $\widehat{Q}^{(k)}$ and $\widehat{\beta}^{(k)}$ are uniformly bounded WPA1. In addition, the estimation error (3) decays to zero, WPA1. According to (2), we have

$$\max_{K/2 < k \leq K} \sup_{a,s} \left| Q^*(a,s) - Q^{(k+1)}(a,s) \right| \xrightarrow{p} 0.$$

By Assumption (**A3**), we obtain that $\arg\max_a \phi_L^\top(a,s)\beta^{(k+1)} = \pi^*(s)$, for any large enough k, WPA1. As such, $\beta^{(k)}$ is the solution to the following estimating equations:

$$\mathbf{0} = \sum_{i,t} \mathbf{\Phi}(A_t^{(i)}, S_t^{(i)}) \left\{ R_t^{(i)} + \gamma \phi(\pi^*(S_{t+1}^{(i)}, S_{t+1}^{(i)})\beta - \phi(A_t^{(i)}, S_t^{(i)})\beta \right\}.$$

Under Assumption (**A2**), WAP1, it follows that

$$\begin{aligned} \left(\beta^{(k)} - \beta^*\right) = & \left\{ \sum_{i,t} \mathbf{\Phi}(A_t^{(i)}, S_t^{(i)}) \left(\phi\left(A_t^{(i)}, S_t^{(i)}\right) - \gamma \phi(\pi^*(S_{t+1}^{(i)}), S_{t+1}^{(i)} \right) \right\}^{-1} \\ & \left[\sum_{i,t} \mathbf{\Phi}(A_t^{(i)}, S_t^{(i)}) \left\{ R_t^{(i)} + \gamma \phi\left(\pi^*\left(S_{t+1}^{(i)}\right), S_{t+1}^{(i)}\right)\beta^* - \phi\left(A_t^{(i)}, S_t^{(i)}\right)\beta^* \right\} \right] \end{aligned} \tag{9}$$

Denote

$$\widehat{W} = \frac{1}{MN} \sum_{i,t} \mathbf{\Phi}(A_t^{(i)}, S_t^{(i)}) \left\{ \phi\left(A_t^{(i)}, S_t^{(i)}\right) - \gamma \phi\left(\pi^*\left(S_{t+1}^{(i)}\right), S_{t+1}^{(i)}\right) \right\}$$

Similar to Lemma 3 of Shi et al. (2021), we can show that $\|\widehat{W}^{-1} - W^{-1}\|_2 = O_p(N^{-1/2}\log(N))$ and $\|W^{-1}\|_2 > 0$. Similar to the step 3 of the proof of Theorem 2 in Li et al. (2022), the norm of the numerator in (9) is $O(N^{1/2}\log(N))$. Therefore, WPA1,

$$\widehat{\beta} - \beta^* = \frac{W^{-1}}{MN} \sum_{i,t} \mathbf{\Phi}(A_t^{(i)}, S_t^{(i)}) \delta_t^{(i)*} + O(N^{-1}\log(N)).$$

Proof of the efficiency property.

We now prove that for any $\Phi(\mathbf{A}, \mathbf{S})$, the asymptotic variance of the solution to (5) satisfies the following inequality:

$$W^{-1}(\mathbf{\Phi})\Sigma(\mathbf{\Phi})W^{-1\top}(\mathbf{\Phi}) \geq W^{-1}(\mathbf{\Phi}^*)\Sigma(\mathbf{\Phi}^*)W^{-1\top}(\mathbf{\Phi}^*) = W^{-1}(\mathbf{\Phi}^*).$$

We remove the dependence on $\mathbf{A}$ and $\mathbf{S}$ for simplification in notation and denote $\mathbb{E}(\phi(\pi^*(\mathbf{S}'), \mathbf{S}') \mid \mathbf{A}, \mathbf{S})$ as ϕ'. It is equivalent to show

$$a^\top \Sigma^{-1/2}(\boldsymbol{\Phi}) W(\boldsymbol{\Phi}) W^{-1/2}(\boldsymbol{\Phi}^*) W^{-1/2\top}(\boldsymbol{\Phi}^*) W^\top(\boldsymbol{\Phi}) \Sigma^{-1/2\top}(\boldsymbol{\Phi}) a \le \|a\|_2^2$$

for any $a \in \mathbb{R}^p$. By Cauchy–Schwarz inequality, the LHS

$$\begin{aligned}
&\|a^\top \mathbb{E}^{-1/2}\left(\boldsymbol{\Phi}\mathbf{V}^*\boldsymbol{\Phi}^\top\right)\mathbb{E}(\boldsymbol{\Phi}\left(\phi - \gamma\phi'\right))\mathbb{E}^{-1/2}\left\{\left(\phi - \gamma\phi'\right)\mathbf{V}^{*-1}\left(\phi - \gamma\phi'\right)^\top\right\}\|_2^2 \\
&\le \|a^\top \mathbb{E}^{-1/2}\left(\boldsymbol{\Phi}\mathbf{V}^*\boldsymbol{\Phi}^\top\right)\mathbb{E}^{1/2}(\boldsymbol{\Phi}V^*\boldsymbol{\Phi})\mathbb{E}^{1/2}\left\{\left(\phi - \gamma\phi'\right)\mathbf{V}^{*-1}\left(\phi - \gamma\phi'\right)^\top\right\}\mathbb{E}^{-1/2}\left\{\left(\phi - \gamma\phi'\right)\mathbf{V}^{*-1}\left(\phi - \gamma\phi'\right)^\top\right\}\|_2^2 \\
&= \|a\|_2^2.
\end{aligned}$$

Therefore, we have shown that the solution to the following GEE

$$\begin{aligned}
&\mathbf{0} \\
&= \sum_{i,t}\left\{\phi(A_t^{(i)}, S_t^{(i)}) - \gamma\mathbb{E}\left(\phi(\pi^*(S_{t+1}^{(i)}), S_{t+1}^{(i)}) \mid A_t^{(i)}, S_t^{(i)}\right)\right\}\mathbf{V}_i^{-1}\left\{R_t^{(i)} + \gamma\phi(\pi^*(S_{t+1}^{(i)}, S_{t+1}^{(i)})\beta - \phi(A_t^{(i)}, S_t^{(i)})\beta\right\}.
\end{aligned} \tag{10}$$

has the minimal variance among the class of estimators computed by solving (5). The results is similar to Theorem 6 in Ueno et al. (2011).

In GFQI, the unknown optimal policy is replaced by the current greedy policy $\pi^{(k)}$ at each iteration. Using similar arguements as previous proof of the robustness properties, we can show that $\pi^{(k)}$ converges to π^*. We next prove that the estimator from GFQI is also the solution to (10). Without loss of generality, suppose K is an even number. Suppose $\pi^{(k)}$ converges to π^* after $K/2$th iteration WPA1, then for $k = K/2, \ldots, K$, $\beta^{(k+1)}$ equals

$$\begin{aligned}
\beta^{(k+1)} - \beta^* = &\left\{\sum_{i,t}\widehat{\boldsymbol{\Phi}}(A_t^{(i)}, S_t^{(i)})\phi\left(A_t^{(i)}, S_t^{(i)}\right)\right\}^{-1} \\
&\left[\sum_{i,t}\widehat{\boldsymbol{\Phi}}(A_t^{(i)}, S_t^{(i)})\left\{R_t^{(i)} + \gamma\phi\left(\pi^*\left(S_{t+1}^{(i)}\right), S_{t+1}^{(i)}\right)\beta^{(k)} - \phi(A_t^{(i)}, S_t^{(i)})\beta^*\right\}\right]
\end{aligned} \tag{11}$$

where $\widehat{\boldsymbol{\Phi}}(\mathbf{A}, \mathbf{S}) = \left\{\phi(\mathbf{A}, \mathbf{S}) - \gamma\widehat{\mathbb{E}}(\phi(\pi^*(\mathbf{S}'), \mathbf{S}') \mid \mathbf{A}, \mathbf{S})\right\}\widehat{\mathbf{V}}^{-1}$, $\widehat{\mathbb{E}}(\phi(\pi^*(\mathbf{S}'), \mathbf{S}') \mid \mathbf{A}, \mathbf{S})$ is some consistent estimator of $\mathbb{E}(\phi(\pi^*(\mathbf{S}'), \mathbf{S}') \mid \mathbf{A}, \mathbf{S})$, and $\widehat{\mathbf{V}}$ is some consistent estimator $\mathbf{V}$. Applying (11) for $k = K, K-1, \ldots, K/2+1$, we obtain that

$$\widehat{\beta} - \beta^* = \widehat{\mathcal{L}}^{K/2}(\beta^{(K/2-1)} - \beta^*) + \sum_{k=0}^{K/2}\widehat{\mathcal{L}}^k\widehat{l},$$

where

$$\widehat{\mathcal{L}} = \gamma\left\{\sum_{i,t}\widehat{\boldsymbol{\Phi}}(A_t^{(i)}, S_t^{(i)})\phi\left(A_t^{(i)}, S_t^{(i)}\right)\right\}^{-1}\left\{\sum_{i,t}\widehat{\boldsymbol{\Phi}}(A_t^{(i)}, S_t^{(i)})\phi\left(\pi^*\left(S_{t+1}^{(i)}\right), S_{t+1}^{(i)}\right)\right\}$$

and

$$\widehat{l} = \left\{\sum_{i,t}\widehat{\boldsymbol{\Phi}}(A_t^{(i)}, S_t^{(i)})\phi\left(A_t^{(i)}, S_t^{(i)}\right)\right\}^{-1}\left[\sum_{i,t}\widehat{\boldsymbol{\Phi}}(A_t^{(i)}, S_t^{(i)})\left\{R_t^{(i)} + \gamma\phi\left(\pi^*\left(S_{t+1}^{(i)}\right), S_{t+1}^{(i)}\right)\beta^* - \phi(A_t^{(i)}, S_t^{(i)})\beta^*\right\}\right].$$

Similar to Lemma 3 of Shi et al. (2021), under Assumption (**A2**), we can show that $\|\widehat{\mathcal{L}}\|_2 \le \bar{\gamma}$ for some $\bar{\gamma} < 1$, WPA1. Under the given conditions on K, we obtain that

$$\widehat{\beta} - \beta^* = (I - \widehat{\mathcal{L}})^{-1}\widehat{l} + O(N^{-b}),$$

for some $b > 1/2$ with probability at least $1 - O(N)^{-1}$. Notice that

$$(I - \widehat{\mathcal{L}})^{-1}\widehat{l} = \left\{\sum_{i,t} \widehat{\mathbf{\Phi}}(A_t^{(i)}, S_t^{(i)}) \left(\phi\left(A_t^{(i)}, S_t^{(i)}\right) - \gamma\phi(\pi^*(S_{t+1}^{(i)}), S_{t+1}^{(i)})\right)\right\}^{-1}$$
$$\left[\sum_{i,t} \widehat{\mathbf{\Phi}}(A_t^{(i)}, S_t^{(i)}) \left\{R_t^{(i)} + \gamma\phi\left(\pi^*\left(S_{t+1}^{(i)}\right), S_{t+1}^{(i)}\right)\beta^* - \phi\left(A_t^{(i)}, S_t^{(i)}\right)\beta^*\right\}\right].$$

The assertion follows using similar arguments to the proof of the robustness property as long as $\widehat{\mathbf{\Phi}}(\mathbf{A}, \mathbf{S})$ converges to $\left\{\phi(\mathbf{A}, \mathbf{S}) - \gamma\widehat{\mathbb{E}}(\phi(\pi^*(\mathbf{S}'), \mathbf{S}') \mid \mathbf{A}, \mathbf{S})\right\}\widehat{\mathbf{V}}^{-1}$ with a rate at least $O(N^{-b}\log^{-1}(N))$.

A.3 Proof of Theorem 2

Theorem 2. *(Formal statement) Suppose Assumption **(A1)-(A6)** are satisfied. The regret of the estimated optimal policy is given by*

$$-\frac{1}{2}\mathrm{tr}(\mathit{Var}(\widehat{\beta})H) + O(N^{-3/2})$$

where $H = \frac{\partial^2\mathcal{V}(\pi(\beta))}{\partial\beta\partial\beta^\top}\Big|_{\beta=\beta^*}$ *and* $\mathcal{V}(\pi(\beta)) = \sum_s V^{\pi(\beta)}(s)\rho(s)$.

We next prove Theorem 2. Consider the regret w.r.t. the optimal policy. By Taylor expansion and the asymptotic result from Theorem 1:

$$\begin{aligned}&\mathbb{E}(V(\beta^*) - V(\widehat{\beta}))\\ &= -\frac{1}{2}\mathbb{E}(\widehat{\beta} - \beta^*)^\top H(\widehat{\beta} - \beta^*) + O(\|\widehat{\beta} - \beta^*\|_2^3)\\ &= -\frac{1}{2}\mathrm{tr}(\mathrm{Var}(\widehat{\beta})H) + +O(N^{-3/2}).\end{aligned}$$

The proof is thus completed.

B Estimation of the TD Covariance

The accurate estimation of the covariance structure of TD errors is crucial for the convergence and stability of GEE-based algorithms. The conditional correlation matrix of TD errors, $\mathrm{Corr}(\boldsymbol{\delta}, \boldsymbol{\delta})$, can be decomposed as a combination of the conditional correlation matrices of $\mathbf{R}$, $\mathrm{Corr}(\mathbf{R}, \mathbf{R})$, and $\boldsymbol{f} := \max_a \gamma Q(a, \mathbf{S}') - Q(\mathbf{A}, \mathbf{S})$, $\mathrm{Corr}(\boldsymbol{f}, \boldsymbol{f})$, along with the correlation between $\mathbf{R}$ and $\boldsymbol{f}$, $\mathrm{Corr}(\mathbf{R}, \boldsymbol{f})$:

$$\begin{aligned}&\mathrm{Corr}(\boldsymbol{\delta}, \boldsymbol{\delta})\\ &= \mathrm{Var}^{-1/2}(\boldsymbol{\delta})\mathrm{Var}^{1/2}(\mathbf{R})\mathrm{Corr}(\mathbf{R})\mathrm{Var}^{1/2}(\mathbf{R})\mathrm{Var}^{-1/2}(\boldsymbol{\delta})\\ &+ \mathrm{Var}^{-1/2}(\boldsymbol{\delta})\mathrm{Var}^{1/2}(\boldsymbol{f})\mathrm{Corr}(\boldsymbol{f})\mathrm{Var}^{1/2}(\boldsymbol{f})\mathrm{Var}^{-1/2}(\boldsymbol{\delta})\\ &+ 2\mathrm{Var}^{-1/2}(\boldsymbol{\delta})\mathrm{Var}^{1/2}(\mathbf{R})\mathrm{Corr}(\mathbf{R}, \boldsymbol{f})\mathrm{Var}^{1/2}(\boldsymbol{f})\mathrm{Var}^{-1/2}(\boldsymbol{\delta}).\end{aligned}$$

Reward Correlation. In the clustered MDP, the reward's correlation does not infringe upon the state-transition dynamics. Any models about $\mathrm{Corr}(\mathbf{R})$ is proper and adheres to the principles of MDPs.

State-Value Discrepancy Correlation. The Delta Method enables us to linearize the impact of nonlinear transformations, such as the state-value discrepancy, around the expected value of states as long as the transitions to subsequent states, $S_{t+1}^{(i,m)}$, given the current state-action pairs, are characterized by minor deviations from their expected outcomes. To facilitate the differentiation of the Q-function with respect to state features, define the softmax approximation of $\max_a Q(s, a)$ as:

$$Q_{\mathrm{softmax}}(s) = \sum_a \frac{\exp(\beta^\top\Phi_L(a, s))}{\sum_b \exp(\beta^\top\Phi_L(a, s))}\beta^\top\Phi_L(a, s)$$

We formalize the approximation of the covariance of state-value discrepancies as follows:

$$\begin{aligned}&\mathrm{Cov}(f_t^{(i,m)}, f_{t'}^{(i,j')} \mid S_t^{(i,m)}, A_t^{(i,m)} S_{t'}^{(i',j')}, A_{t'}^{(i',j')}) \approx \left(\frac{\partial Q_{\mathrm{softmax}}(S)}{\partial S}\right)\Bigg|_{S=\overline{S}_{t+1}^{(i,m)}}^{\mathrm{T}} \\ &\mathrm{Cov}(S_{t+1}^{(i,m)}, S_{t'+1}^{(i,j')} \mid S_t^{(i,m)}, A_t^{(i,m)} S_{t'}^{(i',j')}, A_{t'}^{(i',j')}) \left.\frac{\partial Q_{\mathrm{softmax}}(S)}{\partial S}\right|_{S=\overline{S}_{t+1}^{(i,m)}}\end{aligned} \tag{12}$$

where $\overline{S}_{i,j,t+1} = \mathbb{E}(S_{t+1}^{(i,m)} \mid S_t^{(i,m)}, A_t^{(i,m)})$.

In the special case when the state space is of one dimension, if we further assume the marginal variance of the next state given the current state and action and that of the TD errors is constant, the conditional correlation structure of $\boldsymbol{f}$ simplifies significantly and can closely approximate the conditional correlation structure of the states:

$$\mathrm{Corr}(\boldsymbol{f}, \boldsymbol{f}) \approx \mathrm{Corr}(\mathbf{S}', \mathbf{S}'). \tag{13}$$

TD correlation. Given the $\mathrm{Corr}(\mathbf{R})$ and $\mathrm{Corr}(\boldsymbol{f})$, we can derive a "working" correlation matrix for the TD error by simply adding the structure for $\mathbf{R}$ and $\boldsymbol{f}$ together. For example, as in Section 6, if we believe that the correlation structure for the reward and the next state are all exchangeable, we can use and exchangeable "working" correlation structure, $(1-\rho)I + \rho$ where ρ is the correlation coefficient and I is an identity matrix, for the TD errors.

C Additional Details on Simulation

C.1 More on the Semi-synthetic Study

For CQL and DDQN, the Q function is approximated by a fully connected feed-forward neural network consisting of three sequential layers: the first with 1 input and 8 output features, the second with 8 input and 8 output features, and the third with 8 input and 2 output features, where the first two layers are followed by ReLU activation functions. The standard error in Figure 5 is calculated with respect to the random seed after running experiments 50 times. The simulation were conducted on Google Cloud using ”n2-standard” machine type (CPU only).

C.2 Simulation Study

The primary objective of this simulation study is to evaluate the policy regret under various correlation structures. The study examines the effect of different parameters such as the number of clusters, time horizon, and cluster size on the performance of the policies. The simulations were conducted on the school's high-performance computing cluster using CPUs.

Correlation Design. The correlation structures were designed to reflect different potential dependencies within the data. The state update was modeled as: $S_{t+1}^{(i,m)} = 0.5 S_t^{(i,m)}(2A_t^{(i,m)} - 1) + \beta_t^{(i)}$ where $\beta_t^{(i)} \sim \mathcal{N}(0, \sigma_1^2)$ with $\sigma_1^2 = 0.25$. The reward was expressed as: $R_t^{(i,m)} = 0.25(S_t^{(i,m)})^2(2A_t^{(i,m)} - 1) + S_t^{(i,m)} + \alpha_t^{(i)} + \epsilon_t^{(i,m)}$, with $\epsilon_{ijt} \sim \mathcal{N}(0, \sigma_2^2)$ and $\alpha_t^{(t)} \sim \mathcal{N}(0, \sigma_3^2)$, where $\sigma_2^2 = 0.25$ and $\sigma_3^2 = 4$. This correlation structure can be approximated by the exchangeable strucutre.

Simulation Parameters. The simulation was conducted by varying three key parameters:

- Number of Teams: 5, 10, 15, 20, 25, 30 (with cluster size = 5, horizon = 5)
- Time Horizon: 5, 10, 15, 20, 25, 30 (with number of clusters = 5, cluster size = 5)
- Team Size: 5, 10, 15, 20, 25, 30 (with number of clusters = 5, horizon = 5)

Each scenario was repeated 50 times to account for variability in the results.

Policies Compared. Comparison is made among the following algorithms/policies specified in Section 6. For CQL and DDQN, the Q function is approximated by a fully connected feed-forward neural network consisting

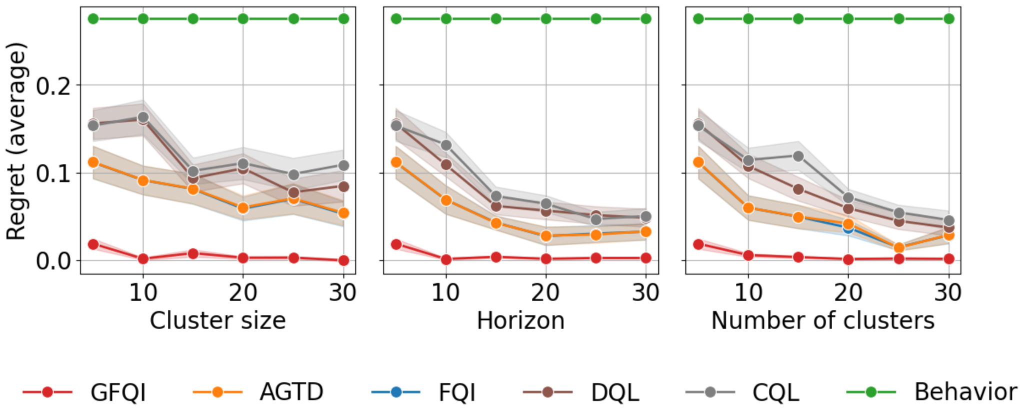

Figure 1: Change in regret of the average reward with varying cluster size, time horizon or number of clusters. The band represents the standard error calculated with respect to the random seed after running experiments 50 times.

of four sequential layers: the first with 1 input and 8 output features, the second and third with 8 input and 8 output features, and the forth with 8 input and 2 output features, where the first two layers are followed by ReLU activation functions.

Evaluation. Policies were evaluated by executing each to generate 100 trajectories of length 1000. The average of the average rewards and the cumulative discounted rewards across these trajectories were calculated.

We compare the average reward of the offline trained policies with a policy trained in an online, uncorrelated environment with Q-learning approach utilizing a neural network approximator with a 64-node hidden layer coupled with ReLu activation functions expect for the last layer. This Q approximator is trained for 5000 episodes, each with a horizon of 100. Evaluation mirrored the method described previously, with 50 repetitions conducted to ascertain the average performance metrics.

Results. The convergence patterns and regret metrics on the average reward with are depicted in the Figure 1. Similar to the results in Section 6, the GFQI estimator achieved the fastest convergence in regret and out performs the neural network based estimator (CQL and DDQN) in most of the cases.

References

Ertefaie, A. and Strawderman, R. L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105(4):963–977.

Li, M., Shi, C., Wu, Z., and Fryzlewicz, P. (2022). Testing stationarity and change point detection in reinforcement learning.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the american statistical association*.

Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N. (2008). Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28:277–289.

Perdomo, J. C., Krishnamurthy, A., Bartlett, P., and Kakade, S. (2022). A complete characterization of linear estimators for offline policy evaluation.

Shi, C., Zhang, S., Lu, W., and Song, R. (2021). Statistical Inference of the Value Function for Reinforcement Learning in Infinite-Horizon Settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):765–793.

Ueno, T., ichi Maeda, S., Kawanabe, M., and Ishii, S. (2011). Generalized td learning. *J. Mach. Learn. Res.*, 12:1977–2020.

Xie, C., Yang, W., and Zhang, Z. (2023). Semiparametrically efficient off-policy evaluation in linear markov decision processes. In *International Conference on Machine Learning*, pages 38227–38257. PMLR.