Supplement to "Doubly Inhomogeneous Reinforcement Learning"

This supplement is organised as follows. We first introduce some notations and additional technical conditions, and provide auxiliary lemmas and proofs of our major theorems. We next detail our experimental settings and implementations.

A Notations, conditions and proofs

A.1 Notations

For each cluster k, let θ_k^0 denote the oracle parameter of the state transition model after the most recent change. Let $T - \tau^{(k)}$ denote the cluster-specific most recent change point of the kth cluster. We have $\tau^{(k)} = \tau_i^*$ for any $i \in \mathcal{C}_k$.

We use θ_k^1 to denote the limit of

$$\arg\max_{\theta} \sum_{i \in \mathcal{C}_k} \sum_{t=T-\tau+1}^{T-\tau^{(k)}} \mathbb{E} \log p(S_{it}|S_{it-1}, A_{it-1}, \theta),$$

as $\tau - \tau^{(k)}$ diverges to infinity. Such a limit exists under the ergodicity assumption imposed in Section A.2. Define the signal strength of the temporal change as $s_{cp} = \min_k \|\theta_k^1 - \theta_k^0\|$. Similarly, define the signal strength of the subject heterogeneity as $s_{cl} = \min_{k_1 \neq k_2} \|\theta_{k_1}^0 - \theta_{k_2}^0\|$.

Let $\mathcal{B}_k^*$ denote the Bellman operator such that for any Q-function Q, $\mathcal{B}_k^* Q$ denotes another function given by

$$\mathcal{B}_k^* Q(s,a) = \mathbb{E}\Big[\max_{a^*} Q(S_{i,T+1}, a^*)|S_{i,T} = s, A_{i,T} = a\Big],$$

for any $i \in \mathcal{C}_k$. Additionally, let $\mathcal{Q}$ denote the Q-function class used to model the Q-function in FQI at each iteration.

Throughout the proof, we use c and C to denote some generic constants whose values are allowed to vary from place to place. Finally, we define a non-negative number $\epsilon = \epsilon(N,T)$ dependent on N and T such that (i) $\epsilon = 0$ when N diverges to infinity with T; (ii) $\epsilon = T^{-1}\log(NT)\sqrt{\log(\log(NT))}$ when N remains finite.

A.2 Additional Assumptions

Assumption 1 (Change point locations and cluster sizes)**.** *$\tau^{(1)}$, $\tau^{(2)}$, $\cdots$, $\tau^{(K)}$ are proportional to T, and $|\mathcal{C}_1|$, $|\mathcal{C}_2|$, $\cdots$, $|\mathcal{C}_K|$ are proportional to N.*

Assumption 2 (Compact space and parameter spaces)**.** *Both the state space and the parameter space Θ are compact.*

Assumption 3 (Differentiability and boundedness of the transition function)**.** *The transition function p is uniformly bounded away from zero, and is twice continuously differentiable with uniformly bounded second-order derivatives on Θ. In addition, there exists some constant $\lambda > 0$ such that*

$$\min_{a} \min_{\theta \in \Theta} \lambda_{\min} \left[- \int_{s,s'} \frac{\partial \log p(s'|s,a,\theta)}{\partial \theta \partial \theta^\top} ds ds' \right] \geq \lambda,$$

where $\lambda_{\min}[\bullet]$ denotes minimum eigenvalue of a given matrix.

Assumption 4 (Signal strength)**.** $s_{cl} \gg \max(T^{-1/4}, \log^{3/2}(NT)/(\sqrt{NT} s_{cp}))$ *and* $s_{cp} \gg N^{-1/2} T^{-1/4} \log^{3/2}(NT)$.

Assumption 5 (Geometric ergodicity)**.** *For each k and any $i \in \mathcal{C}_k$, the sequence $\{S_{i,t}\}_{t \geq T - \tau^{(k)}}$ is geometrically ergodic (see e.g., Bradley, 2005, for the detailed definition). In addition, when N is finite, the sequence $\{S_{i,t}\}_{(1-\bar{\epsilon})T - \tau_i^* \leq t < T - \tau_i^*}$ is geometrically ergodic for any i as well. Finally, the behavior policy is a double homogeneous policy, whose value is bounded away from zero.*

Assumption 6 (Number of iterations)**.** *The number of iterations in FQI is much larger than* $\log(NT)/\log(\gamma^{-1})$.

Assumption 7 (Bounded reward)**.** *There exists some constant $R_{\max} < \infty$ such that $|R_{i,t}| \geq R_{\max}$ almost surely.*

Assumption 8 (Completeness)**.** *For any $1 \leq k \leq K$ and Q that belongs to the Q-function class $\mathcal{Q}$, we have $\mathcal{B}_k^* Q \in \mathcal{Q}$.*

Assumption 9 (Q-function class)**.** *The estimated Q-function belongs to the VC type class (see e.g., Definition 2.1, Chernozhukov et al., 2014) with a finite VC index. Additionally, its envelop function is upper bounded by $R_{\max}/(1-\gamma)$.*

A.3 An Auxiliary Lemma

Recall that the log-likelihood function for a given set of indices $\mathcal{C}$ and a given time interval $[t_1, t_2]$ is given by

$$\ell(\theta; \mathcal{C}, [t_1, t_2]) = \frac{1}{|\mathcal{C}|(t_2 - t_1)} \sum_{i \in \mathcal{C}} \sum_{t=t_1+1}^{t_2} \log p(S_{it}|S_{it-1}, A_{it-1}, \theta).$$

We use ℓ_0 to denote its expectation, i.e., $\ell_0(\theta; \mathcal{C}, [t_1, t_2]) = \ell(\theta; \mathcal{C}, [t_1, t_2])$ for a given $\mathcal{C}$ and $[t_1, t_2]$. Recall that $\widehat{\theta}_{\mathcal{C}_k, [t_1, t_2]}$ denotes the conditional maximum likelihood estimator. The following lemma establishes the uniform consistency and rate of convergence of these estimators. Its proof is similar to the one for standard maximum likelihood estimators (see e.g., Casella and Berger, 2024). The difference lies in that we consider MDP settings with dependent data while allowing the number of parameters to diverge to infinity as well. For completeness, we provide its proof in the following subsection.

Lemma 1 (Uniform consistency and rate of convergence)**.** *Suppose MA, LHE, LSE and Assumptions 1-5 hold. Then the set of estimated parameters* $\{\widehat{\theta}_{\mathcal{C}_k,[t_1,t_2]} : t_2 - t_1 > \epsilon T, t_1 \geq T - \tau^{(k)}\}$ *converges uniformly to* θ_k^0*, and satisfies*

$$\|\widehat{\theta}_{\mathcal{C}_k,[t_1,t_2]} - \theta_k^0\| = O\left(\frac{\sqrt{\log(NT)}}{\sqrt{N(t_2-t_1)}}\right),$$

uniformly in all triplets (k, t_1, t_2) *such that* $t_2 - t_1 > \epsilon T$ *and* $t_1 \geq T - \tau^{(k)}$*, wpa1.*

A.4 Proof of Lemma 1

Proof. Notice that for any k, under the given assumptions, it follows from Jensen's inequality that $\ell_0(\theta; \mathcal{C}_k, [t_1, t_2])$ is uniquely maximised at θ_k^0 whenever $t_1 \geq T - \tau^{(k)}$ (see e.g., Hogg and Craig, 1995). The rest of the proof is divided into three steps. The first step is to establish the uniform consistency of log-likelihood function. The second step is to show the uniform consistency of the estimated parameters. The last step derives the rate of convergence.

A.4.1 Proof of Step 1

In this step, we aim to establish the uniform convergence of $\ell(\theta; \widehat{\mathcal{C}}_k, [t_1, t_2])$. That is, if Assumptions 1-5 hold, then

$$\max_{k,t_1,t_2} \sup_{\theta} |\ell(\theta; \widehat{\mathcal{C}}_k, [t_1, t_2]) - \ell_0(\theta; \widehat{\mathcal{C}}_k, [t_1, t_2])| \xrightarrow{P} 0,$$

where the first maximum is taken over all triplets (k, t_1, t_2) such that $t_2 - t_1 > \epsilon T$ and $t_1 \geq T - \tau^{(k)}$.

We will first prove the point-wise convergence of $\ell(\theta; \mathcal{C}_k, [t_1, t_2])$ at a given parameter value θ. Notice that the likelihood function $\ell(\theta; \mathcal{C}_k, [t_1, t_2])$ can be decomposed into the sum of the following three terms:

$$\begin{aligned}
&\frac{1}{|\mathcal{C}_k|(t_2-t_1)} \sum_{i\in\mathcal{C}_k} \sum_{t=t_1+1}^{t_2} [f(S_{it}|S_{it-1}, A_{it-1}, \theta) - \mathbb{E}_{(\bullet|)} f(S_{it}|S_{it-1}, A_{it-1}, \theta)] \\
+ \quad &\frac{1}{|\mathcal{C}_k|(t_2-t_1)} \sum_{i\in\mathcal{C}_k} \sum_{t=t_1+1}^{t_2} [\mathbb{E}_{(\bullet|)}[f(S_{it}|S_{it-1}, A_{it-1}, \theta) - \mathbb{E}\{f(S_{it}|S_{it-1}, A_{it-1}, \theta)|S_{it-1}\}] \\
+ \quad &\frac{1}{|\mathcal{C}_k|(t_2-t_1)} \sum_{i\in\mathcal{C}_k} \sum_{t=t_1+1}^{t_2} [\mathbb{E}\{f(S_{it}|S_{it-1}, A_{it-1}, \theta)|S_{it-1}\} - \mathbb{E}f(S_{it}|S_{it-1}, A_{it-1}, \theta)],
\end{aligned}$$

where f is a shorthand for $\log p$ and $\mathbb{E}_{(\bullet|)}$ is a shorthand for the conditional expectation of the next state given the current state-action pair. In the following, we will use concentration inequalities designed for martingales and β-mixing processes to bound the first two lines and the third line, respectively. Specifically:

1. For the first line, a key observation is that, under the Markov assumption, the first line forms a sum of martingale difference sequence (see e.g., Step 3 of the proof of Theorem 1 in Shi et al., 2022, for a detailed illustration). In addition, it follows from Assumption 3 that f is uniformly bounded away from infinity. As such, using the Azuma-Hoeffding's inequality[1], we can show

[1] see e.g., `https://galton.uchicago.edu/~lalley/Courses/386/Concentration.pdf`.

with probability at least $1 - O(N^{-1}T^{-3})$, for any $t_2 - t_1 > \epsilon T$ and any $1 \le k \le K$, the absolute value of the first line is upper bounded by $C\sqrt{N(t_2 - t_1)\log(NT)}$ with proper choice of the constant $C > 0$. Using Bonferroni's inequality, we can show that the above event holds uniformly for any t_1, t_2, k such that $t_2 - t_1 > \epsilon T$, with probability at least $1 - O(N^{-1}T^{-1})$.

2. Next, using similar arguments, we can show the supremum of the absolute value of the second line over the triplet (t_1, t_2, k) with the constraint that $t_2 - t_1 > \epsilon T$ is upper bounded by $O(\sqrt{N(t_2 - t_1)\log(NT)})$ with probability at least $1 - O(N^{-1}T^{-1})$.

3. Finally, consider the third line. Under Assumption 3, for any $i \in \mathcal{C}_k$, both the probability density function of $S_{i,T-\tau^{(k)}}$ and the stationary probability density function of $\{S_{i,t}\}_{t \ge T-\tau^{(k)}}$ are bounded away from zero and infinity; see e.g., Part 2 of the proof of Lemma E.2 of Shi et al. (2022). Together with Assumption 5, it follows from Lemma 1 of Meitz and Saikkonen (2019) that $\{S_{i,t}\}_{t \ge T-t^{(k)}}$ is exponentially β-mixing. Denote the resulting β-mixing coefficient by $\{\beta(q)\}_q$. Similar to Theorem 4.2 of Chen and Christensen (2015), we can show that, for any $t \ge 0$ and integer $1 < q < T$,

$$\begin{aligned}
&\max_{\substack{t_2 - t_1 > \epsilon T \\ t_1 \ge T - \tau^k}} \mathbb{P}\left(\Big| \sum_{i \in \mathcal{C}_k} \sum_{t=t_1+1}^{t_2} \mathbb{E}\{f(S_{it}|S_{it-1}, A_{it-1})|S_{it-1}\} - \mathbb{E}f(S_{it}|S_{it-1}, A_{it-1})| \ge 6t \right) \\
&= \max_{\substack{t_2 - t_1 > \epsilon T \\ t_1 \ge T - \tau^k}} \mathbb{P}\Big(\Big| \sum_{(i,t) \in I_r} \mathbb{E}\{f(S_{it}|S_{it-1}, A_{it-1})|S_{it-1}\} - \mathbb{E}f(S_{it}|S_{it-1}, A_{it-1})| \ge t \Big) \\
&+ O(1)\frac{|\mathcal{C}_k|(t_2 - t_1)}{q}\beta(q) + O(1)\exp\left(\frac{-t^2/2}{|\mathcal{C}_k|(t_2 - t_1)qM^2 + qMt/3} \right), \qquad \text{(A.1)}
\end{aligned}$$

where $O(1)$ denotes some positive constant, $I_r = \{q\lceil |\mathcal{C}_k|(t_2 - t_1)/q \rceil, q\lceil |\mathcal{C}_k|(t_2 - t_1)/q \rceil + 1, \ldots, |\mathcal{C}_k|(t_2 - t_1 + 1) - 1\}$ and $2f$ is uniformly upper bounded by M. Suppose $t > 5qM$. Notice that $|I_r| \le q$. We have

$$P\left(\Big| \sum_{(i,t) \in I_r} \mathbb{E}\{f(S_{it}|S_{it-1}, A_{it-1})|S_{it-1}\} - \mathbb{E}f(S_{it}|S_{it-1}, A_{it-1})| \ge t \right) = 0.$$

Under exponential β-mixing, we have $\beta(q) = O(\rho^q)$ for some positive constant $\rho < 1$. Set $q = -6\log(|\mathcal{C}_k|(t_2 - t_1))/\log\rho$, we obtain $|\mathcal{C}_k|(t_2 - t_1)\beta(q)/q = O(N^{-6}T^{-6})$ under Assumption 1. Set $t = \max\{4\sqrt{|\mathcal{C}_k|(t_2 - t_1)qM^2\log(|\mathcal{C}_k|(t_2 - t_1))}, 4qM\log(|\mathcal{C}_k|(t_2 - t_1))\}$, we obtain that

$$\frac{t^2}{2} \ge 8|\mathcal{C}_k|(t_2 - t_1)qM^2\log(|\mathcal{C}_k|(t_2 - t_1)) \text{ and } \frac{t^2}{2} \ge 6qM\log(|\mathcal{C}_k|(t_2 - t_1))\frac{t}{3} \text{ and } t \gg qM,$$

as either $N \to \infty$ or $T \to \infty$. Thus, it follows from (A.1) that the absolute value of the third line is upper bounded by $O(\sqrt{N(t_2 - t_1)}\log(NT))$ with probability at least $1 - O(N^{-6}T^{-6})$. By Bonferroni's inequality, we can show that this event holds uniformly for any triplet (t_1, t_2, k) such that $t_2 - t_1 > \epsilon T$ with probability $1 - O(N^{-1}T^{-1})$.

To summarize, we have shown that with probability at least $1 - O(N^{-1}T^{-1})$,

$$|\ell(\theta; \mathcal{C}_k, [t_1, t_2]) - \ell_0(\theta; \mathcal{C}_k, [t_1, t_2])| = O\left(\frac{\log(NT)}{\sqrt{N(t_2 - t_1)}} \right),$$

for all triplets (k, t_1, t_2) such that $t_2 - t_1 > \epsilon T$ and $t_1 \geq T - \tau^{(k)}$. This proves the pointwise convergence of the log-likelihood function as either N or T diverges to infinity.

To establish the uniform convergence, we need to show that for any $\varepsilon, \eta > 0$, there exists some integer $n(\varepsilon, \eta)$ such that for all $NT > n(\varepsilon, \eta)$,

$$\mathbb{P}[\max_{k, t_2 - t_1 > \varepsilon T} \sup_{\theta \in \Theta} |\ell(\theta; \mathcal{C}_k, [t_1, t_2]) - \ell_0(\theta; \mathcal{C}_k, [t_1, t_2])| > \varepsilon] < \eta. \tag{A.2}$$

Consider open balls of radius δ around $\theta \in \Theta$, i.e., $B(\theta, \delta) = \{\widetilde{\theta} : \|\theta - \widetilde{\theta}\| < \delta\}$ where $\| \bullet \|$ denotes the Euclidean norm. The union of these open balls contains Θ. Since Θ is a compact set, there exists a finite subcover, which we denote by $\{B(\theta^j, \delta), j = 1, \ldots, J\}$. It follows from the triangle inequality that

$$\begin{aligned}
&|\ell(\theta; \mathcal{C}_k, [t_1, t_2]) - \ell_0(\theta; \mathcal{C}_k, [t_1, t_2])| \\
&\leq |\ell(\theta; \mathcal{C}_k, [t_1, t_2]) - \ell(\theta^j; \mathcal{C}_k, [t_1, t_2])| &\text{(A.3)} \\
&+ |\ell(\theta^j; \mathcal{C}_k, [t_1, t_2]) - \ell_0(\theta^j; \mathcal{C}_k, [t_1, t_2])| &\text{(A.4)} \\
&+ |\ell_0(\theta^j; \mathcal{C}_k, [t_1, t_2]) - \ell_0(\theta; \mathcal{C}_k, [t_1, t_2])|. &\text{(A.5)}
\end{aligned}$$

For a given θ, set θ^j such that $\theta \in B(\theta^j, \delta)$. Under Assumption 3, the log-likelihood function is Lipschitz continuous. As such, (A.3) can be upper bounded by $L\delta$ for some constant $L > 0$. Similarly, (A.5) can be upper bounded by $L\delta$ as well. Finally, according to the pointwise convergence results, (A.4) converges to zero in probability as either N or T diverges to infinity. As such, by setting $\delta = \varepsilon/(3L)$ and letting $n(\varepsilon, \eta) \to \infty$, it is immediate to see that (A.2) holds. The proof for Step 1 is hence completed.

A.4.2 Proof of Step 2

In this step, we aim to show that for any positive $\varepsilon > 0$, the event $\max_{k, t_2 - t_1 > \epsilon T} \|\widehat{\theta}_{\mathcal{C}_k, [t_1, t_2]} - \theta_k^0\| \leq \varepsilon$ holds wpa1 as $T \to \infty$.

Consider the objective function $\ell_0(\theta, \mathcal{C}_k, [t_1, t_2])$. For any $t_1 \geq T - \tau^{(k)}$, under Assumption 5, for sufficiently large t_2, the distribution of the state S_{it_2} will converge to its limiting distribution. As discussed in Step 1 of the proof, the process $\{S_{i,t}\}_{t \geq T - \tau^{(k)}}$ is exponentially β-mixing. According to the definition of the β-mixing coefficient, we have

$$\beta(q) = \int_s \sup_{0 \leq \varphi \leq 1} \left| \mathbb{E}[\varphi(S_{iq+t_1}) | S_{it_1} = s] - \int \varphi(s)\mu(s)ds \right| \mu(s)ds,$$

where μ denotes the density function of the limiting distribution. Since p is well bounded away from zero and infinity, so are the marginal distributions of S_{it_1} and μ. Consequently, there exists some universal constant $C > 0$ such that

$$\int_s \sup_{0 \leq \varphi \leq 1} \left| \mathbb{E}[\varphi(S_{iq+t_1}) | S_{it_1}] - \int \varphi(s)\mu(s)ds \right| \leq C\beta(q).$$

Under exponentially β-mixing, this immediately implies that $\ell_0(\theta, \mathcal{C}_k, [t_1, t_2]) \to \ell_0^\infty(\theta, \mathcal{C}_k)$ whenever $t_1 \geq T - \tau^{(k)}$ and $t_2 - t_1 \geq \kappa T$, as $T \to \infty$. Here, $\ell_0^\infty(\theta, \mathcal{C}_k) = \int_s \mathbb{E} \log p(S_{it_1+1} | S_{it_1} = s, A_{it_1}, \theta)\mu(s)ds$ for any $i \in \mathcal{C}_k$. Moreover, the convergence is uniform in k, t_1 and t_2. This together with the proof for Step 1 yields the uniform convergence of $\ell(\theta, \mathcal{C}_k, [t_1, t_2])$ to $\ell_0^\infty(\theta, \mathcal{C}_k)$.

Under the regularity conditions in Assumption 3, for each k, $\ell_0^\infty(\theta, \mathcal{C}_k)$ is uniquely maximised at θ_k^0. In addition, it is a continuous function of θ. Since the parameter space is compact, $\ell_0^\infty(\theta_k^0, \mathcal{C}_k)$ is strictly larger than $\sup_{\|\theta-\theta_k^0\|\leq\varepsilon} \ell_0^\infty(\theta, \mathcal{C}_k)$. This together with the uniform consistency of $\ell(\theta, \mathcal{C}_k, [t_1, t_2])$ yields the uniform consistency of the estimated parameters. □

A.4.3 Proof of Step 3

Proof. By Taylor expansion, we obtain that

$$0 = \ell^{'}(\widehat{\theta}_{\mathcal{C}_k,[t_1,t_2]}; \mathcal{C}_k, [t_1, t_2]) = \ell^{'}(\theta_k^0; \mathcal{C}_k, [t_1, t_2]) + \int_0^1 \ell^{''}(t\widehat{\theta}_{\mathcal{C}_k,[t_1,t_2]} + (1-t)\theta_k^0; \mathcal{C}_k, [t_1, t_2])dt(\widehat{\theta}_{\mathcal{C}_k,[t_1,t_2]} - \theta_k^0),$$

for some θ_k^* lying on the line segment joining $\widehat{\theta}_{\mathcal{C}_k,[t_1,t_2]}$ and θ_k^0. It follows that

$$\|\widehat{\theta}_{\mathcal{C}_k,[t_1,t_2]} - \theta_k^*\| \leq \left[\lambda_{\min}[-\int_0^1 \ell^{''}(t\widehat{\theta}_{\mathcal{C}_k,[t_1,t_2]} + (1-t)\theta_k^0; \mathcal{C}_k, [t_1, t_2])dt]\right]^{-1} \|\ell^{'}(\theta_k^0; \mathcal{C}_k, [t_1, t_2])\|. \tag{A.6}$$

It remains to bound the two terms on the right-hand-side (RHS) of (A.6).

First, consider the first term on the RHS of (A.6). Under Assumption 3, both the transition function p and the marginal state density function is lower bounded by some constant $c > 0$. This together with the minimum eigenvalue assumption in Assumption 3 implies that

$$\begin{aligned}\min_\theta \lambda_{\min}\left[-\mathbb{E}\frac{\partial^2 \log p(S_{it}|S_{it-1}, A_{it-1}, \theta)}{\partial\theta\theta^\top}\right] &\geq \min_{a,\theta} \lambda_{\min}\left[-\mathbb{E}\frac{\partial^2 \log p(S_{i,t}|S_{it-1}, a, \theta)}{\partial\theta\theta^\top}\right] \\ &\geq c^2 \min_{a,\theta} \lambda_{\min}\left[-\int_{s,s'} \frac{\partial^2 \log p(s'|s, a, \theta)}{\partial\theta\theta^\top} dsds'\right]\end{aligned}$$

is well bounded away from zero. Similarly, the minimum eigenvalue of $-\mathbb{E}\ell^{''}(\theta; \mathcal{C}_k, [t_1, t_2])$ is uniformly bounded away from zero as well. In addition, similar to Lemma 1, we can show that the set of difference $\{\ell^{''}(\theta; \mathcal{C}_k, [t_1, t_2]) - \mathbb{E}\ell^{''}(\theta; \mathcal{C}_k, [t_1, t_2]) : t_1 \geq T - \tau^{(k)}, t_2, k, \theta\}$ converge uniformly to 0 in probability. As such, the minimum eigenvalue of $-\int_0^1 \ell^{''}(t\widehat{\theta}_{\mathcal{C}_k,[t_1,t_2]} + (1-t)\theta_k^0; \mathcal{C}_k, [t_1, t_2])dt$ is uniformly bounded away from zero, wpa1. Equivalently, the first term on the RHS of (A.6) is upper bounded by some positive constant.

As for the second term, notice that

$$\mathbb{E}_{(\bullet|)}\frac{\partial \log p(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0)}{\partial\theta} = 0,$$

whenever $t > T - \tau^{(k)}$ and $i \in \mathcal{C}_k$. As such, the second term forms a sum of martingale difference sequence. Using similar arguments in Step 1 of the proof of Lemma 1, it follows from the martingale concentration inequality that the supermum of the second term on the RHS of (A.6) over all triplets (t_1, t_2, k) such that $t_1 \geq T - \tau^{(k)}, t_2 - t_1 > \epsilon T$ is upper bounded by $O(N^{-1/2}(t_2 - t_1)^{-1/2}\sqrt{\log(NT)})$, with probability $1 - O(N^{-1}T^{-1})$. This together with the uniform upper bound for the first term yields the desired uniform rate of convergence. □

A.5 Proof of Theorem 1

Notice that Theorem 1 is automatically implied by the following three lemmas. We focus on proving these lemmas one by one in this section.

Lemma 2. *Suppose MA, LSE, LHE, Assumptions 1 – 3, 5 hold and* $s_{cp} \gg (NT)^{-1/2}\log^{3/2}(NT)$. *When using the oracle cluster memberships as input, the estimated change points computed by the proposed most recent change point detection subroutine satisfy*

$$\max_i \frac{|\widehat{\tau}_i^* - \tau_i^*|}{\tau_i^*} = O\left[\frac{\log^3(NT)}{NTs_{cp}^2}\right], \tag{??}$$

wpa1.

Lemma 3. *Suppose MA, LSE, LHE and Assumptions 1 – 3, 5 hold. Suppose the initial estimators satisfy* $\max_i[\tau_i^0 - \tau_i^*]_+/\tau_i^* \ll s_{cl}$, $s_{cl} \gg T^{-1/2}\sqrt{\log(NT)}$, $\min_i \tau_i^0 \geq \kappa T$ *for some constant* $\kappa > 0$, *and the number of cluster* K *is correctly specified. Then the estimated cluster memberships based on the proposed clustering subroutine achieves a zero clustering error, wpa1.*

Lemma 4. *Suppose MA, LSE, LHE and Assumptions 1 – 3, 5 hold. Suppose the initial estimators satisfy* $\max_i[\tau_i^0 - \tau_i^*]_+/\tau_i^* \ll T^{-1/2}\sqrt{\log(NT)}$, $s_{cl} \gg \max(T^{-1/4}, \log^{3/2}(NT)/(\sqrt{NT}s_{cp}))$, $\min_i \tau_i^0 \geq \kappa T$ *for some constant* $\kappa > 0$. *Then the proposed IC correctly identifies* K, *wpa1.*

A.5.1 Proof of Lemma 2

We first prove Lemma 2. Notice that the clustering error equals exactly zero wpa1. Lemma 2 thus implies that at each iteration, the estimated $\{\widehat{\tau}_i^*\}_i$ will converge at a rate of (**??**), wpa1. Additionally, the condition on s_{cp} is automatically implied by Assumption 4.

Proof. The proof is divided into two steps. In the first step, we aim to show that for all $\tau < \tau^{(k)}$ and k, the threshold used in the likelihood ratio test, as a function of the sample size $|\mathcal{C}_k|\tau$, is greater than the maximum log-likelihood ratio statistics with probability $1 - O(N^{-1}T^{-1})$. This implies that our method will not underestimate $\tau^{(k)}$.

Recall that for a given candidate change point location u, the loglikelihood is given by

$$\begin{aligned}\mathrm{LR}(\mathcal{C}_k, [T-\tau, T], u) = &-\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T} f(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{\mathcal{C}_k,[T-\tau,T]}) \\ &+\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{u} f(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{\mathcal{C}_k,[T-\tau,u-1]}) + \sum_{i\in\mathcal{C}_k}\sum_{t=u+1}^{T} f(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{\mathcal{C}_k,[u,T]}).\end{aligned}$$

To simplify the notation, let $\widehat{\theta}_k^{null} = \widehat{\theta}_{\mathcal{C}_k,[T-\tau,T]}$, $\widehat{\theta}_k^1 = \widehat{\theta}_{\mathcal{C}_k,[T-\tau,u-1]}$, and $\widehat{\theta}_k^2 = \widehat{\theta}_{\mathcal{C}_k,[u,T]}$. Using

Taylor expansion, we obtain that

$$\begin{aligned}
\mathrm{LR}(\mathcal{C}_k, [T-\tau, T], u) = & \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{u} f(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0) \\
& - (\widehat{\theta}_k^1 - \theta_k^0)^\top \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{u} f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_k^{1,*})(\widehat{\theta}_k^1 - \theta_k^0) \\
& + \sum_{i\in\mathcal{C}_k}\sum_{t=u+1}^{T} f(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0) \\
& - (\widehat{\theta}_k^2 - \theta_k^0)^\top \sum_{i\in\mathcal{C}_k}\sum_{t=u+1}^{T} f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_k^{2,*})(\widehat{\theta}_k^2 - \theta_k^0) \\
& - \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T} f(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0) \\
& + (\widehat{\theta}_k^{null} - \theta_k^0)^\top \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T} f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_k^{n,*})(\widehat{\theta}_k^{null} - \theta_k^0),
\end{aligned}$$

for some $\theta_k^{1,*}$, $\theta_k^{2,*}$, $\theta_k^{n,*}$ that lie on the line segments joining θ_k^0 and $\widehat{\theta}_k^1$, θ_k^0 and $\widehat{\theta}_k^2$, θ_k^0 and $\widehat{\theta}_k^{null}$, respectively. It suffices to analyse the second, fourth and last lines in the above expression. Below, we analyse them one by one:

- Under Assumption 1, it follows from Lemma 1 that wpa1, the difference between $\widehat{\theta}_k^{null}$ and θ_k^0 is upper bounded by $O(N^{-1/2}T^{-1/2}\sqrt{\log(NT)})$. Under the boundedness assumption on the second-order derivatives, the last line is upper bounded by $O(\log(NT))$, wpa1.
- Similarly, for any candidate change point location u such that $u+\tau-T \geq \epsilon T$, the difference between $\widehat{\theta}_k^{null}$ and θ_k^0 is upper bounded by $O(N^{-1/2}(u+\tau-T)^{-1/2}\sqrt{\log(NT)})$. Hence, the second term is upper bounded by $O(\log(NT))$. In cases where N is finite and $u+\tau-T < \epsilon T$, it follows from the boundedness of the parameter space in Assumption 2 and the definition of ϵ that the second line is upper bounded by $O(\epsilon^2 T^2) = O(\log^2(NT)\log(\log(NT)))$.
- Using similar arguments in the second bullet point, we can show that the fourth term is upper bounded by $O(\log^2(NT)\log(\log(NT)))$ as well.

To summarize, we have shown that the likelihood ratios are uniformly upper bounded by $O(\log^2(NT)\log(\log(NT)))$, wpa1. According to Assumption 1, $|\mathcal{C}_k|\tau$ approaches infinity as $NT \to \infty$. It follows from Section B.2 that the threshold is much larger than the maximum likelihood ratio. This completes the proof for the first step.

In the second step, we show that the test statistics would exceed the threshold if

$$\tau = \tau^{(k)} + N^{-1}\log^3(NT)s_{cp}^{-2}.$$

Consider the log-likelihood ratio $\mathrm{LR}(\mathcal{C}_k, [T-\tau, T], \tau^{(k)})$. Similarly, it follows from Taylor expansion

that $\mathrm{LR}(\mathcal{C}_k, [T-\tau, T], \tau^{(k)})$ equals

$$\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T-\tau^{(k)}} f(S_{it}|S_{it-1}, A_{it-1}, \theta_k^1) - \frac{1}{2}(\widehat{\theta}_k^1 - \theta_k^1)^\top \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T-\tau^{(k)}} f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_k^{1,*})(\widehat{\theta}_k^1 - \theta_k^1)$$
$$+ \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau^{(k)}+1}^{T} f(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0) - \frac{1}{2}(\widehat{\theta}_k^2 - \theta_k^0)^\top \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau^{(k)}+1}^{T} f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_k^{0,*})(\widehat{\theta}_k^2 - \theta_k^0)$$
$$- \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T} f(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0) + \frac{1}{2}(\widehat{\theta}_k^{null} - \theta_k^0)^\top \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T} f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_k^{n,*})(\widehat{\theta}_k^{null} - \theta_k^0)], \tag{A.7}$$

for some $\theta_k^{1,*}$, $\theta_k^{2,*}$, $\theta_k^{n,*}$ that lie on the line segments joining θ_k^1 and $\widehat{\theta}_k^1$, θ_k^0 and $\widehat{\theta}_k^0$, θ_k^0 and $\widehat{\theta}_k^{null}$, respectively, all converging to θ_k^0 as θ_k^1 is asymptotically equivalent to θ_k^0.

Meanwhile, using similar arguments to the proof of Lemma 1, we obtain that the minimum eigenvalues of $\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T-\tau^{(k)}} f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_k^{1,*})/(|\mathcal{C}_k|(\tau-\tau^{(k)}))$ and $\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau^{(k)}+1}^{T} f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_k^{0,*})/(|\mathcal{C}_k|\tau^{(k)})$ are bounded away from zero, wpa1, uniform in k. This together with the boundedness of f'' yields that

$$(\text{A.7}) \geq \sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T-\tau^{(k)}} \log\frac{p(S_{it}|S_{it-1}, A_{it-1}, \theta_k^1)}{p(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0)} - C|\mathcal{C}_k|\tau\|\widehat{\theta}_k^{null} - \theta_k^0\|^2, \tag{A.8}$$

for some constant $C > 0$.

Using similar arguments in the proof of Lemma 1, we can show that the the convergence rate $\|\widehat{\theta}_k^{null} - \theta_k^0\|_2^2$ is proportional to the order of magnitude of

$$\left\|\frac{1}{|\mathcal{C}_k|\tau}\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T-\tau^{(k)}} f'(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0)\right\|_2^2 + \left\|\frac{1}{|\mathcal{C}_k|\tau}\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau^{(k)}+1}^{T} f'(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0)\right\|_2^2, \tag{A.9}$$

by Cauchy-Schwarz inequality. Notice that according to the Azuma Hoeffding's inequality, the second term in (A.9) is $O(N^{-1}T^{-1}\log(NT))$, with probability $1 - O(N^{-1}T^{-1})$. As for the first term, using similar arguments in the proof for Step 1 of Lemma 1, we can show that it is of the order of magnitude of

$$\frac{1}{N^2T^2}\left[\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T-\tau^{(k)}} \mathbb{E}f'(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0)\right]^2 + \frac{(\tau-\tau^{(k)})\log^2(NT)}{N^2T^2}$$
$$= O\left(\frac{(\tau-\tau^{(k)})^2}{T^2}\|\theta_k^1 - \theta_k^0\|^2\right) + \frac{(\tau-\tau^{(k)})\log^2(NT)}{N^2T^2},$$

with probability $1 - O(N^{-1}T^{-1})$. It follows from (A.8) that the log-likelihood ratio is larger than or equal to

$$\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T-\tau^{(k)}} \log\frac{p(S_{it}|S_{it-1}, A_{it-1}, \theta_k^1)}{p(S_{it}|S_{it-1}, A_{it-1}, \theta_k^0)} - \frac{c(\tau-\tau^{(k)})\log^2(NT)}{NT} - \frac{cN(\tau-\tau^{(k)})^2}{T}\|\theta_k^1 - \theta_k^0\|^2,$$

for some constant $c > 0$.

Next, using similar arguments to Step 1 of the proof of Lemma 1, we can show that the first term in the above expression is larger than or equal to

$$\left[\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T-\tau^{(k)}} \mathbb{E}\log\frac{p(S_{it}|S_{it-1},A_{it-1},\theta_k^1)}{p(S_{it}|S_{it-1},A_{it-1},\theta_k^0)}\right] - O\left(\|\theta_k^1-\theta_k^0\|\sqrt{N(\tau-\tau^{(k)})}\log(NT)\right),$$

with probability $1 - O(N^{-1}T^{-1})$.

Moreover, under Assumption 3, using a second order Taylor expansion and similar arguments in bounding the first term on the RHS of (A.6), we can show that,

$$\sum_{i\in\mathcal{C}_k}\sum_{t=T-\tau+1}^{T-\tau^{(k)}} \mathbb{E}\log\frac{p(S_{it}|S_{it-1},A_{it-1},\theta_k^1)}{p(S_{it}|S_{it-1},A_{it-1},\theta_k^0)} \geq c\lambda N(\tau-\tau^{(k)})\|\theta_k^1-\theta_k^0\|^2,$$

for some constant $c > 0$. Combining these results together, it is immediate to see that under the given specification on τ and the signal strength condition that $s_{cp} \gg (NT)^{-1/2}\log^{3/2}(NT)$, the likelihood ratio is strictly larger than 0 and is strictly larger than the threshold with probability approaching 1. This completes the proof for the second step. Therefore, the change point detection procedure will stop as long as $\tau \geq \tau^{(k)} + N^{-1}s_{cp}^{-2}\log^3(NT)$. This yields the desired rate of convergence. □

A.5.2 Proof of Lemma 3

We next prove Lemma 3. Suppose Lemma 3 is proven. Under the conditions that $s_{cl} \gg (NT)^{-1/2}s_{cp}^{-1}\log^{3/2}(NT)$ and $s_{cp} \gg N^{-1/2}T^{-1/4}\log^{3/2}(NT)$ in Assumption 4, it follows that s_{cl} is much larger than the change point detection error in (**??**). Consequently, when K is correctly specified, it follows that the clustering error will be zero wpa1 during each iteration — not just at the initial iteration.

Proof. We use $k(\bullet)$ to denote a given mapping from the indices of subjects $\{1,\cdots,N\}$ to the indices of clusters $\{1,\cdots,K\}$. Let $\mathcal{K}$ denote the set $\{k(i)\}_i$. For a given set of parameters $\theta = \{\theta_k\}_k$, define

$$Q(\theta,\mathcal{K}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\widehat{\tau}_i}\sum_{t=T-\widehat{\tau}_i+1}^{T}\log p(S_{i,t}|A_{i,t-1},S_{i,t-1};\theta_{k(i)}).$$

Let $(\widehat{\theta},\widehat{\mathcal{K}}) = \arg\max Q(\theta,\mathcal{K})$ where $\widehat{\mathcal{K}} = \{\widehat{k}_i\}_i$. For a given value of θ, define the optimal group assignment for each unit as

$$\widehat{k}_i(\theta) = \underset{k\in\{1,\ldots,K\}}{\arg\max}\sum_{t=T-\widehat{\tau}_i^*+1}^{T}\log p(S_{it}|S_{it-1},A_{it-1},\theta_k).$$

For conciseness, we write $\widehat{k}_i(\widehat{\theta})$ as $\widehat{k}_i$, let k_i^0 denote the oracle group assignment for the ith unit and $\mathcal{K}^0 = \{k_i^0\}_k$. Let $\theta^0 = \{\theta_k^0\}_k$ denote the set of oracle parameters.

In the first step, we establish the rate of convergence of the estimated parameters $N^{-1}\sum\|\widehat{\theta}_{\widehat{k}_i}-\theta^0_{k_i^0}\|^2$. It follows that

$$\begin{aligned}
0 &\geq Q(\theta^0,\mathcal{K}^0)-Q(\widehat{\theta},\widehat{\mathcal{K}}) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\tau_i^0}\sum_{t=T-\min(\tau_i^*,\tau_i^0)+1}^{T}[f(S_{it}|S_{it-1},A_{it-1},\theta^0_{k_i^0})-f(S_{it}|S_{it-1},A_{it-1},\widehat{\theta}_{\widehat{k}_i})] \\
&+ \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\tau_i^0}\sum_{t=T-\tau_i^0+1}^{T-\min(\tau_i^*,\tau_i^0)}[f(S_{it}|S_{it-1},A_{it-1},\theta^0_{k_i^0})-f(S_{it}|S_{it-1},A_{it-1},\widehat{\theta}_{\widehat{k}_i})]\mathbb{I}(\tau_i^0>\tau_i^*).
\end{aligned} \tag{A.10}$$

By Taylor expansion, the second line equals

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\tau_i^0}\sum_{t=T-\min(\tau_i^*,\tau_i^0)+1}^{T}\Big[&-f^{'}(S_{it}|S_{it-1},A_{it-1},\theta^0_{k_i^0})^\top(\widehat{\theta}_{\widehat{k}_i}-\theta^0_{k_i^0}) \\
&-\frac{1}{2}(\theta^0_{k_i^0}-\widehat{\theta}_{\widehat{k}_i})^\top f^{''}(S_{it}|S_{it-1},A_{it-1},\theta_i^*)(\theta^0_{k_i^0}-\widehat{\theta}_{\widehat{k}_i})\Big],
\end{aligned}$$

for some θ_i^* that lies on the line segment joining $\theta^0_{k_i^0}$ and $\widehat{\theta}_{\widehat{k}_i}$.

Under the given conditions, both τ_i^* and τ_i^0 are proportional to T. As $T\to\infty$, under the minimum eigenvalue condition in Assumption 3, using similar arguments to Step 1 of the proof of Lemma 1, we can show that the minimum eigenvalues of the matrices $\{-(\tau_i^0)^{-1}\sum_{t=T-\min(\tau_i^*,\tau_i^0)+1}^{T}f^{''}(S_{it}|S_{it-1},A_{it-1},\theta_i^*)\}_i$ are uniformly bounded away from zero with probability $1-O(N^{-1}T^{-1})$. In addition, using Azuma Hoeffding's inequality, sums of the scores $\{\|\sum_{t=T-\min(\tau_i^*,\tau_i^0)+1}^{T}f'(S_{it}|S_{it-1},A_{it-1},\theta^0_{k_i^0})\|\}_i$ can be uniformly upper bounded by $\sqrt{T\log(NT)}$, with probability $1-O(N^{-1}T^{-1})$. To summarise, we have shown that the second line of (A.10) is lower bounded by

$$\frac{c}{N}\sum_{i=1}^{N}\|\widehat{\theta}_{\widehat{k}_i}-\theta^0_{k_i^0}\|^2-\frac{C\sqrt{\log(NT)}}{N\sqrt{T}}\sum_{i=1}^{N}\|\widehat{\theta}_{\widehat{k}_i}-\theta^0_{k_i^0}\|.$$

with probability approaching 1, for some constants $c,C>0$.

Next, consider the third line of (A.10). Under Assumption 3, the derivative f' is uniformly bounded. Under (A1), τ_i^0 is proportional to τ_i^* for any i. As such, the third line can be lower bounded by

$$-\frac{O(1)}{N}\sum_{i=1}^{N}\frac{[\tau_i^0-\tau_i^*]_+}{\tau_i^*}\|\widehat{\theta}_{\widehat{k}_i}-\theta^0_{k_i^0}\|,$$

where $O(1)$ denotes some positive constant whose value is allowed to vary from place to place.

It follows from (A.10) that with probability $1-O(N^{-1}T^{-1})$,

$$\frac{c}{N}\sum_{i=1}^{N}\|\widehat{\theta}_{\widehat{k}_i}-\theta^0_{k_i^0}\|^2\leq\frac{O(1)}{N}\sum_{i=1}^{N}\left(\frac{[\tau_i^0-\tau_i^*]_+}{\tau_i^*}+\frac{\sqrt{\log(NT)}}{\sqrt{T}}\right)\|\widehat{\theta}_{\widehat{k}_i}-\theta^0_{k_i^0}\|,$$

for some positive constant denoted by $O(1)$. Using Cauchy-Schwarz inequality, it is immediate to see that with probability $1 - O(N^{-1}T^{-1})$, we have

$$\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\theta}_{\widehat{k}_i} - \theta^0_{k^0_i}\|^2 \leq O(1)\frac{\log(NT)}{T} + O(1)\frac{1}{N}\sum_{i=1}^{N}\frac{[\tau^0_i - \tau^*_i]^2_+}{(\tau^*_i)^2}. \tag{A.11}$$

This completes the proof of the first step.

In the second step, we aim to show that the clustering algorithm achieves a zero clustering error, with probability $1-O(N^{-1}T^{-1})$. Toward that end, we notice that under the current conditions, the signal strength s_{cl} is much larger than the square root of the RHS of (A.11). Since K is correctly specified, using similar arguments in the proof of Lemma B.3 in Bonhomme and Manresa (2015), we can show the existence of a permutation $\sigma(\bullet) : \{1, \cdots, K\} \to \{1, \cdots, K\}$ such that for each k,

$$\|\widehat{\theta}_{\sigma(k)} - \theta^0_k\|^2 \leq O(1)\frac{\log(NT)}{T} + O(1)\frac{1}{N}\sum_{i=1}^{N}\frac{[\tau^0_i - \tau^*_i]^2_+}{(\tau^*_i)^2}, \tag{A.12}$$

and that $\sum_{i\in\mathcal{C}_k}\mathbb{I}(i = \sigma(k)) \xrightarrow{P} 1$. Without loss of generality, assume σ is an identity function such that $\sigma(k) = k$ for any k. Notice that at this point, we have shown that the clustering error decays to zero. Below, we show that it is exactly zero with probability $1 - O(N^{-1}T^{-1})$. A key observation is that, since the estimated cluster membership maximise the log-likelihood function, we have for each i that

$$\sum_{t=T-\tau^0_i+1}^{T} f(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{\widehat{k}_i}) \geq \sum_{t=T-\tau^0_i+1}^{T} f(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{k^0_i}).$$

Similar to Step 1 of the proof, this implies that with probability approaching 1, we have for any i that

$$\|\widehat{\theta}_{\widehat{k}_i} - \widehat{\theta}_{k^0_i}\|^2 \leq C\frac{(\tau^0_i - \tau^*_i)^2_+}{(\tau^*_i)^2} + C\left\|\frac{1}{\tau^0_i}\sum_{t=T-\min(\tau^0_i,\tau^*_i)+1}^{T} f^{'}(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{k^0_i})\right\|^2, \tag{A.13}$$

for some positive constant $C > 0$. Using Taylor expansion and Cauchy-Schwarz inequality, the second term on the RHS can be upper bounded by

$$2C\left\|\frac{1}{\tau^0_i}\sum_{t=T-\min(\tau^0_i,\tau^*_i)+1}^{T} f^{'}(S_{it}|S_{it-1}, A_{it-1}, \theta^0_{k^0_i})\right\|^2 + O(1)\|\widehat{\theta}_{k^0_i} - \theta^0_{k^0_i}\|^2,$$

where $O(1)$ denotes some positive constant. Similar to Step 1 of the proof, with probability approaching 1, the first term of the above expression can be upper bounded by $T^{-1}\log(NT)$ and the bound is uniform in i. Meanwhile, the second term can be upper bounded based on (A.12). As such, it follows from (A.13) that

$$\|\widehat{\theta}_{\widehat{k}_i} - \widehat{\theta}_{k^0_i}\|^2 \leq O(1)\max_i\frac{[\tau^0_i - \tau^*_i]^2_+}{(\tau^*_i)^2} + O(1)\frac{\log(NT)}{T}, \tag{A.14}$$

for some positive constant $O(1)$.

Given that $s_{cl} \gg \max_i(\tau^*_i)^{-1}(\tau^0_i - \tau^*_i)_+ + T^{-1/2}\sqrt{\log(NT)}$, according to (A.12), the difference $\|\widehat{\theta}_{k_1} - \widehat{\theta}_{k_2}\|$ is at least $s_{cl}/2$ whenever $k_1 \neq k_2$. As such, (A.14) holds only when $\widehat{k}_i = k^0_i$. This completes the proof for the second step. □

A.5.3 Proof of Lemma 4

Before proving Lemma 4, we remark that to guarantee the results in Lemma 4 hold for later iterations in addition to the first iteration, we require the estimated change point to converge at a rate of $T^{-1/2}\log(NT)$. However, this is achieved by the condition $s_{cp} \gg N^{-1/2}T^{-1/4}\log^{3/2}(NT)$ in Assumption 4. Thus, under Assumption 4, Lemma 4 implies the consistency of the proposed IC at every iteration, not just the first iteration.

Proof. Firstly, consider the case where $K > K^0$ where K^0 is the true number of clusters. For a given K, we rewrite the clustering objective function as $Q(\theta, \mathcal{K}|K)$ and denote $(\widehat{\theta}(K), \widehat{\mathcal{K}}(K)) = \arg\max Q(\theta, \mathcal{K}|K)$ to highlight their dependencies upon K.

Using similar arguments in the proof of Lemma 3, we can show that $\widehat{\theta}_K$ also satisfies the rate of convergence in (A.11) with probability $1 - O(N^{-1}T^{-1})$. Meanwhile, it is easy to show that the upper error bound therein can be refined by replacing the $\log(NT)$ term with on the RHS with $\log(1/\alpha)$ for any fixed $\alpha \in (0, 1]$. However, this refinement comes at the cost of changing the high probability bound from the previous $1 - O(N^{-1}T^{-1})$ to $1 - \alpha - o(1)$. Therefore, the difference between the proposed IC with K^0 clusters and that with K many clusters equals

$$
\begin{aligned}
&IC(\widehat{\theta}_{K^0}, \widehat{\mathcal{K}}_{K^0}) - IC(\widehat{\theta}_K, \widehat{\mathcal{K}}_K) = N[Q(\widehat{\theta}_{K^0}, \widehat{\mathcal{K}}_{K^0}|K^0) - Q(\widehat{\theta}_K, \widehat{\mathcal{K}}_K|K)] + (K - K^0)\frac{N\log(NT)}{T} \\
=&\sum_{i=1}^{N}\frac{1}{\tau_i^0}\sum_{t=T-\min(\tau_i^*,\tau_i^0)}^{T} f^{'}(S_{it}|S_{it-1}, A_{it-1}, \theta_{k_i^0}^0)^\top(\widehat{\theta}_{\widehat{k}_i(K^0)}(K^0) - \widehat{\theta}_{\widehat{k}_i(K)}(K)) \\
+&\frac{1}{2}\sum_{i=1}^{N}\frac{1}{\tau_i^0}\sum_{t=T-\min(\tau_i^*,\tau_i^0)}^{T}(\theta_{k_i^0}^0 - \widehat{\theta}_{\widehat{k}_i(K^0)}(K^0))^\top f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_i^*(K^0))(\theta_{k_i^0}^0 - \widehat{\theta}_{\widehat{k}_i(K^0)}(K^0)) \\
-&\frac{1}{2}\sum_{i=1}^{N}\frac{1}{\tau_i^0}\sum_{t=T-\min(\tau_i^*,\tau_i^0)}^{T}(\theta_{k_i^0}^0 - \widehat{\theta}_{\widehat{k}_i(K)}(K))^\top f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_i^*(K))(\theta_{k_i^0}^0 - \widehat{\theta}_{\widehat{k}_i(K)}(K)) \\
-&\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\tau_i^0}\sum_{t=T-\tau_i^0+1}^{T-\min(\tau_i^*,\tau_i^0)}[f(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{\widehat{k}_i(K^0)}(K^0)) - f(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{\widehat{k}_i(K)}(K))]\mathbb{I}(\tau_i^0 > \tau_i^*) \\
+&(K - K^0)\frac{N\log(NT)}{T},
\end{aligned}
$$

for some $\theta_i^*(K^0)$ and $\theta_i^*(K)$ that lie between the oracle parameter and the estimator. Below, we analyse the terms on the RHS one by one:

- Using similar arguments to the proof of Lemma 3, the second line can be lower bounded by $-O(T^{-1/2}\sqrt{\log(1/\alpha)})\sum_{i=1}^{N}\|\widehat{\theta}_{\widehat{k}_i(K^0)}(K^0) - \widehat{\theta}_{\widehat{k}_i(K)}(K)\|$, with probability $1 - \alpha$. Based on the established convergence rates for $\sum_{i=1}^{N}\|\widehat{\theta}_{\widehat{k}_i(K^0)}(K^0) - \theta_{k_i^0}^0\|^2$ and $\sum_{i=1}^{N}\|\widehat{\theta}_{\widehat{k}_i(K)}(K) - \theta_{k_i^0}^0\|^2$, it follows from Cauchy-Schwarz inequality that the second line can be further lower bounded by

$$
-O(1)\frac{N\log(1/\alpha)}{T} - O(1)\sum_{i=1}^{N}\frac{(\tau_i^0 - \tau_i^*)_+^2}{(\tau_i^0)^2},
$$

  with probability $1 - \alpha$.

- Similarly, based on the established convergence rates for $\sum_{i=1}^{N} \|\widehat{\theta}_{\widehat{k}_i(K^0)}(K^0) - \theta^0_{k_i^0}\|^2$, the third line can be lower bounded by

$$-O(1)\frac{N\log(1/\alpha)}{T} - O(1)\sum_{i=1}^{N}\frac{(\tau_i^0 - \tau_i^*)_+^2}{(\tau_i^0)^2},$$

  with probability $1-\alpha$.

- Using similar arguments to Step 1 of the proof of Lemma 1, we can show that the minimum eigenvalues of the matrices $\{-(\tau_i^0)^{-1}\sum_{t=T-\min(\tau_i^*,\tau_i^0)+1}^{T} f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_i^*(K))\}_i$ are positive semi-definite wpa1. Consequently, the fourth line is non-negative wpa1.

- Similar to the proof of Lemma 3, the fifth line can be lower bounded by $-[\max_i(\tau_i^0 - \tau_i^*)_+/\tau_i^0]\sum_{i=1}^{N}\|\widehat{\theta}_{\widehat{k}_i(K^0)}(K^0) - \widehat{\theta}_{\widehat{k}_i(K)}(K)\|$, which, according to the convergence rates of $\widehat{\theta}_{\widehat{k}_i(K^0)}(K^0)$ and $\widehat{\theta}_{\widehat{k}_i(K)}(K)$, can be further lower bounded by

$$-O(1)\frac{N\log(1/\alpha)}{T} - O(1)\sum_{i=1}^{N}\frac{(\tau_i^0 - \tau_i^*)_+^2}{(\tau_i^0)^2},$$

  with probability $1-\alpha$.

- Finally, the last line is lower bounded by $N\log(NT)/T$, as $K > K^0$.

To summarize, we have shown that $IC(\widehat{\theta}_{K^0}, \widehat{\mathcal{K}}_{K^0}) - IC(\widehat{\theta}_K, \widehat{\mathcal{K}}_K)$ is lower bounded by

$$\frac{N\log(NT)}{T} - O(1)\frac{N\log(1/\alpha)}{T} - O(1)\sum_{i=1}^{N}\frac{(\tau_i^0 - \tau_i^*)_+^2}{(\tau_i^0)^2},$$

with probability $1-\alpha-o(1)$. The above expression is strictly positive, under the given conditions on the initial change point estimator. This suggests that the proposed IC will not over-select the number of clusters with probability $1-\alpha-o(1)$.

Next, consider the case where $K < K^0$. We claim that there are $\Omega(N)$ many subjects being wrongly clustered into the same group. This is because as K is smaller than K^0, there will be at least two true clusters, say the k_1-th and k_2-th clusters, with over $|\mathcal{C}_{k_1}|/K^0$ and $|\mathcal{C}_{k_2}|/K^0$ many subjects, respectively, being assigned to the same cluster. Under Assumption 1, then number of subjects in this wrongly formed cluster is proportional to N. Let $\widetilde{\theta}$ denote the estimated parameter using data from this cluster. The difference between the proposed IC with K^0 clusters and that

with K clusters is given by

$$IC(\widehat{\theta}_{K^0}, \widehat{\mathcal{K}}_{K^0}) - IC(\widehat{\theta}_K, \widehat{\mathcal{K}}_K) = N[Q(\widehat{\theta}_{K^0}, \widehat{\mathcal{K}}_{K^0}|K^0) - Q(\widehat{\theta}_K, \widehat{\mathcal{K}}_K|K)] + (K - K^0)\frac{N\log(NT)}{T}$$

$$= \sum_{i=1}^{N} \frac{1}{\tau_i^0} \sum_{t=T-\min(\tau_i^*, \tau_i^0)}^{T} f^{'}(S_{it}|S_{it-1}, A_{it-1}, \theta_{k_i^0}^0)^\top (\widehat{\theta}_{\widehat{k}_i(K^0)}(K^0) - \widehat{\theta}_{\widehat{k}_i(K)}(K))$$

$$+ \frac{1}{2}\sum_{i=1}^{N} \frac{1}{\tau_i^0} \sum_{t=T-\min(\tau_i^*, \tau_i^0)}^{T} (\theta_{k_i^0}^0 - \widehat{\theta}_{\widehat{k}_i(K^0)}(K^0))^\top f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_i^*(K^0))(\theta_{k_i^0}^0 - \widehat{\theta}_{\widehat{k}_i(K^0)}(K^0))$$

$$- \frac{1}{2}\sum_{i=1}^{N} \frac{1}{\tau_i^0} \sum_{t=T-\min(\tau_i^*, \tau_i^0)}^{T} (\theta_{k_i^0}^0 - \widehat{\theta}_{\widehat{k}_i(K)}(K))^\top f^{''}(S_{it}|S_{it-1}, A_{it-1}, \theta_i^*(K))(\theta_{k_i^0}^0 - \widehat{\theta}_{\widehat{k}_i(K)}(K))$$

$$- \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\tau_i^0} \sum_{t=T-\tau_i^0+1}^{T-\min(\tau_i^*, \tau_i^0)} [f(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{\widehat{k}_i(K^0)}(K^0)) - f(S_{it}|S_{it-1}, A_{it-1}, \widehat{\theta}_{\widehat{k}_i(K)}(K))]\mathbb{I}(\tau_i^0 > \tau_i^*)$$

$$+ (K - K^0)\frac{N\log(NT)}{T}.$$

Again, we analyse the above expression line by line:

- Similarly, the second line can be lower bounded by $-O(NT^{-1/2}\sqrt{\log(1/\alpha)})$, with probability $1-\alpha$;
- with probability $1-\alpha$, the third line is again lower bounded by

$$-O(1)\frac{N\log(1/\alpha)}{T} - O(1)\sum_{i=1}^{N}\frac{(\tau_i^0 - \tau_i^*)_+^2}{(\tau_i^0)^2}.$$

- The fourth line is lower bounded by $cT^{-1}\sum_{i=1}^{N}\|\theta_{k_i^0}^0 - \widehat{\theta}_{\widehat{k}_i(K)}(K)\|^2$ for some constant $c > 0$ wpa1. Considering the $\Omega(N)$ many subjects who originally belong to the k_1-th and k_2-th clusters but are wrongly clustered together, this term is at least $CN\|\theta_{k_1}^0 - \theta_{k_2}^0\|^2 \geq CNs_{cl}^2$.
- The fifth line can be similarly lower bounded by $-O(1)N[\max_i(\tau_i^0 - \tau_i^*)_+/\tau_i^0]$.
- The last line is $-O(T^{-1}N\log(NT))$.

To summarize, we have shown that $IC(\widehat{\theta}_{K^0}, \widehat{\mathcal{K}}_{K^0}) - IC(\widehat{\theta}_K, \widehat{\mathcal{K}}_K)$ is lower bounded by

$$CNs_{cl}^2 - O(1)\frac{\sqrt{\log(1/\alpha)}}{\sqrt{T}} - -O(1)N\max_i\frac{(\tau_i^0 - \tau_i^*)_+}{\tau_i^0}.$$

Under the given signal strength conditions in Assumption 4, it is strictly positive, with probability at least $1 - \alpha - o(1)$.

Consequently, we have shown that the proposed IC is maximized at the true number of clusters K^0, with probability $1-\alpha-o(1)$. Since α can be made arbitrarily small, it follows that the proposed IC is consistent. This completes the proof.

□

A.6 Proof of Corollary 1

Proof. The proof of Corollary 1 is straightforward. Based on the results in Theorem 4.1 and the condition on N, at each iteration, we can show that the change point detection error is smaller than $1/T$ wpa1. Since the change point detection error can only take values 0, $1/T$, $2/T$, etc., it implies that the change point detection error equals exactly 0 wpa1. The proof is hence completed. □

A.7 Proof of Theorem 2

Theorem 2 is concerned with the regrets of various estimated optimal policies. Below, we first derive the regret bound for the proposed algorithm. We next prove the inconsistencies of the Homogeneous, Stationary and Doubly Homogeneous algorithms.

A.7.1 The proposed algorithm

Proof. The proof is very similar to that of Chen and Jiang (2019), who established the regret bound of the FQI algorithm in standard doubly homogeneous environments. Consequently, we provide a sketch of the proof only, focusing on highlighting the difference from that of Chen and Jiang (2019).

Our proof is divided into three parts. In Part 1, we define another regret by assuming the transition never change after time T, and bound the difference between this regret and the original definition. In Part 2, we apply the performance difference lemma to upper bound our newly defined regret using the estimation error of the Q-function. Finally, in Part 3, we derive the Q-function's estimation error.

Part 1. We begin with a new definition of the cumulative reward. Specifically, for a given policy π, let $\mathbb{E}_s^\pi$ denote the expectation by assuming the transition function $p_{i,t}$ remains stationary after time T. This allows us to define the following expected cumulative reward

$$J_s(\pi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=T+1}^{\infty}\gamma^{t-T-1}\mathbb{E}_s^\pi(R_{i,t}),$$

and its associated regret $\sup_{\pi^*} J_s(\pi^*) - J_s(\pi)$. In this part, we focus us providing an upper bound for $\sup_{\pi^*} J(\pi^*) - J(\pi) - [\sup_{\pi^*} J_s(\pi^*) - J_s(\pi)]$.

Recall that T^* corresponds to the most recent change point after T. By definition, $\mathbb{E}_s^\pi(R_{i,t})$ is equal to $\mathbb{E}^\pi(R_{i,t})$ for any $T < t < T + T^*$. Let π^{**} denote the argmax of $J(\pi^*)$, we have

$$\begin{aligned}\sup_{\pi^*} J(\pi^*) - J(\pi) - [\sup_{\pi^*} J_s(\pi^*) - J_s(\pi)] &\le J(\pi^{**}) - J(\pi) - J_s(\pi^{**}) + J_s(\pi)\\ &\le \sup_{\pi}\frac{2}{N}\sum_{i=1}^{N}\sum_{t=T^*}^{\infty}\gamma^{t-T-1}|\mathbb{E}_s^\pi(R_{i,t}) - \mathbb{E}^\pi(R_{i,t})| \le 2R_{\max}\frac{\gamma^{T^*-T-1}}{1-\gamma},\end{aligned}$$

where the last inequality follows from the bounded reward assumption in Assumption 7. Under the assumption that $T^* - T \gg \log(T)/\log(\gamma^{-1})$ and N is at most proportional to TR, this term is of the order $T^{-C}R_{\max}/(1-\gamma)$, or equivalently $O(N^{-c}T^{-c}R_{\max}/(1-\gamma))$ for any sufficiently large constant $c > 0$. Notice that this term is negligible as the constant c can be made arbitrarily large. Without this condition, it will incur an additional term in the regret bound, given by

$$\frac{2R_{\max}\mathbb{E}(\gamma^{T^*-T})}{\gamma(1-\gamma)}.$$

Part 1 of the proof is thus completed.

Part 2. Based on the results in Part 1, it suffices to upper bound the newly defined regret $\sup_{\pi^*} J_s(\pi^*) - J_s(\pi)$. Under LHE, $J_s(\pi)$ can be represented by $N^{-1}\sum_{k=1}^{K}|\mathcal{C}_k|J_{k,s}(\pi)$ where

$$J_{k,s}(\pi) = \sum_{t=T+1}^{\infty} \gamma^{t-T-1}\mathbb{E}_s^{\pi}(R_{i,t}),$$

for any $i \in \mathcal{C}_k$. Since K is finite, it suffices to show that for each k, the regret $\sup_{\pi^*} J_{s,k}(\pi^*) - J_{s,k}(\widehat{\pi})$ is of the order of magnitude specified in Theorem 2.

By the definition of $J_{s,k}$, this result can be established using the arguments from proofs in doubly homogeneous environments; see e.g., the proof of Theorem 11 of Chen and Jiang (2019). We summarise the main steps below.

1. First, using the performance difference lemma (see e.g., Lemma 13 of Chen and Jiang, 2019), we can upper bound $\sup_{\pi^*} J_{s,k}(\pi^*) - J_{s,k}(\widehat{\pi})$ by $(1-\gamma)^{-1}[\|\widehat{Q}_{s,k} - Q^*_{s,k}\|_{\eta^{\widehat{\pi}}\times\widehat{\pi}} + \|\widehat{Q}_{s,k} - Q^*_{s,k}\|_{\eta^{\widehat{\pi}}\times\pi^*_s}]$ where $Q^*_{s,k}$ denotes the optimal Q-function for the k-th cluster assuming their transition function remains stationary after time point T, $\widehat{Q}_{s,k}$ denotes its estimator and for any policies π_1, π_2 and function $f(S, A)$,

$$\|f\|_{\eta^{\pi_1}\times\pi_2} := \sqrt{\mathbb{E}_{S\sim\eta^{\pi_1}, A\sim\pi_2} f^2(S, A)},$$

   where the expectation $\mathbb{E}_{S\sim\eta^{\pi_1}, A\sim\pi_2}$ is defined by assuming the state follows the discounted visitation distribution under π_1 and the action follows π_2.

2. Next, notice that under Assumptions 3 and 5, both the transition function and the behavior policy are bounded away from zero. This implies that Assumption 1 of Chen and Jiang (2019) is automatically satisfied with a finite concentratability coefficient. Now, using Lemmas 14 & 15 and the proof of Theorem 11 of Chen and Jiang (2019), we can further upper bound $(1-\gamma)^{-1}[\|\widehat{Q}_{s,k} - Q^*_{s,k}\|_{\eta^{\widehat{\pi}}\times\widehat{\pi}} + \|\widehat{Q}_{s,k} - Q^*_{s,k}\|_{\eta^{\widehat{\pi}}\times\pi^*_s}]$ by

$$O\Big(\frac{\gamma^J R_{\max}}{(1-\gamma)^2}\Big) + O\Big(\frac{\max_{1\le j\le J}\|\widehat{Q}^{(j)}_{s,k} - \mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}\|_{\mu_k}}{(1-\gamma)^2}\Big), \tag{A.15}$$

   where recall that J denotes the number of iterations in FQI, and $\widehat{Q}^{(j)}_{s,k}$ denotes the estimated Q-function computed at the jth iteration. Under the condition $J \gg \log(NT)/\log(\gamma^{-1})$ in Assumption 6, the first term in (A.15) becomes negligible. It remains to upper bound the second term in (A.15). We derive this upper bound in Part 3.

Part 3. As commented in Step 2 of the proof, we aim to upper bound $\max_{1\le j\le J}\|\widehat{Q}^{(j)}_{s,k} - \mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}\|_{\mu_k}$ in this step. The proof is again, similar to that of Lemma 16 of Chen and Jiang (2019). Below, we focus on highlighting their differences.

For any functions $Q_1, Q_2, Q_3 \in \mathcal{Q}$, define a function $g(s, a, r, s'; Q_1, Q_2, Q_3) = [r + \gamma\max_a \gamma Q_1(s', a) - Q_2(s, a)]^2 - [r + \gamma\max_a \gamma Q_1(s', a) - Q_3(s, a)]^2$. A key step in the proof is to establish a uniform

upper bound the following empirical process

$$\frac{1}{|\widehat{\mathcal{C}}_k|\widehat{\tau}^{(k)}}\sum_{i\in\widehat{\mathcal{C}}_k}\sum_{t=T-\widehat{\tau}^{(k)}}^{T-1}\Big[g(S_{i,t},A_{i,t},R_{i,t},S_{i,t+1};Q_1,Q_2,Q_3) \\ -\mathbb{E}[g(S_{i,t},A_{i,t},R_{i,t},S_{i,t+1};Q_1,Q_2,Q_3)]\Big], \tag{A.16}$$

indexed by $g\in\mathcal{G}=\{g:Q_1,Q_2,Q_3\in\mathcal{Q}\}$.

We next summarise the differences between our setting and the setting considered in Lemma 16 of Chen and Jiang (2019):

1. The error bound in Chen and Jiang (2019) is derived in stationary environments. To the contrary, we consider potentially non-stationary environments where the transition function can be non-stationary when $\widehat{\tau}^{(k)}$ over-estimates its oracle value $\tau^{(k)}$.

2. Chen and Jiang (2019) imposed a finite hypothesis class assumption on $\mathcal{Q}$ whereas our VC-class condition in Assumption 9 is more reasonable as it allows Q to be an infinite hypothesis class.

3. Chen and Jiang (2019) required the state-action-reward-next-state tuples to i.i.d. whereas we consider the more realistic setting by taking their temporal dependence into account.

To handle non-stationary environments, we decompose (A.16) into two terms, given by

$$\frac{1}{|\widehat{\mathcal{C}}_k|\widehat{\tau}^{(k)}}\sum_{i\in\widehat{\mathcal{C}}_k}\sum_{t=T-\widehat{\tau}^{(k)}}^{T-\min(\widehat{\tau}^{(k)},\tau^{(k)})}\Big[g(S_{i,t},A_{i,t},R_{i,t},S_{i,t+1};Q_1,Q_2,Q_3) \\ -\mathbb{E}[g(S_{i,t},A_{i,t},R_{i,t},S_{i,t+1};Q_1,Q_2,Q_3)|S_{i,t},A_{i,t}]\Big]$$

and

$$\frac{1}{|\widehat{\mathcal{C}}_k|\widehat{\tau}^{(k)}}\sum_{i\in\widehat{\mathcal{C}}_k}\sum_{t=T-\min(\widehat{\tau}^{(k)},\tau^{(k)})}^{T-1}\Big[g(S_{i,t},A_{i,t},R_{i,t},S_{i,t+1};Q_1,Q_2,Q_3) \\ -\mathbb{E}[g(S_{i,t},A_{i,t},R_{i,t},S_{i,t+1};Q_1,Q_2,Q_3)|S_{i,t},A_{i,t}]\Big]. \tag{A.17}$$

Under Assumptions 1 and 9, it follows from the conclusions in Theorem 1 that the first term can be upper bounded by

$$O\Big(\frac{\log^3(NT)R_{\max}^2}{NTs_{cp}^2(1-\gamma)^2}\Big). \tag{A.18}$$

To handle infinite hypothesis classes, we notice that under Assumption 9, the composite function g is Lipschitz as a function of Q_1, Q_2 and Q_3, with the Lipschitz constant upper bounded by $O(R_{\max}/(1-\gamma))$. According to e.g., Lemma A.6 of Chernozhukov et al. (2014), the function class $\mathcal{G}$ belongs to the VC type class as well, with the envelop function upper bounded by $O(R_{\max}^2/(1-\gamma)^2)$. Consequently, we can find an ϵ-net of $\mathcal{G}$, denoted by $\mathcal{G}_0$, with ϵ proportional to $(NT)^{-1}R_{\max}^2/(1-\gamma)^2$,

such that restricting to the finite hypothesis class $\mathcal{G}_0$ provides an reasonable approximation for $\mathcal{G}$ with the approximation error upper bounded by

$$O\Big(\frac{R_{\max}^2}{(1-\gamma)^2 NT}\Big). \tag{A.19}$$

Meanwhile, the number of elements in $\mathcal{G}_0$ is of the order $O(N^V T^V)$ for some constant $V > 0$.

Finally, to handle non i.i.d data, we notice that the sequence $\{S_{i,t}\}_{t>T}$ is exponentially β-mixing; see Step 1 of the proof of Lemma 1. Consequently, we can invoke the Bernstein's inequality designed for martingales (see e.g., Dzhaparidze and Van Zanten, 2001) and exponentially β-mixing time series (Chen and Christensen, 2015, Theorem 4.2) to obtain a uniform upper bound for (A.16), when restricting to the class of functions $g \in \mathcal{G}_0$. Other arguments are similar to those in the proof of Lemma 16 of Chen and Jiang (2019) and the proof of Step 1 of Lemma 1. More specifically, it can be shown that wpa1, (A.17) can be upper bounded by

$$O\Big(\frac{R_{\max}^2 \log^2(NT)}{(1-\gamma)^2 NT}\Big) + O\Big(\frac{R_{\max}\log(NT)\|Q_2 - Q_3\|_{\mu_k}}{(1-\gamma)\sqrt{NT}}\Big), \tag{A.20}$$

uniformly in $g \in \mathcal{G}_0$. This together with (A.18) and (A.19) yields that, when considering the unrestricted function class $\mathcal{G}$, (A.16) can be upper bounded by

$$O\Big(\frac{\log^3(NT) R_{\max}^2}{NT s_{cp}^2 (1-\gamma)^2}\Big) + O\Big(\frac{R_{\max}^2 \log^2(NT)}{(1-\gamma)^2 NT}\Big) + O\Big(\frac{R_{\max}\log(NT)\|Q_2 - Q_3\|_{\mu_k}}{(1-\gamma)\sqrt{NT}}\Big), \tag{A.21}$$

Under the completeness assumption in Assumption 8, we have $\mathcal{B}_k^* \widehat{Q}_{s,k}^{(j-1)} \in \mathcal{Q}$ for any j. Let Q_1 denote $\widehat{Q}_{s,k}^{(j-1)}$, Q_2 denote $\widehat{Q}_{s,k}^{(j)}$ and Q_3 denote $\mathcal{B}_k^* \widehat{Q}_{s,k}^{(j-1)}$. Since $\widehat{Q}_{s,k}^{(j)}$ is the empirical risk minimiser, it follows that

$$\begin{aligned}
&\frac{1}{|\widehat{\mathcal{C}}_k|\widehat{\tau}^{(k)}} \sum_{i\in\widehat{\mathcal{C}}_k} \sum_{t=T-\widehat{\tau}^{(k)}}^{T-1} \Big[R_{i,t} + \gamma \max_a \widehat{Q}_{s,k}^{(j-1)}(S_{i,t+1}, a) - \widehat{Q}_{s,k}^{(j)}(S_{i,t}, A_{i,t})\Big]^2 \\
\leq\ & \frac{1}{|\widehat{\mathcal{C}}_k|\widehat{\tau}^{(k)}} \sum_{i\in\widehat{\mathcal{C}}_k} \sum_{t=T-\widehat{\tau}^{(k)}}^{T-1} \Big[R_{i,t} + \gamma \max_a \widehat{Q}_{s,k}^{(j-1)}(S_{i,t+1}, a) - \mathcal{B}_k^* \widehat{Q}_{s,k}^{(j-1)}(S_{i,t}, A_{i,t})\Big]^2.
\end{aligned}$$

This together with the established uniform upper error bound implies that

$$-\frac{1}{|\widehat{\mathcal{C}}_k|\widehat{\tau}^{(k)}} \sum_{i\in\widehat{\mathcal{C}}_k} \sum_{t=T-\widehat{\tau}^{(k)}}^{T-1} g_{i,t}^*(\widehat{Q}_{s,k}^{(j-1)}, \widehat{Q}_{s,k}^{(j)}, \mathcal{B}_k^* \widehat{Q}_{s,k}^{(j-1)})$$

where $g_{i,t}^*(Q_1, Q_2, Q_3) = \mathbb{E}[g(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1}; Q_1, Q_2, Q_3)]$, is upper bounded by (A.21), wpa1.

The above expression can be decomposed into two terms, given by

$$-\frac{1}{|\widehat{\mathcal{C}}_k|\widehat{\tau}^{(k)}} \sum_{i\in\widehat{\mathcal{C}}_k} \sum_{t=T-\widehat{\tau}^{(k)}}^{T-\min(\tau^{(k)},\widehat{\tau}^{(k)})} g_{i,t}^*(\widehat{Q}_{s,k}^{(j-1)}, \widehat{Q}_{s,k}^{(j)}, \mathcal{B}_k^* \widehat{Q}_{s,k}^{(j-1)})$$

and

$$-\frac{1}{|\widehat{\mathcal{C}}_k|\widehat{\tau}^{(k)}}\sum_{i\in\widehat{\mathcal{C}}_k}\sum_{t=T-\min(\tau^{(k)},\widehat{\tau}^{(k)})}^{T-1} g^*_{i,t}(\widehat{Q}^{(j-1)}_{s,k},\widehat{Q}^{(j)}_{s,k},\mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}).$$

Under similar arguments in proving (A.18), the first term can be lower bounded by

$$-O\Big(\frac{\log^3(NT)R^2_{\max}}{NTs^2_{cp}(1-\gamma)^2}\Big). \tag{A.22}$$

As for the second term, notice that $\mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}$ is the argmin of $g^*_{i,t}(\widehat{Q}^{(j-1)}_{s,k},\widehat{Q}^{(j)}_{s,k},\bullet)$ whenever $i\in\mathcal{C}_k$ and $t\geq\tau^{(k)}$. According to Theorem 1, we have $\widehat{\mathcal{C}}_k=\mathcal{C}_k$ wpa1. It follows that the second term can be represented by

$$\frac{1}{|\mathcal{C}_k|\widehat{\tau}^{(k)}}\sum_{i\in\mathcal{C}_k}\sum_{t=T-\min(\tau^{(k)},\widehat{\tau}^{(k)})}^{T-1}\mathbb{E}\Big[Q_3(S_{i,t},A_{i,t})-Q_2(S_{i,t},A_{i,t})\Big]^2,$$

with Q_2 being $\widehat{Q}^{(j)}_{s,k}$ and Q_3 being $\mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}$. Under the boundedness assumption on the transition function in Assumption 3, the density function $S_{i,t}$ is bounded away from zero. It follows from the change of measure theorem that the second term can be lower bounded by $c\|\widehat{Q}^{(j)}_{s,k}-\mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}\|^2_{\mu_k}$ for some constant $c>0$.

Combining these results together with (A.21) yields that

$$c\|\widehat{Q}^{(j)}_{s,k}-\mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}\|^2_{\mu_k}\leq\frac{CR^2_{\max}\log^3(NT)}{(1-\gamma)^2NTs^2_{cp}}+\frac{CR_{\max}\log(NT)\|\widehat{Q}^{(j)}_{s,k}-\mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}\|_{\mu_k}}{(1-\gamma)\sqrt{NT}},$$

for some constant $C>0$. Using Cauchy-Schwarz inequality, the last term can be upper bounded by

$$\frac{c\|\widehat{Q}^{(j)}_{s,k}-\mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}\|^2_{\mu_k}}{2}+\frac{C^2R^2_{\max}\log^2(NT)}{2c(1-\gamma)^2NT}.$$

Consequently, $\|\widehat{Q}^{(j)}_{s,k}-\mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}\|_{\mu_k}$ can be upper bounded by

$$O\Big(\frac{R_{\max}\log^{3/2}(NT)}{(1-\gamma)\sqrt{NT}s_{cp}}\Big).$$

This together with (A.15) yields the desired upper error bound.

Finally, under the assumption that $N\gg s^{-2}_{cp}\log^3(NT)$, the change point error equals exactly zero, as proven in Corollary 1. In this case, both the first term in (A.21) and the lower bound in (A.22) become zero. Using the same arguments, it can be shown that $\|\widehat{Q}^{(j)}_{s,k}-\mathcal{B}^*_k\widehat{Q}^{(j-1)}_{s,k}\|_{\mu_k}$ can be upper bounded by

$$O\Big(\frac{R_{\max}\log(NT)}{(1-\gamma)\sqrt{NT}}\Big),$$

instead. The proof is hence completed. □

A.8 Inconsistencies of the Homogeneous and Doubly Homogeneous algorithms

Consider an MDP where the states are i.i.d. over time and population, independent of the rewards and actions. Consider settings with binary reward. Assume there exist two subgroups. For subjects belonging to the first subgroup, their reward equals 2 if they select action 1, and 0 otherwise. For those in the second subgroup, their reward equals 1 if they select action 0, and 0 otherwise. By definition, it is immediate to see that the optimal policy will assign all subjects within the first subgroup action 1, and all subjects within the second subgroup action 0. This yields an expected cumulative reward of

$$\frac{1}{2}\sum_{t=0}^{\infty}\gamma^t\times 2+\frac{1}{2}\sum_{t=0}^{\infty}\gamma^t\times 1=\frac{3}{2(1-\gamma)}.$$

Since the transition function does not change, no change point will be identified by the Homogeneous algorithm. As such, both the Homogeneous and the Doubly Homogeneous algorithms will apply FQI to learn the optimal policy based on the entire offline dataset. Suppose the initial Q-function is a zero function. At the first iteration, its learned Q-function will converge to

$$Q^{(1)}(a,s)=\begin{cases}\frac{1}{2}\times 2+\frac{1}{2}\times 0=1 & a=1\\ \frac{1}{2}\times 1+\frac{1}{2}\times 0=\frac{1}{2} & a=0\end{cases}$$

as either N or $T\to\infty$. At the second iteration, its learned Q-function will converge to

$$Q^{(2)}(a,s)=\begin{cases}\frac{1}{2}\times 2+\frac{1}{2}\times 0+\gamma\max(1,1/2)=1+\gamma & a=1\\ \frac{1}{2}\times 1+\frac{1}{2}\times 0+\gamma\max(1,1/2)=\frac{1}{2}+\gamma & a=0\end{cases}$$

Following the same logic, it can be shown that the learned optimal Q-function will converge to

$$\begin{cases}\frac{1}{1-\gamma} & a=1\\ \frac{1}{1-\gamma}-\frac{1}{2} & a=0\end{cases}$$

Consequently, the learned policy will assign action 1 to all subjects, as it ignores heterogeneity over population. Apparently, this policy is sub-optimal, which would incur a regret of

$$\frac{1}{2}\sum_{t=0}^{\infty}\gamma^t\times 1-\frac{1}{2}\sum_{t=0}^{\infty}\gamma^t\times 0=\frac{1}{2(1-\gamma)}.$$

This completes the proof.

A.9 Inconsistency of the Stationary algorithm

The MDP can be similarly constructed to that in Section A.8. Specifically, we assume the action is binary and all states are i.i.d. over time and population. Additionally, suppose all subjects share the same transition function. However, there exists a single change point at location $T/2+1$. Specifically, prior to the change, the reward equals 2 if action 1 is selected, and 0 otherwise. After the change, the reward equals 1 if action 0 is selected, and 0 otherwise.

Suppose there is no change point after time T. Then the optimal policy will assign action 0 to all subjects, leading to an expected cumulative reward of

$$\sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}.$$

Since all subjects share the same data distribution, the Stationary algorithm will only identify one cluster in the offline data. Similar to Section A.8, the optimal policy selected by FQI will assign action 0 to all subjects. Again, this policy is apparently sub-optimal, with a regret of $(1-\gamma)^{-1}$. This completes the proof.

B Semi-synthetic data simulation

B.1 Offline Data Generating Process

The semi-synthetic data setup is designed based on the analysis of IHS conducted in Section **??**, which identified two distinct clusters of interns and associated change points at $T_{train} = 82$. At each time point $t = 0, \ldots, T$, the binary action for the i-th individual is randomly generated with $P(A_{i,t} = 1) = 1 - P(A_{i,t} = -1) = 0.5$. The state vector $S_{i,t}$ comprises three variables. The state variables are initiated at $t = 0$ as independent normal distributions with $S_{i,0,1} \sim \mathcal{N}(0, 1)$ for $i = 1, 2, 3$. The kth transition dynamic takes this form: $S_{i,t+1} = (\boldsymbol{B} + \delta \boldsymbol{G}_k)\,(A_{i,t}, S_{i,t}^\top, A_{i,t} * S_{i,t}^\top)^\top + \epsilon_{i,t}$. Here, $\boldsymbol{B}$ represents the effect of the current state on the next state in a reference dynamic (in our example, the $\mathcal{P}_1$), $\boldsymbol{G}_k$ represents the difference between the effect of the current state in dynamic k and the reference dynamic, and $\epsilon_{i,t} \sim \mathcal{N}_3(0, \mathrm{diag}(0.25, 0.25, 0.25))$. Here, $\mathrm{diag}(0.25, 0.25, 0.25)$ denotes a diagonal matrix with diagonal elements 0.25, 0.25, and 0.25. In addition, $\delta \in \{2, 1, 0.6\}$ is a factor that controls strong, moderate, and weak signal in the change of transition dynamics, respectively.

The base transition matrix is

$$\mathbf{B} = \begin{pmatrix} 0.107 & 0.005 & 0.372 & 0.025 & -0.002 & -0.005 & 0.013 & 0.028 \\ -0.099 & -0.014 & 0.038 & 0.1 & 0.005 & -0.006 & -0.007 & 0.008 \\ 0.005 & -0.007 & 0.002 & -0.015 & 0.475 & -0.002 & 0.005 & -0.002 \end{pmatrix}$$

and

$$\mathbf{G}_2 = \begin{pmatrix} 0.048 & 0.015 & 0.053 & 0.037 & -0.000 & -0.036 & -0.048 & -0.013 \\ -0.127 & 0.030 & 0.085 & -0.039 & 0.016 & 0.024 & 0.024 & 0.018 \\ -0.033 & 0.021 & 0.037 & -0.004 & -0.012 & -0.038 & -0.000 & 0.041 \end{pmatrix},$$

$$\mathbf{G}_3 = \begin{pmatrix} -0.234 & -0.010 & -0.094 & -0.011 & 0.002 & 0.021 & -0.044 & -0.038 \\ 0.198 & 0.010 & -0.026 & -0.001 & 0.019 & -0.005 & 0.037 & -0.023 \\ -0.011 & 0.006 & 0.000 & -0.001 & 0.053 & -0.016 & 0.006 & 0.008 \end{pmatrix},$$

$$\mathbf{G}_4 = \begin{pmatrix} -0.358 & -0.092 & -0.187 & 0.012 & 0.003 & 0.020 & 0.112 & -0.042 \\ 0.427 & 0.058 & 0.018 & -0.035 & -0.004 & 0.007 & -0.143 & -0.006 \\ 0.044 & -0.022 & -0.006 & -0.023 & 0.035 & 0.007 & 0.040 & 0.017 \end{pmatrix}.$$

The transition functions for all subjects of the are specified as the followings. For the first 20 subjects, they follow $\mathcal{P}_1$ from $t = 0$ to $t = 49$ and switch to $\mathcal{P}_2$ after $t = 49$:

$$S_{i,t+1} = \begin{cases} \mathbf{B}(A_{i,t}, S_{i,t}^\top, A_{i,t} * S_{i,t}^\top)^\top + \epsilon_{i,t} & \text{if } t \in [0, 49] \\ (\mathbf{B} + \delta \mathbf{G}_2)(A_{i,t}, S_{i,t}^\top, A_{i,t} * S_{i,t}^\top)^\top + \epsilon_{i,t} & \text{if } t \in [50, 99] \end{cases}$$

Subjects 21 to 40 follow $\mathcal{P}_3$ from $t = 0$ to $t = 49$ and switch to $\mathcal{P}_4$ after $t = 49$:

$$S_{i,t+1} = \begin{cases} (\mathbf{B} + \delta\mathbf{G}_3)(A_{i,t}, S_{i,t}^\top, A_{i,t} * S_{i,t}^\top)^\top + \epsilon_{i,t} & \text{if } t \in [0, 49] \\ (\mathbf{B} + \delta\mathbf{G}_4)(A_{i,t}, S_{i,t}^\top, A_{i,t} * S_{i,t}^\top)^\top + \epsilon_{i,t} & \text{if } t \in [50, 99] \end{cases}$$

The remaining 20 subjects follow $\mathcal{P}_4$ throughout the entire period:

$$S_{i,t+1} = (\mathbf{B} + \mathbf{G}_4)(A_{i,t}, S_{i,t}^\top, A_{i,t} * S_{i,t}^\top)^\top + \epsilon_{i,t} \quad \text{for } t \in [0, 99]$$

B.2 Implementation of change point detection and clustering

To construct the log-likelihood ratio test (LRT) statistic proposed in Section **??**, we need to estimate the conditional distribution $p(S_{it}|S_{it-1}, A_{it-1})$. We parameterised $p(S_{it}|S_{it-1}, A_{it-1})$ using a Gaussian distribution, by fitting linear regression model $S_{i,t+1} = \boldsymbol{\beta}\boldsymbol{X}_{i,t} + \epsilon_{i,t}$, where $\boldsymbol{X}_{i,t} = (A_{i,t}, S_{i,t}^\top, A_{i,t}S_{i,t}^\top)^\top$ consists of the state, action, and their interaction, and $\boldsymbol{\beta}$ is the regression coefficient. The variance of the residual $\epsilon_{i,t}$ is assumed constant in time.

When calculating the rejection threshold of the LRT, we noticed that the asymptotic distribution of the CUSUM statistic depends on the data generating process only through its degree of freedom, thanks to the Markov assumption. As such, to approximate the threshold, we first sampled i.i.d. d-dimensional standard multivariate Gaussian vectors $\{Z_{i,t}\}_{i\in\widehat{\mathcal{C}}_k, \tau\le t\le T-1}$, where d equals the degrees of freedom, or equivalently the number of nonzero coefficients in $\widehat{\boldsymbol{\beta}}$. For a given τ, we next computed the maximum of the squared weighted Euclidean distances: $\max\left\{(\bar{Z}_{[T-\tau,u)} - \bar{Z}_{[u,T]})^2; T - \tau < u < T\right\}$, where $\bar{Z}_I$ is the sample mean of $Z_{i,t}$'s on the time interval I over all i's. We repeated this process to generate 10000 samples of the asymptotic reference distribution. We then used the 0.01 empirical upper quantile of the distribution as the rejection threshold. For theoretical analysis, we set the threshold to be proportional to $\bar{C}\log^2(NT)\log(\log(NT))$ for some sufficiently constant $\bar{C} > 0$ so that the resulting test is consistent theoretically.

To implement the clustering algorithm, we adopt the proposal by Bonhomme and Manresa (2015) which minimises a least square objective function. Given a set of initial change points and a number of clusters, we iterate between the two subroutines twice to update the estimated change points and cluster membership.

B.3 Implementation Details of Online Evaluation

To estimate the optimal policy, we coupled FQI with decision tree regression to compute the Q-estimator. The discounted factor γ is set to be 0.9. The hyperparameters in the decision tree model, the maximum tree depth and the minimum number of samples on each leaf node, were selected using 5-fold cross validation from $\{3, 5, 6\}$ and $\{50, 60, 80\}$, respectively. We considered an online setting and simulated potential change points start from $t = 100$ from a Poisson process with rate 1/60. Accordingly, a new change point was expected to occur every 40 time points. We set the terminal time to 400, yielding 4 to 5 potential change points in most replications.

The online data were simulated in the following manner. We consider the most recent two transition dynamics described above with $S' = (\boldsymbol{B} + \boldsymbol{G}_k)(A_{i,t}, S_{i,t}^\top, A_{i,t} * S_{i,t}^\top)^\top + \epsilon_{i,t}$ for $k \in \{2, 4\}$. The subjects $i \in [1, 20]$, $i \in [21, 40]$ and $i \in [41, 60]$ are considered to belong to three underlying clusters within which all cluster members share the same dynamic at any time during the whole online data generating process. Whenever a new change point occurs in a cluster, the first two underlying clusters separately decide to change its transition dynamics change into another dynamics

with probability 0.5, or otherwise remain constant. The last underlying cluster adopt dynamic $\mathcal{P}_2$ for all time. This yields a total of five possible scenarios, namely, merge, split, switch, evolution and constancy (see Figure **??** in the main text). In addition, the reward function is assumed doubly homogeneous and equals $R_{i,t} = S_{i,t+1,1}$, i.e., the first dimension of the state variable for all individuals and time points.

Finally, we assumed that online data came in batches regularly at every $L = 25$ time points starting from $t = 100$. This yielded a total of 12 batches of data. The first data batch was generated according to the estimated optimal policy $\widehat{\pi}_k$ computed based on the data subset $\{O_{i,t} : i \in \widehat{\mathcal{C}}_k, t \in [T - \widehat{\tau}_i^*, T]\}, k = 1, \ldots, K$. Let $T_0^* = \max_i T - \widehat{\tau}_i^*$. Suppose b batches of data were received. We first applied the proposed change point detection and clustering method on the data subset in $[T_{b-1}^*, T + bL]$ to identify new change points and clusters. If there was at least one change point, we set $T_b^* = \max_i\{T - \widehat{\tau}_{i,b}^*\}$. If no changes were detected, we set $T_b^* = T_{b-1}^*$. We next updated the optimal policy using the data subset $\{O_{i,t} : i \in \widehat{\mathcal{C}}_k, t \in [T - \widehat{\tau}_{i,b}^*, T + bL]\}, k = 1, \ldots, K$ for all the current clusters and used the updated optimal policy (combined with the ϵ-greedy algorithm) to generate the $(b + 1)$-th data batch. We repeated this procedure until all 12 batches of data were received. Finally, we computed the average rewards obtained from time 100 to 400 for the 60 subjects.

C Additional implementation details for analysing the IHS dataset

All data is standardized to have a mean of zero and unit variance before training. The implementation of the proposed algorithm and fitted-Q iterations follows the specifications in Section B. For off-policy evaluation (OPE), we assume that cluster membership in the testing set remains the same as detected by the proposed method in the training set, and there is no change point in the testing set. Therefore, we conduct separate OPE for each cluster in the testing set and report the average OPE result, weighted by the number of subjects in each cluster. We implement the fitted-Q evaluation algorithm using decision tree regression to compute the Q-estimator. The hyperparameters and tuning method are the same as those used for the FQI in Section B. The average value reported in Table **??** is calculated by multiplying the discounted cumulative step counts by $1 - \gamma = 0.1$.

References

Bonhomme, S. and Manresa, E. (2015) Grouped patterns of heterogeneity in panel data. *Econometrica*, **83**, 1147–1184.

Bradley, R. C. (2005) Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, **2**, 107–144.

Casella, G. and Berger, R. (2024) *Statistical inference*. CRC Press.

Chen, J. and Jiang, N. (2019) Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 1042–1051. PMLR.

Chen, X. and Christensen, T. M. (2015) Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, **188**, 447–465. URL: `https://ideas.repec.org/a/eee/econom/v188y2015i2p447-465.html`.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2014) Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, **42**, 1564–1597.

Dzhaparidze, K. and Van Zanten, J. (2001) On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, **93**, 109–117.

Hogg, R. V. and Craig, A. T. (1995) Introduction to mathematical statistics. *Englewood Hills, New Jersey*.

Meitz, M. and Saikkonen, P. (2019) Subgeometric ergodicity and beta-mixing. *arXiv preprint arXiv:1904.07103*.

Shi, C., Zhang, S., Lu, W. and Song, R. (2022) Statistical inference of the value function for reinforcement learning in infinite horizon settings. *Journal of Royal Statistical Society: Series B*, **84**.