Heterogeneous Reinforcement Learning with Offline Data: Estimation and Inference

Elynn Y. Chen [*1], Zhiyue T. Hu [†2], Rui Song [‡3], and Michael I. Jordan [§4]

[1,4]EECS, University of California, Berkeley.
[2]Biostatistics, University of California, Berkeley.
[3]Statistics, North Carolina State University

December 16, 2020

Abstract

Reinforcement Learning (RL), equipped with large-scale datasets, will provide powerful data-driven supports to a wide range of decision making problems in healthcare, education, business, and more. Classical RL methods focus on the mean of total return and, thus, may provide misleading results for the heterogeneous populations in large scale datasets. This paper introduces K-heterogeneous MDP to characterize the sequential decision problems with heterogeneous data, and proposes an Auto-Clustered Policy Iteration (ACPI) that can automatically detect and identify homogeneous sub-population, while learning the Q function and the optimal policy for each sub-population. We establish convergence rates and construct confidence intervals for the estimators obtained by the ACPI. Simulations are conducted to back up our theoretical findings. Empirical study on the well-recognized MIMIC-III dataset shows evidences of value heterogeneity and confirms the advantage of our new method.

1 Introduction

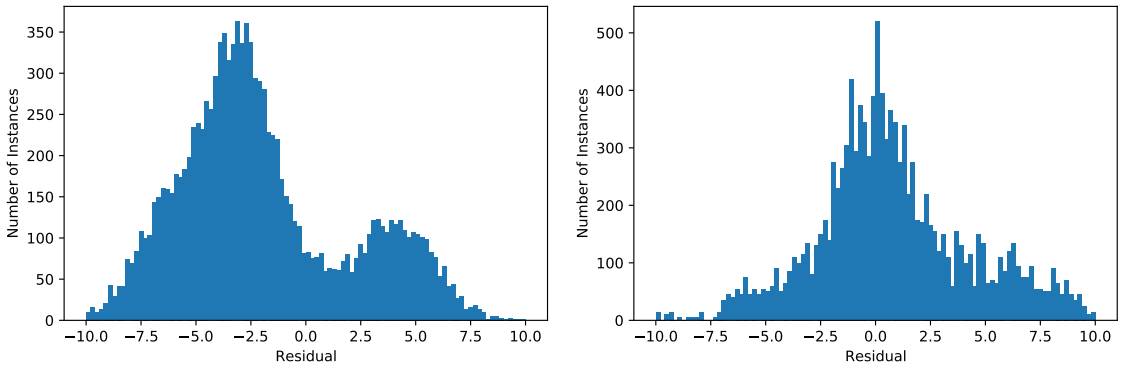
Many real world problems involve making decisions sequentially based on available data that is influenced by the previous decisions. For example, in clinical practice, physi-

[*]Supported in part by NSF Grants DMS-1803241. Email: elynn.chen@berkeley.edu
[†]Email: zyhu95@berkeley.edu.
[‡]Supported in part by NSF grant DMS-1555244 and NCI grant P01 CA142538. Email: rsong@ncsu.edu
[§]Email: jordan@cs.berkeley.edu

cians make a series of treatment decisions over the course of a patient's disease based on his/her baseline and evolving characteristics (Schulte et al., 2014). In education, human and automated tutors attempt to choose pedagogical activities that will maximize student learning, informed by their estimates of the student's current knowledge (Rafferty et al., 2016; Reddy et al., 2017). The framework of reinforcement learning (RL), powered up with large scale dataset, may provide enormous supports to such a wide range of decision-making domains, from healthcare, education, and business to scientific researches (Schulte et al., 2014; Mandel, 2017; Zhou et al., 2017).

Reinforcement learning utilizes Markov Decision Processes (MDP) $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \Pr, r, \gamma)$ to mathematically formalize such dynamic decision problems. The interaction between an agent and an environment evolves as follows. At time t, an agent observes the current system state $\boldsymbol{x}_t$ supported on the set of states $\mathcal{X}$. She chooses a decision a_t supported on the set of actions $\mathcal{A}$ and transit to the next state $\boldsymbol{x}_{t+1}$ according to the unknown system transition probability $\Pr(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, a_t)$. At the same time, she receives an immediate reward $r(\boldsymbol{x}_t, a_t)$ which serves as a *partial* signal of the goodness of her choice at time t. The real performance metric is the sum of discounted rewards, or *return*, defined as $Y_T = \sum_{t=1}^{T} \gamma^t r(\boldsymbol{x}_t, a_t)$ where the horizon T can be finite or infinite. The discount factor $\gamma \in [0, 1)$ reflects a trade-off between immediate and future rewards. An agent's decision making rule is characterized by a policy $\pi(a_t|\boldsymbol{x}_t)$ that defines a distribution over actions conditioned on states and the final goal of reinforcement learning is to learn a near-optimal policy that maximizes the return Y_T.

Classical RL methods learns a state (-action) value function $V(\boldsymbol{x})$ (or $Q(\boldsymbol{x}, a)$), which is an *expectation* of the return $\sum_{t=1}^{T} \gamma^t r(\boldsymbol{x}_t, a_t)$ and chooses optimal policy to maximize this expected return. Such mean value-based RL methods have made tremendous advances in algorithms, theories, as well as applications. See Sutton and Barto (2018) and references therein. However, naive applications of the mean value-based RL methods to large-scale datasets may generate misleading results because they are often created via aggregating many heterogeneous data sources. Each sub-population may exhibit different associations between state variable, action and outcomes (such as state transition, immediate rewards, and return). Take the MIMIC-III dataset as a example (Johnson et al., 2016; Komorowski et al., 2018) . We learned a homogeneous linear Q function with the sepsis cohort data and plot the estimation residual in Figure 1 (a). It is obvious that there are two sub-population in the dataset. The mean value-based Q learning fails to capture the heterogeneity in the dataset. The results based on integrated value over the entire population is misleading, even dangerous in high-risk applications such as health care, finance and autonomous driving.



(a) Residual distribution of simple linear regression is clearly bi-model, with two peaks at around -4 and 4.

(b) Residual of our algorithm is bell-shaped, with one peak at around 0. It also performs better in terms of prediction error.

Figure 1: Heterogeneity in the MIMIC-III dataset and the improvement achieved by ACPI.

In this paper, we introduce a framework of K-heterogeneous MDP (K-hetero MDP) to characterize the problem of dynamic decision making with heterogeneous population and propose an Auto-Clustered Policy Iteration (ACPI) algorithm that can automatically detect and identify homogeneous sub-populations from a data set while learning the Q function and optimal policy for each sub-population. We establish statistical guarantees for the estimators of the proposed ACPI. Specifically, we obtained convergence rates and construct confidence intervals (CIs) for the estimated model parameters and value functions. These theoretical results are especially important in high-risk applications. We apply the proposed method on the widely-recognized MIMIC-III dataset. Figure 1 (b) presents the residual after applying the ACPI to the same MIMIC-III dataset. It is clear that the ACPI captures all the heterogeneity in the population. More results in the application section further confirm the advantage of the our new method.

The contribution of our work are summarized as follows. Methodologically, we introduce a formal definition of K-hetero MDP and propose the new ACPI algorithm to deal with this heterogeneity automatically. The efficacy of the method is validated through simulations and on a real large-scale dataset. While this paper focus on the batch-mode estimations, the framework of K-hetero MDP is generally applicable and the ACPI can be extended to deal with online estimation. Theoretically, we establish convergence rates and construct CI under the bidirectional-asymptotic framework that allows either N or T goes to infinity. Our CI is valid as long as either the number of trajectories N or the length of trajectory T diverge to infinity. It covers a wide variety of real applications and is especially useful when the number of trajectories in a sub-population is small. We also

offer theoretical guidance to practitioners on the choice of tuning parameters in ACPI. Technically, in order to study the properties of the estimators by the ACPI, we study the TD error and derive its infinity norm convergence result. This result is of independent importance, especially in studying RL algorithms with penalization (Song et al., 2015).

1.1 Related works

This paper is closely related to three different research areas in Machine Learning and Statistics, namely distributional RL, dynamic treatment regime and mixture models.

Distributional RL In contrast to classic mean-value-based RL, Bellemare et al. (2017) introduces the *distributional RL (DRL)* that emphasis on the full distribution of the random return Y_T. Rowland et al. (2019) presents an example of multimodality in value distributions. The main approaches to DRL include learning discrete *categorical* distribution (Bellemare et al., 2017; Rowland et al., 2018; Barth-Maron et al., 2018), learning distribution *quantiles* (Dabney et al., 2018; Yang et al., 2019), and learning distribution *expectiles* Rowland et al. (2019). However, these work mainly focus on the computational issues and empirical performances. No statistical properties of the estimators are provided, partly because of the fundamental difficulty of estimating a full distribution.

An important concept introduced in the series of DRL paper is *Bellman closeness* of a statistics. Lemma 4.2 and Theorem 4.3 in Rowland et al. (2019) show that collections of moments are effectively the only finite sets of statistics that are Bellman closed. This inspires us to estimate the sub-population means instead of a full distribution of the super-population: the means are Bellman closed but the discretized distribution or the quantiles is not. Between the spectrum of mean-value-based RL and distributional RL, our method is shown to capture the heterogeneity while avoid the sample inefficiency in estimating a full distribution.

Dynamic treatment regime Researchers in dynamic treatment regime (DTR) have used the RL framework and MDP to derive a set of sequential treatment decision rules, or treatment regimes, to optimize patients' clinical outcomes over a fixed period of time. See Murphy (2003); Robins (2004); Schulte et al. (2014); Luedtke and Van Der Laan (2016); Ertefaie and Strawderman (2018); Zhu et al. (2019); Shi et al. (2020) and references therein. Most methods in DTR are designed for finite horizon T and are implemented through a backward recursive fitting procedure that is related to dynamic programming algorithm (Schulte et al., 2014). For infinite horizon T, Ertefaie and Strawderman (2018)

propose a variant of greedy gradient Q learning (GGQ) to estimate optimal dynamic treatment regimes. It requires modeling a non-smooth function of the data, which creates complications. Luckett et al. (2019) propose a V learning that models the state value function and directly search for the optimal policy among a pre-specified class of policies. Although their setting allows for infinite horizon, the information accumulated over horizon T does not enter explicitly in their results. In contrast, we show that the convergence rate of our estimators are of the order $(NT)^{-1/2}$ under the similar assumptions. Shi et al. (2020) focus on policy evaluation and confidence interval (CI) of a (possibly data-dependent) policy's value function but did not consider the control problem.

Moreover, the personalized treatment regime in DTR aims to offer individualized treatment to each patient i according to different state variable $\boldsymbol{x}_{i,t}$. The model coefficients are the same for all the patients. In contrast, the heterogeneity in this paper stems from different model coefficients. Our problem is more difficult since we need to estimate the model coefficients and, at the same time, to group them into a few clusters.

Mixture models Under the linear regression setting, researchers have studied the problem of supervised clustering that explores homogeneous effects of covariates along coordinates and across samples. The mixture model-based approach needs to specify an underlying distribution for data, and also requires specification of the number of mixture components in the population. Alternative approach to model-based clustering analysis employ grouping penalization on pairwise differences of the data points. For example, assuming parameter homogeneity over individuals, Tibshirani et al. (2005), Bondell and Reich (2008), Shen and Huang (2010) and Ke et al. (2013) employ different penalty functions to each pair of coordinates of the coefficient vector to group similar-effect covariates along coordinate. Assuming parameter heterogeneity over individuals, Hocking et al. (2011), Pan et al. (2013), and Ma and Huang (2017) adopt a fusion-type penalty with either an L_p-shrinkage or a non-convex penalty function to formulate clusters across samples with the same regression coefficients.

Under the RL setting, the model-based approaches are not suitable because the distribution of the value function can not be easily characterized by Gaussian or Mixture Gaussian (Sobel, 1982). Therefore, the ACPI algorithm is developed based on the pairwise fusion-type penalty. However, our problem in RL is more challenging than that in linear regression. First, it is common to assume the noise in regression be i.i.d sub-Gaussian, however, the observations over T in RL are not uncorrelated which imposes great difficulties in our analysis. Second, we consider the semi-parametric approximation with diverging number of basis functions instead of the linear models. Lastly, our

ultimate goal is to learn an optimal policy, while the mixture regression focuses entirely on parameter estimation.

1.2 Notation and organization

We use x, $\boldsymbol{x}$, $\boldsymbol{X}$ to represent scalars, vectors and matrix, respectively. Capital letter X represents random variables. For any matrix $\boldsymbol{X}$, we use $\boldsymbol{x}_{i\cdot}$, $\boldsymbol{x}_{\cdot j}$, and x_{ij} to refer to its i-th row, j-th column, and ij-th entry, respectively. All vectors are column vectors and row vectors are written as $\boldsymbol{x}^\top$ for any vector $\boldsymbol{x}$. We denote the matrix ℓ_2-norm as $\|\boldsymbol{X}\|_2$ and max norm as $\|\boldsymbol{X}\|_{\max} \triangleq \max_{i,j} \left|x_{ij}\right|$. When $\boldsymbol{X}$ is a square matrix, we denote by $\mathrm{Tr}(\boldsymbol{X})$, $\lambda_{max}(\boldsymbol{X})$, and $\lambda_{min}(\boldsymbol{X})$ the trace, maximum and minimum eigenvalues of $\boldsymbol{X}$, respectively. We denote $\textsc{mat}(\boldsymbol{x})$ as the matricization of a vector $\boldsymbol{x}$ and $\textsc{vec}(\boldsymbol{X})$ as the vectorization of a matrix $\boldsymbol{X}$. We let $C, c, C_0, c_0, \ldots$ denote generic constants whose actual values may vary. As a convention, we denote the true, oracle and estimated values of β as $\mathring{\beta}$, $\widetilde{\beta}$, and $\widehat{\beta}$, respectively.

The rest of this paper is organized as follows. Section 2 introduces the problem setting and a few important concepts. Section 3 proposes the ACPI. Section 5 describes the computation procedure. Section 4 establishes the statistical properties of the estimators. Section 6 and Section 7 presents empirical results with synthetic and real datasets. Section 8 concludes. All proofs and technique lemmas are relegate to the supplementary material.

2 Statistical Framework

2.1 Markov decision process

Reinforcement learning addresses the problem of learning to control a dynamical system. The dynamic system is fully defined by a Markov decision process (MDP).

Definition 2.1 (Markov decision process.)**.** *The Markov decision process is defined as a tuple* $\mathcal{M} = \{\mathcal{X}, \mathcal{A}, \Pr, r, \gamma\}$*, where* $\mathcal{X}$ *is a set of state* $\boldsymbol{x} \in \mathcal{X}$*;* $\mathcal{A}$ *is a set of actions* $a \in \mathcal{A}$*;* $\Pr$ *is a Markovian state transition model –* $\Pr(\boldsymbol{x}_t, a_t, \boldsymbol{x}_{t+1})$ *is the probability of transition to state* $\boldsymbol{x}_{t+1}$ *when taking action* a_t *in state* $\boldsymbol{x}_t$*;* $r(\boldsymbol{x}_t, a_t)$ *defines the reward function; and* $\gamma \in (0, 1]$ *is a scalar discount factor.*

Let $t \in [T]$ index the time or events in the decision process that necessitate an action decision. At each time point t, the current status of the i-th individual, $i \in [N]$, is characterized by covariates $X_{i,t} \in \mathcal{X}$ and a finite set of all possible actions $\mathcal{A}$ is available. A

decision $A_{i,t} \in \mathcal{A}$ is selected and the state of the i-th individual changes to $X_{i,t+1}$ according to the transition probability $\Pr(X_{i,t+1} \mid X_{i,t}, A_{i,t})$. An immediate random scalar reward $R_{i,t}$ is observed. Overtime, these comprise a sample trajectory for the i-th individual and we have in total N sample trajectories $\{(X_{i,t}, A_{i,t}, R_{i,t})\}_{0 \le t \le T}$ for $i \in [N]$.

With respect to the data generating model, we assume that it is a *Time-Homogeneous Markov process (TH Markov)* and the observed rewards satisfies the Conditional Mean Independence (CMI) assumption.

Assumption 2.2 (Data generating MDP.)**.** *The sample trajectories* $\left\{\{(X_{i,t}, A_{i,t}, R_{i,t})\}_{t \ge 0}\right\}_{i \in [N]}$ *are generated from (possibly N different) Markov Decision Process satisfying*

(a) (Markov) $X_{i,t} \perp \{(X_{i,s}, A_{i,s}, Y_{i,s})\}_{0 \le s \le t-1} \mid (X_{i,t-1}, A_{i,t-1})$ *for any* $i \in [N]$.

(b) (Time-Homogeneous) The conditional density $\Pr(X_{i,t+1} \mid X_{i,t}, A_{i,t})$ *is the same over* t *for any* $i \in [N]$.

(c) (CMI) For any $i \in [N]$*, it holds that*

$$\mathbb{E}\left[R_{i,t} | X_{i,t} = \boldsymbol{x}, A_{i,t} = a, \{X_{i,s}, A_{i,s}, R_{i,s}\}_{0 \le s < t}\right] = \mathbb{E}\left[R_{i,t} | X_{i,t} = \boldsymbol{x}, A_{i,t} = a\right] = r_i(\boldsymbol{x}, a)$$

for some reward function $r_i(\boldsymbol{x}, a)$.

Remark 1. *The assumptions of the TH Markov and CMI are common in the literature. However, Assumption 2.2 is different because* $r_i(\boldsymbol{x}, a)$ *can be different across N trajectories.*

Remark 2. *Assumption 2.2 implies that* $R_{i,t} = r_i(X_{i,t}, A_{i,t}) + \eta_{i,t}$ *where* $\mathbb{E}[\eta_{i,t} \mid X_{i,t}, A_{i,t}] = 0$. *The* ℓ_2 *convergence and the asymptotic normality of the oracle estimators do not need any additional assumption on the noise term* $\eta_{i,t}$*. However, the uniform convergence that is needed for the feasible estimators requires* $\eta_{i,t}$ *be an sub-Gaussian random variable.*

2.2 Policy and value functions

A policy $\pi(a \mid \boldsymbol{x}) : \mathcal{X} \mapsto \mathcal{P}(\mathcal{A})$ is a function that maps the covariate space $\mathcal{X}$ to probability mass functions on the action space $\mathcal{A}$. It satisfies $\pi(a|\boldsymbol{x}) \ge 0$, for any $a \in \mathcal{A}$, $\boldsymbol{x} \in \mathcal{X}$ and $\sum_{a \in \mathcal{A}} \pi(a \mid \boldsymbol{x}) = 1$ for any $\boldsymbol{x} \in \mathcal{X}$. Under a policy π, an agent chooses action $A_{i,t}$ at state $X_{i,t}$ with probability $\pi(A_{i,t} \mid X_{i,t})$ and receives an immediate reward $R_{i,t}$. Over time, we observe a trajectory of length T, given by $\mathcal{H}_i = \{X_{i,0}, A_{i,0}, \cdots, X_{i,T}, A_{i,T}\}$, where T may be infinite. The trajectory distribution $\mathcal{P}_\pi$ for a given MDP $\mathcal{M}$ and a policy π is

$$\mathcal{P}_\pi(\mathcal{H}_i) = \mu_0(X_{i,0}) \prod_{t=0}^{T} \pi(A_{i,t}|X_{i,t}) \Pr(X_{i,t+1}|X_{i,t}, A_{i,t}). \tag{1}$$

The total return collected over time on a trajectory $\mathcal{H}_i$ is the accumulated discounted rewards defined as

$$Y(\mathcal{H}_i) = \sum_{t=0}^{T} \gamma^t R_{i,t}, \tag{2}$$

where the discount rate γ reflects a trade-off between immediate and future rewards. If $\gamma = 0$, the decision maker choose actions that maximize the immediate reward. As γ increase, the decision maker will put more weights on future rewards. The goal of the reinforcement learning is to learn the optimal policy $\pi^*(a \mid \boldsymbol{x})$ that maximize the total return $Y(\mathcal{H}_i)$.

Classic RL measures the goodness of a policy π by value functions defined as follows. Given a policy π and a discount factor $\gamma \in [0,1)$, the *state value function* is the expectation of the total return starting from a state $\boldsymbol{x}$:

$$V^{\pi}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{H}\sim\mathcal{P}_{\pi}(\mathcal{H})}\left[\sum_{t=0}^{T} \gamma^t R_{i,t} \middle| X_{i,0} = \boldsymbol{x}\right], \tag{3}$$

where the expectation is taken under the trajectory distribution generated by policy π on MDP $\mathcal{M}$. The *action value function* or *Q-function* of a given policy π is the expectation of the accumulated discounted rewards starting from a state $\boldsymbol{x}$ and taking action a:

$$Q^{\pi}(\boldsymbol{x}, a) = \mathbb{E}_{\mathcal{H}\sim\mathcal{P}_{\pi}(\mathcal{H})}\left[\sum_{t=0}^{T} \gamma^t R_{i,t} \middle| X_{i,0} = \boldsymbol{x}, A_{i,0} = a\right], \tag{4}$$

where the expectation is taken by assuming that the dynamic system follows the given policy π afterwards.

Similar to Theorem 6.2.12 of Puterman (1994), we can show that under the given conditions there exists at least on optimal policy π^* such that $V(\pi^*, \boldsymbol{x}) \geq V(\pi, \boldsymbol{x})$, $\forall\ \pi$ and $\boldsymbol{x}$. For continuous state set $\mathcal{X}$, it is also common to define an optimal policy using the *integrated* value function

$$V_{\mathcal{P}}(\pi) = \int V^{\pi}(\boldsymbol{x})\, d\mathcal{P}_x(\boldsymbol{x}), \tag{5}$$

which is defined with respect to a reference distribution $\mathcal{P}_x$ on the domain of $\boldsymbol{x}$. For a pre-specified class of policies, Π, the optimal policy, $\pi^*_{\mathcal{P}} \in \Pi$, satisfies $V_{\mathcal{P}}\left(\pi^*_{\mathcal{P}}\right) \geq V_{\mathcal{P}}(\pi)$ for all $\pi \in \Pi$. The estimated optimal policy is defined as

$$\pi^*_{\mathcal{P}} = \underset{\pi\in\Pi}{\arg\max}\ \widehat{V}_{\mathcal{P}}(\pi). \tag{6}$$

The goal is to estimate π^* or $\pi^*_{\mathcal{P}}$ using data collected previously from N trajectories, each of which has length $T_i = T$ for simple presentation and T may diverge to infinity.

2.3 Heterogeneous MDP and Bellman consistent equation

Since the learning objective is optimizing the total return, we define the heterogeneity in terms of the value functions – the expectation of return (2) in homogeneous sub-populations. Different from the classic RL, we allow different values, and thus different optimal policies, across heterogeneous super-population.

Definition 2.3 (K-Hetero MDP)**.** *A K-heterogeneous MDP is defined as a dynamic system with a latent variable $\omega \in [K]$ such that*

(i) For $\forall k \in [K]$, $\Pr(\omega = k) = P_k$ and it satisfies $P_k \in (0,1)$ and $\sum_{k=1}^{K} P_k = 1$.

(ii) The action value functions for a given π conditioned on $\omega = k$ are

$$V^{\pi,(k)}(\boldsymbol{x}) \equiv \mathbb{E}_{\mathcal{H}\sim\mathcal{P}_\pi(\mathcal{H})}\left[\sum_{t=0}^{T}\gamma^t R_{i,t}\middle| X_{i,0} = \boldsymbol{x}, w_i = k\right],$$
$$Q^{\pi,(k)}(\boldsymbol{x},a) \equiv \mathbb{E}_{\mathcal{H}\sim\mathcal{P}_\pi(\mathcal{H})}\left[\sum_{t\geq 0}\gamma^t R_{i,t}\middle| X_{i,0} = \boldsymbol{x}, A_{i,0} = a, w_i = k\right].$$

(iii) For any $k \neq k'$, the value functions of a given policy π are different, that is, $V^{\pi,(k)}(\boldsymbol{x}) \neq V^{\pi,(k')}(\boldsymbol{x})$ and $Q^{\pi,(k)}(\boldsymbol{x},a) \neq Q^{\pi,(k')}(\boldsymbol{x},a)$.

Remark 3. *Heterogeneity may also be defined in terms of MDP tuple (Definition 2.1), that is, $\mathcal{M}_k = (\mathcal{S}_k, \mathcal{A}_k, \Pr_k, r_k, \gamma_k)$, for $k \in [K]$, or in terms of trajectory distribution $\mathcal{P}^{(k)}_\pi(\mathcal{H}_i)$ for $k \in [K]$. However, Definition 2.3 provides a clearer framework to work with for the purpose of estimating optimal policy that maximizing the expected sum of discounted rewards. Since we may have the same value functions of some policy π for different $\mathcal{M}_k$ or $\mathcal{P}^{(k)}_\pi(\mathcal{H}_i)$, heterogeneity defined by $\mathcal{M}_k$ or $p^{(k)}_\pi(\mathcal{H}_i)$ unnecessarily complicates the problem.*

Now we allow for heterogeneity over different trajectories. Without knowing the true subgroups, it is safe for now to use different $Q^\pi_i(\boldsymbol{x},a)$ functions for different trajectory $i \in [N]$. The following Bellman consistent equation is the first moment condition that we will use to construct an auto-clustered estimation of the Q^π function.

Lemma 2.4. *Under the assumptions of TH Markov and CMI on $\mathcal{H}_i$, $1 \le i \le N$, we have the first moment condition of the Q_i^π function, for any function $\psi : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$,*

$$\mathbb{E}\left[R_{i,t} + \gamma \sum_{a' \in \mathcal{A}} Q_i^\pi(X_{i,t+1}, a')\, \pi(a'|X_{i,t+1}) - Q_i^\pi(X_{i,t}, A_{i,t}) \middle| X_{i,t}, A_{i,t}\right] = 0, \tag{7}$$

$$\mathbb{E}\left[\left(R_{i,t} + \gamma \sum_{a' \in \mathcal{A}} Q_i^\pi(\pi, X_{i,t+1}, a')\, \pi(a'|X_{i,t+1}) - Q_i^\pi(\pi, X_{i,t}, A_{i,t})\right) \psi(X_{i,t}, A_{i,t})\right] = 0. \tag{8}$$

3 Auto-Clustered Policy Iteration

The Auto-Clustered Policy Iteration is summarized in Algorithm 1. The algorithm contains two major parts: policy evaluation with heterogeneity, group-wise policy improvement. We explain each part of the algorithm in the sequel.

Algorithm 1: Auto-Clustered Q-learning (with parametric random policies)

Input: Number of sub-population K.
Model $Q(\pi, \boldsymbol{x}, a; \boldsymbol{\beta})$.
Parametric policies $\Pi = \{\pi(\alpha_k) : \boldsymbol{\alpha}_k \in \mathcal{S}_\alpha, 1 \le k \le K\}$.
Tuning parameter λ_{NT}.
Output: Optimal policies for group $\mathcal{G}_1, \cdots \mathcal{G}_K$ with parameter $\alpha_1, \cdots, \alpha_K$.
Data: Sample trajectories $\{(\boldsymbol{x}_{i,t}, a_{i,t}, r_{i,t})\}_{0 \le t \le T, 1 \le i \le N}$.

Initialize group $\mathcal{G}_1, \cdots \mathcal{G}_K$
Set step $s = 1$ and initialized α_k^1, $1 \le k \le K$, to K random starting values in $\mathcal{S}_\alpha$:
while *Not converged* **do**
  for $1 \le k \le K$ **do**
    Estimate

$$\widehat{\boldsymbol{\beta}}^{\pi(\boldsymbol{\alpha}_k^s)} = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \boldsymbol{G}\left(\pi\left(\boldsymbol{\alpha}_k^s\right), \boldsymbol{\beta}\right)^\top \boldsymbol{G}\left(\pi\left(\boldsymbol{\alpha}_k^s\right), \boldsymbol{\beta}\right) + \frac{1}{N^2} \sum_{1 \le i < j \le N} \mathcal{P}\left((JM)^{-1/2} \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|_2, \lambda_{NT}\right).$$

  Cluster and align N trajectories to $\mathcal{G}_1, \cdots \mathcal{G}_K$ using $\left\{\widehat{\boldsymbol{\beta}}^{\pi(\boldsymbol{\alpha}_1^s)}, \cdots, \widehat{\boldsymbol{\beta}}^{\pi(\boldsymbol{\alpha}_K^s)}\right\}$.
  for $1 \le k \le K$ **do**
    Evaluate on $\mathcal{P}_k$ for
    $\widehat{V}_{\mathcal{P}_k}(\pi(\boldsymbol{\alpha})) = \int \widehat{V}(\pi(\boldsymbol{\alpha}), \boldsymbol{x})\, d\mathcal{P}_k(\boldsymbol{x}) = \int \boldsymbol{u}(\pi(\boldsymbol{\alpha}), \boldsymbol{x})^\top \widehat{\boldsymbol{\beta}}^{\pi(\boldsymbol{\alpha}_k^s)} d\mathcal{P}_k(\boldsymbol{x})$.
    Update $\boldsymbol{\alpha}_k^{s+1} \leftarrow \arg\max_{\boldsymbol{\alpha}} \widehat{V}_{\mathcal{P}_k}(\pi(\boldsymbol{\alpha}))$
  $s \leftarrow s + 1$.

3.1 Policy evaluation with heterogeneity

Let $Q^{\pi}\left(\boldsymbol{x}, a; \boldsymbol{\beta}_i\right)$ denote a model for $Q_i^{\pi}(\boldsymbol{x}, a)$ indexed by $\boldsymbol{\beta}_i = \left[\boldsymbol{\beta}_{1,i}^{\top}, \cdots, \boldsymbol{\beta}_{M,i}^{\top}\right]^{\top}$, $1 \leq i \leq N$, where $\boldsymbol{\beta}_{m,i} \in \mathbb{R}^J$ is the coefficient for the m-th action in the action set $\mathcal{A} = \{1, 2, \cdots, M\}$. We assume that the map $\boldsymbol{\beta}_i \mapsto Q_i^{\pi}\left(\boldsymbol{x}, a; \boldsymbol{\beta}_i\right)$ is differentiable everywhere for each fixed $\boldsymbol{x}$, a and π. We stack all individual coefficients in a long vector $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_1^{\top} \cdots \boldsymbol{\beta}_N^{\top}\right]^{\top}$ and define a trajectory indicator $\boldsymbol{\Lambda}_i = [\mathbf{0} \cdots \mathbf{0}\ \boldsymbol{I}\ \mathbf{0} \cdots \mathbf{0}]$ such that $\boldsymbol{\beta}_i = \boldsymbol{\Lambda}_i \boldsymbol{\beta}$. Let $\nabla Q^{\pi}(\boldsymbol{x}, a; \boldsymbol{\Lambda}_i \boldsymbol{\beta})$ denote the gradient of $Q^{\pi}(\boldsymbol{x}, a; \boldsymbol{\Lambda}_i \boldsymbol{\beta})$ with respect to the unknown coefficients. With observed trajectories $\{\boldsymbol{x}_{i,t}, a_{i,t}, r_{i,t}\}$, $1 \leq t \leq T$, $1 \leq i \leq N$, we define a sample version quantity:

$$\boldsymbol{G}(\pi, \boldsymbol{\beta}) = \frac{1}{NTJ} \sum_{i=1}^{N} \sum_{t=1}^{T} \Bigg(r_{i,t} + \gamma \sum_{a' \in \mathcal{A}} Q^{\pi}(\boldsymbol{x}_{i,t}, a_{i,t}; \boldsymbol{\Lambda}_i \boldsymbol{\beta}) \pi(a' | \boldsymbol{x}_{i,t+1}) - Q^{\pi}(\boldsymbol{x}_{i,t}, a_{i,t}; \boldsymbol{\Lambda}_i \boldsymbol{\beta}) \Bigg) \cdot \nabla Q^{\pi}(\boldsymbol{x}_{i,t}, a_{i,t}; \boldsymbol{\Lambda}_i \boldsymbol{\beta}). \tag{9}$$

By Lemma 2.4, under certain mild conditions, there exist some $\mathring{\boldsymbol{\beta}}$ that satisfy $\mathbb{E}\left[\boldsymbol{G}^{\pi}\left(\mathring{\boldsymbol{\beta}}\right)\right] = \mathbf{0}$. Thus, an estimator of $\mathring{\boldsymbol{\beta}}$ can be obtained by minimizing $\boldsymbol{G}(\pi, \boldsymbol{\beta})^{\top} \boldsymbol{G}(\pi, \boldsymbol{\beta})$. At the same time, the true model coefficients should satisfy $\mathring{\boldsymbol{\beta}}_i = \mathring{\boldsymbol{\beta}}_j$ if $i, j \in \mathcal{G}_k$, $1 \leq k \leq K$ under the K-hetero MDP assumption. When the group $\mathcal{G}_k$ is unknown, it is crucial and beneficial to encourage grouping individuals with the same model coefficients together to achieve efficient estimation. Given a penalty function $p : \mathbb{R} \mapsto \mathbb{R}$, the policy evaluation step of ACPI estimates $\mathring{\boldsymbol{\beta}}$ as:

$$\widehat{\boldsymbol{\beta}}^{\pi} = \arg\max_{\boldsymbol{\beta}} \left[\boldsymbol{G}(\pi, \boldsymbol{\beta})^{\top} \boldsymbol{G}(\pi, \boldsymbol{\beta}) + \frac{1}{N^2} \sum_{1 \leq i < j \leq N} p\left((JM)^{-1/2} \left\| \boldsymbol{\beta}_i - \boldsymbol{\beta}_j \right\|_2, \lambda_{NT} \right) \right] \tag{10}$$

The fusion type penalty shrinks some of the pairs $\widehat{\boldsymbol{\beta}}_i^{\pi} - \widehat{\boldsymbol{\beta}}_j^{\pi}$ to zero, based on which we can partition the sample into subgroups simultaneously when we estimate the coefficients. The choice of tuning parameter λ_{NT} is given in Theorem 4.12.

Semiparametric Q approximation The ACPI requires a class of models for the Q value function indexed by parameter $\boldsymbol{\beta}$. We use a basis function approximation. Let $\boldsymbol{\phi}(\cdot) = \left[\phi_1(\cdot), \cdots, \phi_J(\cdot)\right]^{\top}$ be a vector of pre-specified basis functions such as Gaussian basis function or splines. We allow J to grow with the sample size to reduce the bias of the resulting estimates. The heterogeneous Q function writes

$$Q^{\pi}(\boldsymbol{x}_{i,t}, a_{i,t}; \boldsymbol{\Lambda}_i \boldsymbol{\beta}) \approx \boldsymbol{z}_{i,t}^{\top} \boldsymbol{\Lambda}_i \boldsymbol{\beta}. \tag{11}$$

where

$$\boldsymbol{z}_{i,t} = \boldsymbol{z}(\boldsymbol{x}_{i,t}, a_{i,t}) = \left[\boldsymbol{\phi}(\boldsymbol{x}_{i,t})^\top \mathbb{1}(a_{i,t}=1), \cdots, \boldsymbol{\phi}(\boldsymbol{x}_{i,t})^\top \mathbb{1}(a_{i,t}=M)\right]^\top. \tag{12}$$

Plugging (11) in (10), we have

$$\boldsymbol{G}(\pi, \boldsymbol{\beta}) = \frac{1}{NTJ} \sum_{i=1}^{N} \sum_{t=0}^{T} \left(\boldsymbol{\Lambda}_i^\top \boldsymbol{z}_{i,t} r_{i,t} - \boldsymbol{\Lambda}_i^\top \boldsymbol{z}_{i,t} (\boldsymbol{z}_{i,t} - \gamma \boldsymbol{U}_{\pi,i,t+1} \boldsymbol{\Lambda}_i)^\top \boldsymbol{\Lambda}_i \boldsymbol{\beta}\right), \tag{13}$$

where

$$\boldsymbol{u}_{i,t}^{\pi} = \boldsymbol{u}(\pi, \boldsymbol{x}_{i,t}) = \left[\boldsymbol{\phi}(\boldsymbol{x}_{i,t})^\top \pi(1|\boldsymbol{x}_{i,t}), \cdots, \boldsymbol{\phi}(\boldsymbol{x}_{i,t})^\top \pi(M|\boldsymbol{x}_{i,t})\right]^\top. \tag{14}$$

Concave penalty function The ACPI also requires a penalty function $p(\cdot)$. We consider concave penalties satisfying Assumption 4.10 that can produce unbiased estimates. Examples includes SCAD (Fan and Li, 2001) and MCP (Zhang et al., 2010), which are defined respectively as

$$p_\eta^{MCP}(t, \lambda) = \lambda \int_0^t (1 - x/(\eta\lambda))_+ \, dx, \quad \eta > 1, \tag{15}$$

$$p_\eta^{SCAD}(t, \lambda) = \lambda \int_0^t \min\left\{1, (\gamma - x/\lambda)_+/(\gamma - 1)\right\} dx, \quad \eta > 2, \tag{16}$$

where η is a parameter that controls the concavity of the penalty function. In particular, both penalties converge to the L_1 penalty as $\eta \to \infty$.

Heterogeneous policy evaluation Under semi-parametric Q function approximation, we obtain $\widehat{\boldsymbol{\beta}}$ by solving (10) with instantiated $\boldsymbol{G}(\pi, \boldsymbol{\beta})$ given in (13) and a concave penalty $p(\cdot)$ satisfying Assumption 4.10. The individual-wise value functions can be estimated by

$$\widehat{Q}_i^{\pi}(\boldsymbol{x}, a) = \boldsymbol{z}(\boldsymbol{x}, a)^\top \widehat{\boldsymbol{\beta}}_i^{\pi}, \quad \text{and} \quad \widehat{V}_i^{\pi}(\boldsymbol{x}, a) = \boldsymbol{u}(\pi, \boldsymbol{x})^\top \widehat{\boldsymbol{\beta}}_i^{\pi}, \quad \text{for} \quad 1 \le i \le N, \tag{17}$$

where $\boldsymbol{z}(\boldsymbol{x}, a)$ and $\boldsymbol{u}(\pi, \boldsymbol{x})$ are defined in (12) and (14), respectively.

The dimension of the unknown parameters $\boldsymbol{\beta}$ is NJM, which will diverge as sample size N increase. Without the K hetero MDP assumption and the penalty term, the ℓ_2 convergence of $\widehat{\boldsymbol{\beta}}_i^{\pi}$ the can be shown to be $1/\sqrt{T}$ under the TH Markov and CMI assumption, employing arguments similar to those in Luckett et al. (2019); Shi et al. (2020). The information accumulated along N does not help. In contrast, the ACPI obtain $1/\sqrt{TNP_{\min}}$ convergence rate where $P_{\min} = \min\{P_1, \cdots, P_K\}$.

Different from Luckett et al. (2019), our method estimate Q function and derive the

corresponding value estimators. Our method does not require correct specification of the behavior policy. Nor do we need to estimate it from the observed dataset. This can be viewed as an implicit importance weighting.

3.2 Cluster-wise policy improvement

We consider finite action space $\mathcal{A}$ and a parametric class of policies commonly used in the literature. Assuming M possible actions $\mathcal{A} = \{1, \cdots, M\}$, we define a parametric class of policies Π as

$$\pi(\boldsymbol{\alpha}) \equiv \pi(j, \boldsymbol{x}; \boldsymbol{\alpha}) = \begin{cases} \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\alpha}_j)}{1+\sum_{j=1}^{M-1} \exp(\boldsymbol{x}^\top \boldsymbol{\alpha}_j)}, & \text{for } j = 1, \cdots, M-1 \\ \frac{1}{1+\sum_{j=1}^{M-1} \exp(\boldsymbol{x}^\top \boldsymbol{\alpha}_j)}, & \text{for } j = M, \end{cases} \tag{18}$$

where $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_1^\top, \cdots, \boldsymbol{\alpha}_M^\top\right]^\top$ and $\boldsymbol{\alpha}_j$ is a vector of parameters for the j-th action.

Given estimated $\widehat{\boldsymbol{\beta}}_i^\pi$ and K from Section 3.1, we can estimate the $N \times K$ group membership matrix $\widehat{\boldsymbol{W}} = [\widehat{\boldsymbol{w}}_{1\cdot} \cdots \widehat{\boldsymbol{w}}_{N\cdot}]^\top$ by using any chosen clustering algorithm. The k-th sub-population coefficient is estimated as $\widehat{\boldsymbol{\theta}}_k^\pi = \left(\widehat{\boldsymbol{w}}_{\cdot k}^\top \widehat{\boldsymbol{w}}_{\cdot k}\right)^{-1} \text{MAT}\left(\widehat{\boldsymbol{\beta}}^\pi\right) \widehat{\boldsymbol{w}}_{\cdot k}$. Note that we only need to use clustering algorithms when N and T are small. Asymptotically, by the oracle property in Theorem 4.12, we have $\Pr\left(\widehat{\boldsymbol{\beta}}_i^\pi = \widehat{\boldsymbol{\beta}}_j^\pi\right) \to 1$ if i and j belong to the same sub-population. The sub-population coefficients $\boldsymbol{\theta}_k^\pi$, $1 \le k \le K$, can be chosen as the distinct values of $\widehat{\boldsymbol{\beta}}_i^\pi$, $1 \le i \le N$. The value functions for the k sub-population can be estimated by

$$\widehat{Q}^{\pi,(k)}(\boldsymbol{x}, a) = \boldsymbol{z}(\boldsymbol{x}, a)^\top \widehat{\boldsymbol{\theta}}_k^\pi, \quad \text{and} \quad \widehat{V}^{\pi,(k)}(\boldsymbol{x}) = \boldsymbol{u}(\pi, \boldsymbol{x})^\top \widehat{\boldsymbol{\theta}}_k^\pi, \quad \text{for} \quad 1 \le k \le K, \tag{19}$$

where $\boldsymbol{z}(\boldsymbol{x}, a)$ and $\boldsymbol{u}(\pi, \boldsymbol{x})$ are defined in (12) and (14), respectively. Let $\mathcal{P}_x^{(k)}$ be a reference distribution on the covariate space $\mathcal{X}_k$ of the k-th group. The integrated value function can be estimated by

$$\widehat{V}_{\mathcal{P}}^{(k)}(\pi(\alpha)) = \int \sum_{j \in \mathcal{A}} \widehat{Q}^{\pi,(k)}(\boldsymbol{x}, j)\, \pi(j, \boldsymbol{x}; \boldsymbol{\alpha})\, \mathcal{P}_x^{(k)}(d\boldsymbol{x}), \tag{20}$$

where $\widehat{Q}^{(k)}(\pi, \boldsymbol{x}, j)$ and $\pi(j, \boldsymbol{x}; \boldsymbol{\alpha})$ are given in (19) and (18), respectively. For each sub-population k, group-wise policies $\pi(\boldsymbol{\alpha}_k)$ are improved to maximize integrated value (20), according to line 7 - 9 in Algorithm 1.

4 Theory

In this section, we lay out the theoretical framework for individual-wise modeling inference and population-wise clustering analysis in a double-divergence structure, which allows either sample size N or decision horizon T goes to infinity. We establish asymptotic properties of the offline estimations of the coefficients, value functions and optimal policy. Let $\boldsymbol{W} = [\boldsymbol{w}_{1\cdot} \cdots \boldsymbol{w}_{N\cdot}]^\top \in \mathbb{R}^{N\times K}$ be the group membership matrix where $w_{ik} = 1$ for $i \in \mathcal{G}_k$ and $w_{ik} = 0$ otherwise. Now we consider theoretical results for the *oracle* estimator when the true $\boldsymbol{W}$ is *known a priori*. Section 4.1 derives the oracle properties such as ℓ_2 and ℓ_∞ convergence rates when the true sub-population information $\{\mathcal{G}_k, 1 \le k \le K\}$ is known. Section 4.2 established the ℓ_2 convergence rate and the asymptotic normality of the parameters estimated from (10) when $\{\mathcal{G}_k, 1 \le k \le K\}$ is unknown. The proofs of all theorems are provided in the supplementary materials.

4.1 Properties of the oracle coefficients

We denote $\boldsymbol{\theta}^\pi = \left[\boldsymbol{\theta}_1^{\pi\top} \cdots \boldsymbol{\theta}_K^{\pi\top}\right]^\top$ as the group coefficient matrix where $\boldsymbol{\theta}_k^\pi$ is the coefficient for $\mathcal{G}_k$. When the $\boldsymbol{W}$ is known, the oracle estimator refers to an estimator minimizing the objective function in (10) with respect to $\boldsymbol{\theta}^\pi$ without penalty, that is

$$\widetilde{\boldsymbol{\theta}}^\pi = \arg\max_{\boldsymbol{\theta}} \widetilde{\boldsymbol{G}}(\pi, \boldsymbol{\theta})^\top \widetilde{\boldsymbol{G}}(\pi, \boldsymbol{\theta}), \tag{21}$$

where

$$\widetilde{\boldsymbol{G}}(\pi, \boldsymbol{\theta}) = \boldsymbol{G}\Big(\pi, (\boldsymbol{W} \otimes \boldsymbol{I}_{JM})\boldsymbol{\theta}\Big) = \frac{1}{NTJ}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big(\widetilde{\boldsymbol{\Lambda}}_i^\top \boldsymbol{z}_{i,t} r_{i,t} - \widetilde{\boldsymbol{\Lambda}}_i^\top \Big(\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top - \gamma \boldsymbol{z}_{i,t}\boldsymbol{u}_{\pi,i,t+1}^\top\Big)\widetilde{\boldsymbol{\Lambda}}_i\boldsymbol{\theta}\Big), \tag{22}$$

and $\widetilde{\boldsymbol{\Lambda}}_i = \boldsymbol{\Lambda}_i(\boldsymbol{W} \otimes \boldsymbol{I}_{JM})$ and $\widetilde{\boldsymbol{\beta}}^\pi = (\boldsymbol{W} \otimes \boldsymbol{I}_{JM})\widetilde{\boldsymbol{\theta}}^\pi$.

We first layout the assumptions that are necessary for the convergence and asymptotic normality of the heterogeneous policy evaluation.

Definition 4.1 (κ-Smooth functions)**.** *Let $f(\cdot)$ be an arbitrary function on $\mathcal{X} \in \mathbb{R}^p$. For a p-tuple $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)$ fo non-negative integers, let D^α denote the differential operator:*

$$D^\alpha f(\boldsymbol{x}) = \frac{\partial^{\|\boldsymbol{\alpha}\|_1} f(\boldsymbol{x})}{\partial x_1^{\alpha_1} \cdots \partial x_p^{\alpha_p}},$$

where $\boldsymbol{x} = (x_1, \cdots, x_p)^\top$. *The class of* κ*-smooth functions is defined as*

$$\mathcal{H}(\kappa, c) = \left\{ f : \sup_{\|\boldsymbol{\alpha}\|_1 \le \lfloor \kappa \rfloor} \sup_{\boldsymbol{x} \in \mathcal{X}} |D^{\boldsymbol{\alpha}} f(\boldsymbol{x})| \le c; \text{ and } \sup_{\|\boldsymbol{\alpha}\|_1 \le \lfloor \kappa \rfloor} \sup_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \boldsymbol{x}_1 \ne \boldsymbol{x}_2} \frac{|D^{\boldsymbol{\alpha}} f(\boldsymbol{x}_1) - D^{\boldsymbol{\alpha}} f(\boldsymbol{x}_2)|}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^{\kappa - \lfloor \kappa \rfloor}} \right\}.$$

Assumption 4.2. *There exists some* κ, $c > 0$ *such that* $r(\boldsymbol{x}, a)$, $\Pr(\boldsymbol{x}'|\boldsymbol{x}, a)$ *belong to the class of* κ*-smooth function of* $\boldsymbol{x}$ *for any* $a \in \mathcal{A}$ *and* $\boldsymbol{x}' \in \mathcal{X}$.

Lemma 1 in Shi et al. (2020) shows that under Assumption 4.2, there exists some constant $c' > 0$ such that the Q function $Q(\pi, \boldsymbol{x}, a)$ belongs to the class of κ-smooth function of $\boldsymbol{x}$ for any policy π and any action $a \in \mathcal{A}$. This implies that the Q function has bounded derivatives up to order $\lfloor \kappa \rfloor$.

Assumption 4.3. *Let* $BSpline(J, r)$ *denote a tensor-product B-spline basis of dimension* J *and of degree* r *on* $[0,1]^p$ *and* $Wav(J, r)$ *denote a tensor-product Wavelet basis of regularity* r *and dimension* J *on* $[0,1]^p$. *The sieve* $\boldsymbol{\phi}_J$ *is either* $BSpline(J, r)$ *or* $Wav(J, r)$ *with* $r > \max(\kappa, 1)$.

Assumption 4.4. *The density function* μ *and* ν_0 *are uniformly bounded away from* 0 *and* ∞ *on* $\mathcal{X}$.

Assumption 4.5. *Under the setting that* T *is bounded, we assume that there exists some constant* $C_1 > 0$ *such that*

$$\lambda_{\min}\left(T^{-1} \sum_{t=1}^{T-1} \mathbb{E}\left[\boldsymbol{z}_{i,t} \boldsymbol{z}_{i,t}^\top - \gamma^2 \boldsymbol{u}_\pi(\boldsymbol{X}_{i,t}, A_{i,t}) \boldsymbol{u}_\pi(\boldsymbol{X}_{i,t}, A_{i,t})^\top \right] \right) \ge C_1,$$

where $\boldsymbol{u}_\pi(\boldsymbol{x}, a) = \mathbb{E}\left[\boldsymbol{u}(\pi, \boldsymbol{X}_{i,t+1}) \mid \boldsymbol{X}_{i,t} = \boldsymbol{x}, A_{i,t} = a \right]$ *and* $\lambda_{\min}(\cdot)$ *denotes the minimal eigenvalue of a matrix.*

Assumption 4.6. *Under the setting that* $T \to \infty$, *we assume that*

(i) There exists some constant $C_1 > 0$ *such that*

$$\lambda_{\min}\left(\int_{\boldsymbol{x} \in \mathcal{X}} \sum_{a \in \mathcal{A}} \left(\boldsymbol{z}(\boldsymbol{x}, a) \boldsymbol{z}(\boldsymbol{x}, a)^\top - \gamma^2 \boldsymbol{u}_\pi(\boldsymbol{x}, a) \boldsymbol{u}_\pi(\boldsymbol{x}, a)^\top \right) b(a|\boldsymbol{x}) \mu(\boldsymbol{x}) d\boldsymbol{x} \right) \ge C_1,$$

where $\boldsymbol{u}_\pi(\boldsymbol{x}, a) = \mathbb{E}\left[\boldsymbol{u}(\pi, \boldsymbol{X}_{i,t+1}) \mid \boldsymbol{X}_{i,t} = \boldsymbol{x}, A_{i,t} = a \right]$ *and* $\lambda_{\min}(\cdot)$ *denotes the minimal eigenvalue of a matrix.*

(ii) The Markov chain $\{\boldsymbol{X}_{i,t}\}_{t \ge 0}$ *is geometrically ergodic, that is, there exists some function* $f(\boldsymbol{x})$ *on* $\mathcal{X}$ *and some constant* $c \le 1$ *such that* $\int_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) \mu(\boldsymbol{x}) d\boldsymbol{x} < +\infty$ *and*

$$\left\| P_t(\cdot|\boldsymbol{x}) - \mu(\boldsymbol{x}) \right\|_{TV} \le f(\boldsymbol{x}) \rho^t, \qquad \forall t \ge 0,$$

where $\|\cdot\|_{TV}$ denotes the total variation norm.

When T is finite, Assumption 4.5 guarantees that $\widetilde{\Sigma}$ (23) is invertible with probability approaching one when $N \to \infty$. When $T \to \infty$, 4.6 (i) guarantees that the matrix $\mathbb{E}\left[\widetilde{\Sigma}\right]$ and (ii) enables us to derive matrix concentration inequalities for $\widetilde{\Sigma}$. Together, they implies that $\widetilde{\Sigma}$ is invertible with probability approaching one. Assumption 4.5 and 4.6 are mild and can be verified empirically by checking that certain data-dependent matrices are invertible. More discussion on similar assumptions can be found in Luckett et al. (2019); Shi et al. (2020).

Theorem 4.7 (Oracle estimator asymptotic normality)**.** *Suppose Assumption 4.2 – 4.6 hold. Let $N_{min} = \min_{1\le k\le K} N_k$ and $N_{max} = \max_{1\le k\le K} N_k$. If $K = o(N_{\min} T)$, $J \ll \sqrt{N_{\min} T}/\log(N_{\min} T)$, $J^{\kappa/p} \gg \sqrt{N_{\max} T}$. For any $\boldsymbol{\nu} \in \mathbb{R}^{JMK}$ satisfying $J^{\kappa/p} \gg \sqrt{N_{\max} T\left(1 + \|\boldsymbol{\nu}\|_2^{-2}\right)}$, we have as either $N_{\min} \to \infty$ or $T \to \infty$,*

$$\sqrt{NT}\,\widetilde{\sigma}_\theta(\pi, \boldsymbol{\nu})^{-1}\,\boldsymbol{\nu}^\top\left(\widetilde{\boldsymbol{\theta}}^\pi - \mathring{\boldsymbol{\theta}}^\pi\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\widetilde{\sigma}_\theta(\pi, \boldsymbol{\nu}) = \sqrt{\boldsymbol{\nu}^\top \widetilde{\Sigma}^{-1} \widetilde{\boldsymbol{\Omega}} (\widetilde{\Sigma}^\top)^{-1} \boldsymbol{\nu}}$,

$$\widetilde{\Sigma} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{\boldsymbol{\Lambda}}_i^\top \boldsymbol{z}_{i,t}\left(\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1}\right)^\top \widetilde{\boldsymbol{\Lambda}}_i, \tag{23}$$

$$\widetilde{\boldsymbol{\Omega}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{\boldsymbol{\Lambda}}_i^\top \boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top \widetilde{\boldsymbol{\Lambda}}_i \left(r_{i,t} - \left(\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1}\right)^\top \widetilde{\boldsymbol{\Lambda}}_i \widetilde{\boldsymbol{\theta}}^\pi\right)^2. \tag{24}$$

Remark 4. *Under the condition in Theorem 4.7, we can show that $\widetilde{\sigma}_\theta(\pi, \boldsymbol{\nu})$ converges almost surely to $\sigma_\theta(\pi, \boldsymbol{\nu}) = \sqrt{\boldsymbol{\nu}^\top \Sigma^{-1} \boldsymbol{\Omega} (\Sigma^\top)^{-1} \boldsymbol{\nu}}$, where $\boldsymbol{\Omega} = \mathbb{E}\left[\widetilde{\boldsymbol{\Omega}}\right]$ and $\Sigma = \mathbb{E}\left[\widetilde{\Sigma}\right]$.*

Applying Theorem 4.7 to value functions estimators defined in (19), we establish the asymptotic distribution for state and action value functions V^π and Q^π.

Corollary 4.8 (Value asymptotic normality)**.** *Suppose Assumption 4.2 – 4.6 hold. If $K = o(N_{\min} T)$, $J \ll \sqrt{N_{\min} T}/\log(N_{\min} T)$, and $J^{\kappa/p} \gg \sqrt{N_{\max} T\left(1 + \|\boldsymbol{U}(\pi, \boldsymbol{x})\|_2^{-2}\right)}$, we have as either $N_{\min} \to \infty$ or $T \to \infty$,*

$$\sqrt{NT}\,\widetilde{\sigma}_v(\pi, \boldsymbol{u})^{-1}\left(\widetilde{V}^\pi - \mathring{V}^\pi\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$
$$\sqrt{NT}\,\widetilde{\sigma}_q(\pi, \boldsymbol{z})^{-1}\left(\widetilde{Q}^\pi - \mathring{Q}^\pi\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where

$$\widetilde{\sigma}_v(\pi,\boldsymbol{u}) = \sqrt{\boldsymbol{u}(\pi,\boldsymbol{x})^\top \widetilde{\Sigma}^{-1}\widetilde{\boldsymbol{\Omega}}(\widetilde{\Sigma}^\top)^{-1}\boldsymbol{u}(\pi,\boldsymbol{x})}, \quad \widetilde{\sigma}_q(\pi,\boldsymbol{z}) = \sqrt{\boldsymbol{z}(\boldsymbol{x},a)^\top \widetilde{\Sigma}^{-1}\widetilde{\boldsymbol{\Omega}}(\widetilde{\Sigma}^\top)^{-1}\boldsymbol{z}(\pi,\boldsymbol{x})},$$

$\widetilde{\Sigma}$ *and* $\widetilde{\boldsymbol{\Omega}}$ *are given in* (23) *and* (24).

Remark 5. *Theorem 4.7 and Corollary 4.8 imply that the group-wise estimator converges at a rate of* $(NTP_k)^{-1/2}$, $k \in [K]$. *The estimators in Luckett et al. (2019); Jiang and Li (2016); Thomas et al. (2015) typically converge at a rate of* $(NP_k)^{-1/2}$ *and are not suitable for settings when one sub-population has only a few trajectories. Our estimation also aggregate information along horizon* T. *The theoretical property along* T *is obtained by treating finite and infinite* T *separately, using the matrix concentration and martingale center limit theorem.*

To finally establish the large sample theory for the ACPI, we need a stronger *uniform consistency* regarding the unknown parameters when either $N \to \infty$ or $T \to \infty$.

Theorem 4.9 (Oracle estimator uniform convergence.)**.** *Suppose Assumption 4.2 – 4.6 hold. If* $K = o(N_{\min}T)$, $J \ll \sqrt{N_{\min}T}/\log(N_{\min}T)$, $J^{-\kappa/p} \ll 1/\sqrt{N_{\max}T}$, *we have with probability at least* $1 - 2JMK(N_{\min}T)^{-2} - \mathcal{O}\left((N_{\min}T)^{-2}\right)$ *that*

$$\left\|\widetilde{\boldsymbol{\theta}}^\pi - \mathring{\boldsymbol{\theta}}^\pi\right\|_\infty \le \phi_{NT},$$

where

$$\phi_{NT} = \left\|\widetilde{\boldsymbol{\theta}}^\pi - \mathring{\boldsymbol{\theta}}^\pi\right\|_\infty \le 6cC^{-1}\frac{N_{\max}}{N_{\min}}\sqrt{2J\frac{\log(N_{\max}T)}{N_{\max}T}}. \tag{25}$$

Remark 6. *By Definition 2.3,* $N_{\max}/N_{\min} = P_{\max}/P_{\min}$ *are bounded way from infinity. Bound* (25) *can be simplified to* $\mathcal{O}_p\left(\sqrt{J\frac{\log(NT)}{NT}}\right)$.

4.2 Properties of the feasible estimator

We now study the theoretical properties of the feasible estimator when the true group membership $\boldsymbol{W}$ is not known. We introduce the following assumptions on the penalty function and the minimum signal difference between groups.

Assumption 4.10. *The penalty function* $p(x,\lambda)$ *is a symmetric function of* x, *and it is non-decreasing and concave in* x *for* $x \in [0,+\infty)$. *Let* $\rho(x) = \lambda^{-1}p(x,\lambda)$, *there exists a constant* $0 < c < \infty$ *such that* $\rho(0) = 0$ *and* $\rho(x)$ *is a constant for all* $x \ge c\lambda$. *Its derivative* $\rho'(x)$ *exists and is continuous except for a finite number of* x *and* $\rho'(0+) = 1$.

Assumption 4.11. *For $K > 2$, define the minimal difference of the common values between any pair of groups as*

$$d_{NT} = (JM)^{-1/2} \min_{k \neq l} \left\| \mathring{\boldsymbol{\theta}}_k - \mathring{\boldsymbol{\theta}}_l \right\|_2 .$$

We assume that $d_{NT} \gg \phi_{NT}$ where ϕ_{NT} is given in (25).

Assumption 4.10 is commonly assumed in high-dimensional settings. It is satisfied by the concave penalties such as MCP (15) and SCAD (16). Assumption 4.11 is the separability condition on the minimum signal difference between groups that is needed to recover the true groups.

Theorem 4.12 (Feasible estimator)**.** *Suppose the assumptions in Theorem 4.9 hold and $K \geq 2$. If $d_{NT} \geq C\lambda$ and $\lambda \gg \max\left(N_{\min}^{-1}\sqrt{J}, \phi_{NT}\right)$, where C, d_{NT}, and ϕ_{NT} are defined in Assumption 4.10, 4.11 and Equation (25). Then there exists a local minimizer $\widehat{\boldsymbol{\beta}}^{\pi}$ of the objective function $\mathcal{L}_{NT}$ given in (10) satisfying*

$$\Pr\left(\widehat{\boldsymbol{\beta}}^{\pi} = \widetilde{\boldsymbol{\beta}}^{\pi}\right) \to 1.$$

Remark 7. *Recall that we define $\widehat{\boldsymbol{W}}$ as an estimator of $\boldsymbol{W}$ that is obtained by applying any clustering method on the column vector of $\textsc{mat}(\boldsymbol{\beta}^{\pi})$. A direct conclusion from Theorem 4.12 is that $\Pr\left(\widehat{\boldsymbol{W}} = \boldsymbol{W}\right) \to 1$ since the oracle estimator $\widetilde{\boldsymbol{\beta}}_i^{\pi} = \widetilde{\boldsymbol{\beta}}_j^{\pi}$ for any $i, j \in \mathcal{G}_k$.*

The oracle property in Theorem 4.12 together with Theorem 4.7 directly leads to the asymptotic distribution of $\widehat{\beta}^{\pi}$, which is presented in the following corollary.

Corollary 4.13. *Under the conditions in Theorem 4.7 and 4.12, for any $\boldsymbol{\nu} \in \mathbb{R}^{JMN}$ satisfying $J^{\kappa/p} \gg \sqrt{N_{\max} T \left(1 + \|\boldsymbol{\nu}\|_2^{-2}\right)}$, we have as either $N_{\min} \to \infty$ or $T \to \infty$,*

$$\sqrt{NT}\, \widehat{\sigma}_{\beta_i}^{-1} \boldsymbol{\nu}^{\top} \left(\widehat{\boldsymbol{\beta}}_i^{\pi} - \mathring{\boldsymbol{\beta}}_i^{\pi}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where

$$\begin{aligned}
\widehat{\sigma}_{\beta_i} &= \boldsymbol{\nu}^{\top} (\widehat{\boldsymbol{w}}_{i\cdot}^{\top} \otimes \boldsymbol{I}_{JM}) \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\Omega}} (\widehat{\boldsymbol{\Sigma}}^{\top})^{-1} (\widehat{\boldsymbol{w}}_{i\cdot} \otimes \boldsymbol{I}_{JM}) \boldsymbol{\nu} \\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\boldsymbol{\Lambda}}_i^{\top} \boldsymbol{z}_{i,t} \left(\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1}\right)^{\top} \widehat{\boldsymbol{\Lambda}}_i, \\
\widehat{\boldsymbol{\Omega}} &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\boldsymbol{\Lambda}}_i^{\top} \boldsymbol{z}_{i,t} \boldsymbol{z}_{i,t}^{\top} \widehat{\boldsymbol{\Lambda}}_i \left(r_{i,t} - \left(\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1}\right)^{\top} \widehat{\boldsymbol{\Lambda}}_i \widehat{\boldsymbol{\theta}}^{\pi}\right)^2 .
\end{aligned}$$

Asymptotic normality of the value functions similar to Corollary 4.8 can be established by applying Corollary 4.13 and thus is omit here.

4.3 Optimal policy

In this section, we establishes the convergence result of the estimated optimal policy for each homogeneous sub-population assuming that the parametric class Π satisfy properties as follows.

Assumption 4.14. *The map $\boldsymbol{\alpha} \to V_{\mathcal{P}}(\pi(\boldsymbol{\alpha}))$ has a unique and well separated maximum $\boldsymbol{\alpha}^*$ in the interior of the support of $\boldsymbol{\alpha}$, where $V_{\mathcal{P}}(\pi(\boldsymbol{\alpha}))$ is defined in (5).*

Assumption 4.15. *We have as $\delta \downarrow 0$,*

$$\sup_{\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_2 \le \delta} \mathbb{E}\|\pi(a, \boldsymbol{x}; \boldsymbol{\alpha}_1) - \pi(a, \boldsymbol{x}; \boldsymbol{\alpha}_1)\| \to 0.$$

Assumption 4.14 requires that the true optimal decision in each state is unique (see also the Assumption A.8 of Ertefaie and Strawderman (2018) and Assumption 6 of Luckett et al. (2019)) and is a standard assumption in M-estimation. Assumption 4.15 requires smoothness on the class of the polices. The following lemma shows that the parametric class of policies Π defined in (18) satisfies Assumption 4.14 and 4.15.

Lemma 4.16. *The parametric class of policies Π defined in (18) satisfies Assumption 4.14 and 4.15.*

Theorem 4.17 establishes that the estimated optimal policy for each homogeneous sub-population converges in probability to the true optimal policy over π and that the estimated value of the estimated optimal policy converges to the true value of the estimated optimal policy.

Theorem 4.17. *Suppose the conditions in Theorem 4.7 and 4.12 hold. Let $\widehat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathcal{S}_\alpha} \widehat{V}_{\mathcal{P}}(\pi(\alpha))$ and $\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha} \in \mathcal{S}_\alpha} V_{\mathcal{P}}(\pi(\alpha))$, we have as either $N_{\min} \to \infty$ or $T \to \infty$,*

(i) $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 \xrightarrow{\mathcal{P}} 0.$

(ii) $|V_{\mathcal{P}}(\pi(\widehat{\boldsymbol{\alpha}})) - V_{\mathcal{P}}(\pi(\boldsymbol{\alpha}^*))| \xrightarrow{\mathcal{P}} 0.$

5 Computation

The optimization problem of (10) is challenging because of the coupling of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ in the penalty term. To achieve computational scalability, we propose an ADMM-based algorithm (Boyd et al., 2011), which decomposes the original optimization into several smaller pieces that can be solved easily. Specifically, let λ be the value of the tuning

parameter selected based on a data-driven procedure such as the BIC. For brevity, we suppress the superscript π in this section.

We decouple β_i and β_j by reparameterizing and recasting (10) as the following constrained optimization problem:

$$\begin{aligned} \max_{\beta,\delta} \quad & \mathcal{L}(\beta,\delta) = \boldsymbol{G}(\pi,\beta)^\top \boldsymbol{G}(\pi,\beta) + \frac{1}{N^2}\sum_{1\le i<j\le N} \mathcal{P}\left((JM)^{-1/2}\left\|\delta_{ij}\right\|_2, \lambda\right) \\ s.t. \quad & \delta_{ij} = \beta_i - \beta_j, \quad \forall\{i,j\} : 1 \le i \le j \le N, \end{aligned}$$

where $\delta = \left[\delta_{ij}^\top, i<j\right]^\top$. Applying the augmented Lagrangian method (ALM), the solution of the constrained optimization problem can be obtained by minimizing

$$\mathcal{L}(\beta,\delta) + \sum_{i<j}\langle \nu_{ij}, \beta_i - \beta_j - \delta_{ij}\rangle + \frac{\rho}{2}\left\|\beta_i - \beta_j - \delta_{ij}\right\|_2^2$$

where $\nu = \left[\nu_{ij}^\top, i<j\right]^\top$ is the lagrangian multipliers and $\rho > 0$ is the penalty coefficient. We can hence iteratively optimize over three parameters (β,δ,ν) to obtain the solution through ADMM. Specifically, we start at initial values $\left(\beta^0,\delta^0,\nu^0\right)$ and update $\left(\beta^t,\delta^t,\nu^t\right)$ at the t-th iteration as follows.

Step 1. Given $\left(\delta^t,\nu^t\right)$, update β^{t+1} by solving β from $\nabla_\beta \mathcal{L}\left(\beta,\delta^t\right) = 0$.

Step 2. Given β^{t+1}, update δ^{t+1} using the analytical forms, for MCP, with

$$\delta_{ij}^{t+1} = \begin{cases} \frac{\mathcal{S}\left(\beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t, \lambda/\rho\right)}{1-1/(\gamma\rho)} & \text{if } \|\beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t\| \le \gamma\lambda \\ \beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t & \text{if } \|\beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t\| > \gamma\lambda \end{cases}$$

and, for SCAD, with

$$\delta_{ij}^{t+1} = \begin{cases} \mathcal{S}\left(\beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t, \lambda/\rho\right) & \text{if } \|\beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t\| \le \lambda+\lambda/\rho \\ \frac{\mathcal{S}\left(\beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t, \gamma\lambda/((\gamma-1)\rho)\right)}{1=1/((\gamma-1)/\rho)} & \text{if } \lambda+\lambda/\rho < \|\beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t\| \le \gamma\lambda \\ \beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t & \text{if } \|\beta_i^{t+1}-\beta_j^{t+1}+\rho^{-1}\nu_{ij}^t)\| > \gamma\lambda \end{cases},$$

where $\mathcal{S}(x,c) = \text{sign}(x)(|x|-c)_+$ is the soft thresholding rule and $(x)_+ = x$ if $x > 0$ and 0 otherwise.

Step 3. Update $\nu_{ij}^{t+1} = \nu_{ij}^t + \rho(\beta_i^{t+1} - \beta_j^{t+1} - \delta_{ij}^{t+1})$

The iteration terminates when the norm of the primal residual is smaller than some pre-specified small tolerance ϵ, that is, when $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\| < \epsilon$.

Remark 8. *A practical note is that each $\boldsymbol{\nu}_{ij}$ is a vector, which will cause trouble in the first step when it has a closed-form solution. However, if all ν_{ij} are concatenated into a big vector, this closed-form solution can be written. In this case we can simply implement a "difference" matrix, e.g. $\Delta = \{(\boldsymbol{e}_i - \boldsymbol{e}_j)^T\}_{1\le i\le j\le N}$ where $\boldsymbol{e}_i$ is an N-dimensional vector whose i th coordinate is 1 and all others zero. And our objective function will be then in terms of matrix - vector product form instead of summation form. However, when the dimension is very large, such concatenation will become infeasible as the dimension of Δ will explode. In such case, we need to do iterative methods for continuously differentiable functions (i.e. gradient descent) to find an optimal $\boldsymbol{\beta}$ or an accurate approximation for such $\boldsymbol{\beta}$ if resource-constrained. There are of course, other possible ways to solve such problem, for example the general iterative shrinkage and thresholding proposed in Gong et al. (2013).*

6 Simulations

In this section, we compare the performance of the proposed heterogeneous policy evaluation and estimation with the mean value based RL on simulated data.

We generate the initial state variable $\boldsymbol{X}_{i,0}$ from standard normal distribution $\mathcal{N}(\mathbf{0}_2, \boldsymbol{I}_2)$. The available action set is $\mathcal{A} = \{0, 1\}$. The system evolves according to

$$\boldsymbol{X}_{i,t+1} = \begin{bmatrix} 0.75(2A_{i,t}-1) & 0 \\ 0 & 0.75(1-2A_{i,t}) \end{bmatrix} \boldsymbol{X}_{i,t} + \boldsymbol{\sigma}_{i,t},$$

where $\boldsymbol{\sigma}_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}, 0.1 \cdot \boldsymbol{I}_2)$. The data-generating behavior actions are i.i.d. Bernoulli random variables with expectation 0.5 and are independent of $\boldsymbol{X}_{i,t}$ for any $t \ge 0$. We consider $K = 2$ homogeneous sub-populations. The immediate reward $Y_{i,t}$ is defined by

$$Y_{i,t} = \boldsymbol{X}_{i,t}^\top \boldsymbol{b}_k - 0.25(2A_{i,t}-1) + \eta_{i,t}, \quad \text{for} \quad \forall i \in \mathcal{G}_k,\ k \in 1, 2.$$

where $\boldsymbol{b}_1 \ne \boldsymbol{b}_2 \in \mathbb{R}^2$ are real coefficient vectors and $\eta_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0, 0.05)$.

Section 6.1 and 6.3 report results for policy estimation and estimated optimal policies, respectively.

6.1 Coefficients of the value function of a given policy π

The target policy π to be evaluated is specified as

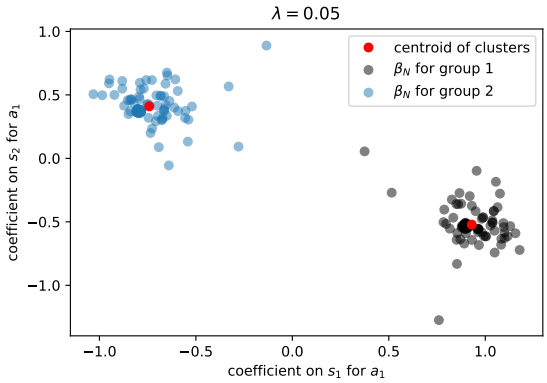
$$\pi(a|\boldsymbol{x}) = \begin{cases} 0, & x_1 > 0 \text{ and } x_2 > 0; \\ 1, & \text{otherwise,} \end{cases}$$

where x_i denotes the i-th element of a vector $\boldsymbol{x}$.

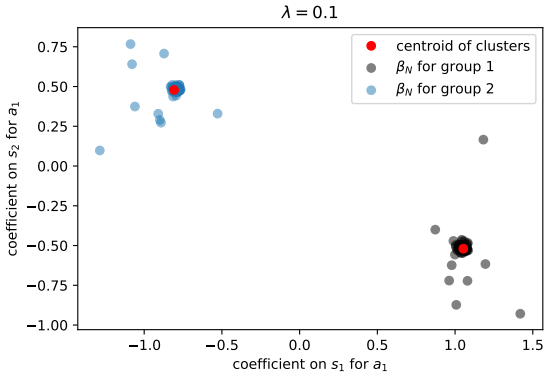
We set $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = [2,-1],[-2,1]$. For each setting, we try different values of tuning parameters λ. We simulate 100 trajectories of length $T = 20$ for each group with different values of $\boldsymbol{\theta}_k$, $k = 1,2$. The estimated values of $\boldsymbol{\beta}_i$ for all $N = 100$ trajectories and cluster centroids of $\boldsymbol{\theta}_k$, $k = 1,2$ under $T = 10$ and tuning parameters are plotted in Figure 2. The proposed algorithm is able to recover mostly correct coefficients with an appropriate choice of λ even when the group centroids are relatively close to each other. The tuning parameter λ controls the "focus of the group" Smaller λ encourages heterogeneity, that is different values for different trajectories, while large λ enforce homogeneity. When λ is too small, we loss the efficiency by grouping trajectories together. When λ is too big, it would reduce to the homogeneous policy evaluation with the same coefficients and hence lose the heterogeneity that we are seeking.

6.2 Confidence intervals of the value function of a given policy π

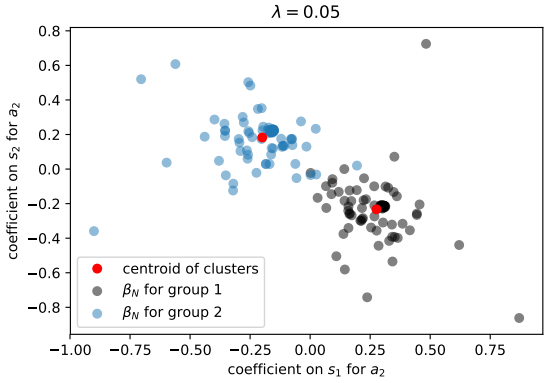
We compare the estimated values of $Q(\pi, \boldsymbol{x}, a)$ obtained by mean-valued RL and heterogeneous RL with the true values of a given policy π. The true value function $Q(\pi, \boldsymbol{x}, a)$ is computed by Monte Carlo approximation. Specifically, we simulate $N = 5,000$ independent trajectories (2,500 trajectories for each group) with initial states variable distributed according to $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_2)$. For each trajectory, we simulate T steps to obtain $\{Y_{i,t}\}$ for $1 \le i \le N$ and $1 \le t \le T$. For each sub-group, the true value is approximated by $\mathring{V}_{\mathcal{R}}(\pi) \approx N^{-1} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \gamma^t \boldsymbol{Y}_{i,t}$. We compare the true value with our estimated values with $n = 20, 50, 100$ and $T = 10, 30, 40$. The coverage probability is plotted in figure 3. We can see that heterogeneous estimation performs constantly well in terms of coverage probability. Due to the stationarity of this Markov chain, homogenous estimation starts to perform better when N and T are both large as the estimated V between two groups are close, but is still behind heterogenous estimation.



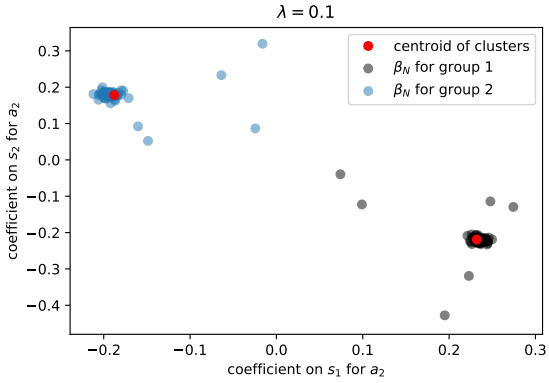
(a) Estimated $\boldsymbol{\beta}_N$ for action a_1, $\lambda = 0.05$



(b) Estimated $\boldsymbol{\beta}_N$ for action a_1, $\lambda = 0.1$



(c) Estimated $\boldsymbol{\beta}_N$ for action a_2, $\lambda = 0.05$



(d) Estimated $\boldsymbol{\beta}_N$ for action a_2, $\lambda = 0.1$

Figure 2: Coefficient centroids and coefficients fitted by our algorithm. Colors represent the sample's true membership. We can see that with an appropriate choice of λ, the majority of fitted coefficients would fall in a very small neighborhood of the centroids and can be hence correctly classified by simple clustering algorithms.

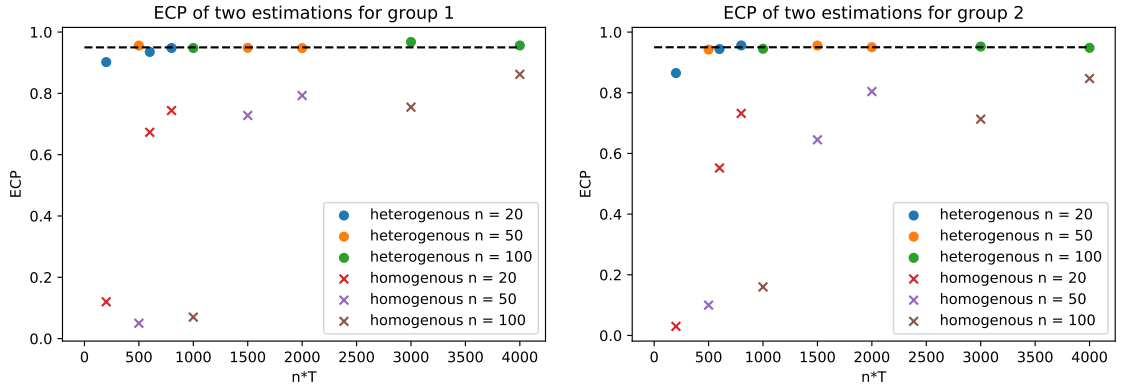
	A = 0	A = 1
$\pi(\boldsymbol{\alpha}_{G1})$	0.14	0.86
$\pi(\boldsymbol{\alpha}_{G2})$	0.75	0.25
$\pi(\boldsymbol{\alpha}_{G})$	0.58	0.42

Table 1: Outputs of three policies on the sample.

6.3 Parametric optimal policy

We train 100 steps of algorithm 1 to obtain two policies, which we denote $\pi(\boldsymbol{\alpha}_{G1})$ and $\pi(\boldsymbol{\alpha}_{G2})$ with auto-clustered heterogenous estimation. We then conduct the same to obtain one policy, which we denote $\pi(\boldsymbol{\alpha}_{G})$ with the standard mean-value estimation.

First, we would like to visualize the difference between optimal policies between two groups. We consider a test sample: $\boldsymbol{X}_{t=1} = \begin{bmatrix} 1.2506 \\ 0.77477 \end{bmatrix}, \boldsymbol{X}_{t=2} \begin{bmatrix} 1.0277 \\ -0.52615 \end{bmatrix}, A_{t=1} = 1, y_{t=1} = 1.6559$. Based on our value estimation, we cluster this sample as in group 1. We compare the results of all three policies in table 1. Based on our setting, we can see the optimal action should have a favor to $A_{t=1} = 1$ as it will yield a larger $y_{t=2}$. However, if we apply the mean-value estimated policy without considering the heterogeneity, we will wrongly weight relative evenly on both actions (or even leaning towards action 0), yielding a lower expected outcome.



(a) Empirical covering probability for group 1 of the 95% confidence interval

(b) Empirical covering probability for group 2 of the 95% confidence interval

Figure 3: Empirical covering probabilities

We then generate 5,000 samples distributed according to $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_2)$ (2,500 samples for each group) and estimate the value of all three policies $\mathring{V}_{\mathcal{R}}(\pi) \approx N^{-1} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \gamma^t \boldsymbol{Y}_{i,t}$ (break tie arbitrarily if there is any) at $T = 50$. We get $\mathring{V}(\pi(\boldsymbol{\alpha}_G)) = 0.0032395$, $\mathring{V}(\pi(\boldsymbol{\alpha}_{G_1})) = 0.13659$ and $\mathring{V}(\pi(\boldsymbol{\alpha}_{G_2})) = 0.13351$. This shows that policies produced from heterogenous

Q estimation are indeed better in terms of expected total sum of rewards.

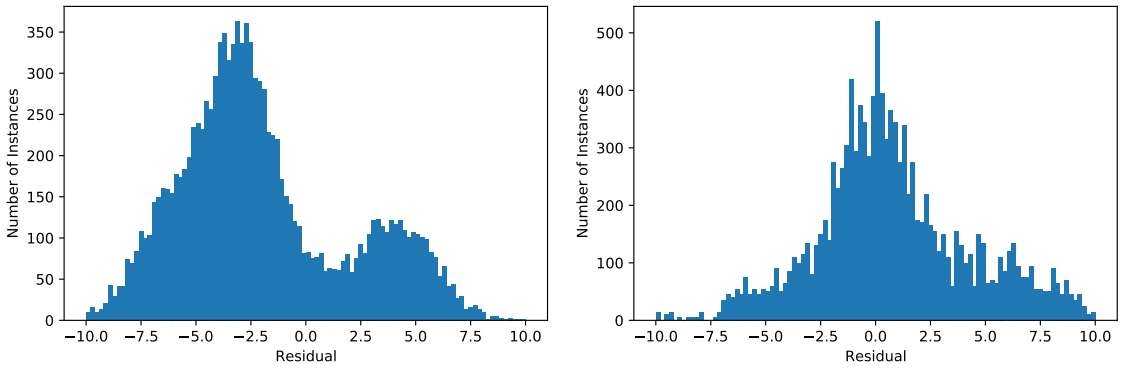
7 Real data application: MIMIC-III

In this section, we illustrate the advantages of the Auto-Clustered Q learning in the Medical Information Mart for Intensive Care version III (MIMIC-III) Database (Johnson et al., 2016), which is a freely available source of de-identified critical care data from 2001 – 2012 in six ICUs at a Boston teaching hospital.

We consider a cohort of sepsis patients, following the same data processing procedure as in Komorowski et al. (2018). Each patient in the cohort is characterized by a set of 47 variables, including demographics, Elixhauser premorbid status, vital signs, and laboratory values. Patients' data were coded as multidimensional discrete time series $\boldsymbol{X}_{i,t} \in \mathbb{R}^{47}$ for $1 \leq i \leq N$ and $1 \leq t \leq T_i$ with 4-hour time steps. The actions of interests are the total volume of intravenous (IV) fluids and maximum dose of vasopressors administrated over each 4-hour period. We discretize two action variables into three levels, respectively. Our low corresponds to 1 - 2, medium corresponds to 3 and high corresponds to 4 - 5 in Komorowski et al. (2018). The combination of the two drugs makes $M = 3 \times 3 = 9$ possible actions in total. In the final processed dataset, we sampled 1000 unique adult ICU admissions, corresponding to unique trajectories to be fed into our algorithms. The observation length T_i varies across trajectories, with 12 987 records in total.

The reward signal is important and need crafted carefully in real applications. For the final reward, we follow Komorowski et al. (2018) and use hospital mortality or 90-day mortality. Specifically, when a patient survived, a positive reward was released at the end of each patient's trajectory; a negative reward was issued if the patient died. For the intermediate rewards, we follow Prasad et al. (2017) and associates reward to the health measurement of a patient. The detailed description of the data pre-processing is presented in Section A of the supplemental material. More information about the dataset can also be found in Komorowski et al. (2018); Prasad et al. (2017).

As demonstrated in Figure 4a, we can see that the residual plot of a simple regression is bi-model. This hints that there are at least two different coefficient groups in this population. Figure 4b is the residual plot of our proposed method. We can see that the residual plot for our algorithm is well-shaped and perform better in terms of prediction error. We then perform a principle component analysis on our coefficients matrix (averaged out across actions) as shown in figure 5. We can observe that there are two clear separate clusters within this coefficient matrix, which is a confirmation of the bi-model residual and a sign that simple k-means clustering would work well when we cluster the



(a) Residual distribution of simple linear regression is clearly bi-model, with two peaks at around -4 and 4.

(b) Residual of our algorithm is bell-shaped, with one peak at around 0. It also performs better in terms of prediction error.

Figure 4: Performance of our algorithm on MIMIC-III dataset.

β's into groups.

We summarize the estimated values of heterogeneously trained policy and homogeneously trained policy with various γ in table 2. The homogeneously trained policy is trained on whole data and evaluated on two groups that heterogenous methods returned. We can see that heterogeneously trained policy outperforms homogeneously trained policy by almost 40 % when $\gamma = 0.7$ on group 1 and around 30 % on group 2.

We then demonstrate two sample patients in table 3 and 4 to show the effectiveness of estimated policies. Patient 1 eventually died in hospital, and this policy is evaluated at two time points before his / her death. At this time point, many of his / her indices are extremely abnormal. For example, his / her Glasgow Coma Scale (GCS) decreased sharply from 15 to 6. This is a critical sign of urgent treatment. Indeed, policy trained under heterogeneous estimation recommends high intravenous volume and high / medium vasopressors, while policy trained under homogenous estimation still puts significant weights on medium intravenous volume and medium vasopressors. This conservative policy recommended by homogenous estimation would be dangerous. Patient 2 survives and all his / her indices are fairly normal at the evaluated time point, except for his / her relatively high respiratory rate (RR). He / she is also one of the youngest patients in the sample, aged around 56 years old. Still, the expert decides to injects little / no amount of both intravenous and vasopressors, and heterogeneously estimated policy also recommends so. However, homogeneously estimated policy also suggests medium volume of intravenous, which is unnecessary for such patient.

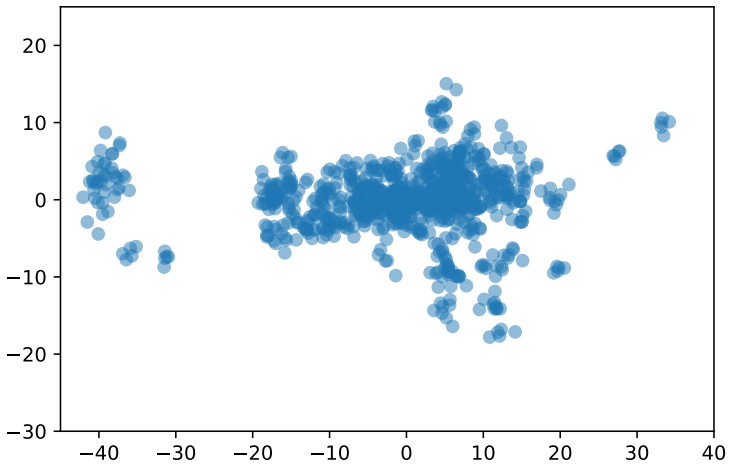


Figure 5: Plot of the first two principle components of the coefficient matrix (averaged across actions) produced by our algorithm. We can see there is a clear separation between two groups.

	Group 1		Group 2	
	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 0.5$	$\gamma = 0.7$
Heterogenous	-2.2140	-3.1851	-1.4433	-1.7793
Homogenous	-3.0964	-4.5735	-2.0634	-2.6487

Table 2: Estimated values for both groups by homogeneously estimated policy and heterogeneously estimated policy.

	Heterogenous	Homogenous	Expert
iv low, vaso low	< 0.0001	< 0.0001	0
iv low, vaso med	< 0.0001	0.071	0
iv low, vaso high	< 0.0001	<0.0001	0
iv med, vaso low	< 0.0001	0.10	0
iv med, vaso med	< 0.0001	0.14	0
iv med, vaso high	0.042	0.21	0
iv high, vaso low	0.23	< 0.0001	0
iv high, vaso med	0.36	0.26	0
iv high, vaso high	0.37	0.22	1

Table 3: Sample patient from group 1.

	Heterogenous	Homogenous	Expert
iv low, vaso low	0.72	0.36	1
iv low, vaso med	0.18	0.018	0
iv low, vaso high	< 0.0001	<0.0001	0
iv med, vaso low	< 0.0001	0.43	0
iv med, vaso med	< 0.0001	<0.0001	0
iv med, vaso high	< 0.0001	0.11	0
iv high, vaso low	0.09	< 0.0001	0
iv high, vaso med	< 0.001	0.075	0
iv high, vaso high	<0.001	<0.0001	0

Table 4: Sample patient from group 2.

8 Summary

Classical RL methods model and optimize the (action-)value function defined as the expectation of total return. Direct applications of such mean-value based RL to large-scale datasets may generate misleading results because of data heterogeneity. In this paper, we go beyond the mean-value based RL to allow for heterogeneity which is characterized by different values across sub-populations. We proposed ACPI for both the policy evaluation and control. We establish convergence rates and construct confidence intervals (CIs) for the estimators obtained by the ACPI. Our theoretical findings are validated on synthetic and real datasets. Particularly, the experiments on the well-recognized MIMIC-III dataset shows evidences of data heterogeneity and confirms the advantage of our new method.

Statistical analysis of policy evaluation and optimal control under the framework of RL and MDP have great potentials to facilitate dynamic decision making in a variety of real applications. We have demonstrated the importance of data heterogeneity when combining RL and large-scale dataset. There are several interesting directions for future research in this area. First, our method is based on the Bellman consistency equation to approximate the Q^π function. Its performance is guaranteed under the assumption of bounded state distribution shift, which is caused by the discrepancy between the behavior policy and the target policy. However, this assumption may not hold generally in real applications. It is of great interest to investigate various ways to relax such an assumption. Second, the theoretical results in this paper is developed for offline batch estimation. Developing online RL estimation with heterogeneity is an interesting topic for future research. Lastly, we estimate the optimal policy by a variation of Q learning. It would also be worthwhile to investigate the ways to incorporate heterogeneity in other RL methods such as policy gradient.

References

Barth-Maron, G., M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, D. TB, A. Muldal, N. Heess, and T. Lillicrap (2018). Distributed distributional deterministic policy gradients. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bellemare, M. G., W. Dabney, and R. Munos (2017). A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 449–458. JMLR. org.

Bondell, H. D. and B. J. Reich (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics 64*(1), 115–123.

Boyd, S., N. Parikh, and E. Chu (2011). *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

Burman, P., K.-W. Chen, et al. (1989). Nonparametric estimation of a regression function. *The Annals of Statistics 17*(4), 1567–1596.

Chen, X. and T. M. Christensen (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics 188*(2), 447–465.

Dabney, W., M. Rowland, M. G. Bellemare, and R. Munos (2018). Distributional reinforcement learning with quantile regression. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Gong, P., C. Zhang, Z. Lu, J. Z. Huang, and J. Ye (2013). A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13. JMLR.org.

Hocking, T. D., A. Joulin, F. Bach, and J.-P. Vert (2011). Clusterpath an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, pp. 1.

Huang, J. Z. et al. (1998). Projection estimation in multiple regression with application to functional anova models. *The annals of statistics 26*(1), 242–272.

Jiang, N. and L. Li (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661.

Johnson, A. E., T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark (2016). Mimic-iii, a freely accessible critical care database. *Scientific data 3*, 160035.

Ke, T., J. Fan, and Y. Wu (2013). Homogeneity in regression. *arXiv preprint arXiv:1303.7409*.

Komorowski, M., L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine 24*(11), 1716–1720.

Luckett, D. J., E. B. Laber, A. R. Kahkoska, D. M. Maahs, E. Mayer-Davis, and M. R. Kosorok (2019). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 1–34.

Luedtke, A. R. and M. J. Van Der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics 44*(2), 713.

Ma, S. and J. Huang (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association 112*(517), 410–423.

Mandel, T. S. (2017). *Better Education Through Improved Reinforcement Learning*. Ph. D. thesis.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 331–355.

Pan, W., X. Shen, and B. Liu (2013). Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *The Journal of Machine Learning Research 14*(1), 1865–1889.

Prasad, N., L. F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt (2017). A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. In *33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017*.

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rafferty, A. N., E. Brunskill, T. L. Griffiths, and P. Shafto (2016). Faster teaching via pomdp planning. *Cognitive science 40*(6), 1290–1332.

Reddy, S., S. Levine, and A. Dragan (2017). Accelerating human learning with deep reinforcement learning. In *NIPS workshop: teaching machines, robots, and humans*.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, pp. 189–326. Springer.

Rowland, M., M. Bellemare, W. Dabney, R. Munos, and Y. W. Teh (2018). An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37.

Rowland, M., R. Dadashi, S. Kumar, R. Munos, M. G. Bellemare, and W. Dabney (2019). Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 5528–5536.

Schulte, P. J., A. A. Tsiatis, E. B. Laber, and M. Davidian (2014). Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics 29*(4), 640.

Shen, X. and H.-C. Huang (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association 105*(490), 727–739.

Shi, C., S. Zhang, W. Lu, and R. Song (2020). Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*.

Sobel, M. J. (1982). The variance of discounted markov decision processes. *Journal of Applied Probability 19*(4), 794–802.

Song, R., W. Wang, D. Zeng, and M. R. Kosorok (2015). Penalized q-learning for dynamic treatment regimens. *Statistica Sinica 25*(3), 901.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Thomas, P. S., G. Theocharous, and M. Ghavamzadeh (2015). High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics 12*(4), 389–434.

Yang, D., L. Zhao, Z. Lin, T. Qin, J. Bian, and T.-Y. Liu (2019). Fully parameterized quantile function for distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 6190–6199.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics 38*(2), 894–942.

Zhou, Z., X. Li, and R. N. Zare (2017). Optimizing chemical reactions with deep reinforcement learning. *ACS central science 3*(12), 1337–1344.

Zhu, W., D. Zeng, and R. Song (2019). Proper inference for value function in high-dimensional q-learning for dynamic treatment regimes. *Journal of the American Statistical Association 114*(527), 1404–1417.

Appendix A Description of MIMIC-III Dataset

A.1 Intensive Care Unit Data

The data we use is the Medical Information Mart for Intensive Care version III (MIMIC-III) Database (Johnson et al., 2016), which is a freely available source of de-identified critical care data from 53,423 adult admissions and 7,870 neonates from 2001 – 2012 in six ICUs at a Boston teaching hospital. The database contain high-resolution patient data, including demographics, time-stamped measurements from bedside monitoring of vital signs, laboratory tests, illness severity scores, medications and procedures, fluid intakes and outputs, clinician notes and diagnostic coding.

We extract a cohort of sepsis patients, following the same data processing procedure as in Komorowski et al. (2018). Specifically, the adult patients included in the analysis satisfy the international consensus sepsis-3 criterion. The data includes 17,083 unique ICU admissions from five separate ICUs in one tertiary teaching hospital. Patient demographics and clinical characteristics are shown in Table 1 and Supplementary Table 1 of Komorowski et al. (2018).

Each patient in the cohort is characterized by a set of 47 variables, including demographics, Elixhauser premorbid status, vital signs, and laboratory values. Demographic information includes age, gender, weight. Vital signs include heart rate, systolic/diastolic blood pressure, respiratory rate et al. Laboratory values include glucose, total bilirubin, (partial) thromboplastin time et al. Patients' data were coded as multidimensional discrete time series with 4-hour time steps. The actions of interests are the total volume of intravenous (IV) fluids and maximum dose of vasopressors administrated over each 4-hour period.

All features were checked for outliers and errors using a frequency histogram method and uni-variate statistical approaches (Tukey's method). Errors and missing values are corrected when possible. For example, conversion of temperature from Fahrenheit to Celsius degrees and capping variables to clinically plausible values.

In the final processed data set, we have 17 621 unique ICU admissions, corresponding to unique trajectories fed into our algorithms.

A.1.1 Irregular Observational Time Series Data

For each ICU admission, we code patient's data as multivariate discrete time series with 4 hours time step. Each trajectory covers from up to 24h preceding until 48h following the estimated onset of sepsis, in order to capture the early phase of its management, in-

cluding initial resuscitation. The medical treatments of interest are the total volume of intravenous fluids and maximum dose of vasopressors administered over each 4 hour period. We use a time-limited parameter specific sample-and-hold approach to address the problem of missing or irregularly sampled data. The remaining missing data were interpolated in MIMIC-III using multivariate nearest-neighbor imputation. After processing, we have in total 278 598 sampled data points for the entire sepsis cohort.

A.1.2 State and Action Space Characterization

The state $\boldsymbol{X}_{i,t}$ is a 47-dimensional feature vector including fixed demographic information (age, weight, gender, admit type, ethnicity et al), vitals signs (heart rate, systolic/diastolic blood pressure, respiratory rate et al), and laboratory values (glucose, Creatinine, total bilirubin, partial thromboplastin time, paO_2, $paCO_2$ et al.).

For action space, we discretize two variables into five actions respectively according to Table in Komorowski et al. (2018). The combination of the two drugs makes $5 \times 5 = 25$ possible actions in total. The action A_t is a 2-dimensional vector, of which the first entry $a_t[0]$ specifies the dosages of IV fluids and the second $a_t[1]$ indicates the dosages of IV fluids and vasopressors, to be administrated over the next 4h interval.

A.2 Reward design

The reward signal is important and need crafted carefully in real applications. Komorowski et al. (2018) uses hospital mortality or 90-day mortality as the sole defining factor for the penalty and reward. Specifically, when a patient survived, a positive reward was released at the end of each patient's trajectory (a reward of 100); while a negative reward (a penalty of -100) was issued if the patient died. However, this reward design is sparse and provide little information at each step. Also, mortality may correlated with respect to the health statues of a patient. So it is reasonable to associate reward to the health measurement of a patient after an action is taken.

In this application, we build our reward signal based on physiological stability. Specifically, in our design, physiological stability is measured by vitals and laboratory values v_t with desired ranges $[v_{\min}, v_{\max}]$. Important variables related to sepsis include heart rate (HR), systolic blood pressure (SysBP), mean blood pressure (MeanBP), diastolic blood pressure (DiaBP), respiratory rate (RR), peripheral capillary oxygen saturation (SpO2), arterial lactate, creatinine, total bilirubin, glucose, white blood cell count, platelets count, (partial) thromboplastin time (PTT), and International Normalized Ratio (INR). We encode a penalty for exceeding desired ranges at each time step by a truncated Sigmoid

function, as well as a penalty for sharp changes in consecutive measurements.

$$r_{t+1} = \sum_{v} C_1 \left[\frac{1}{1+e^{-(v_t - v_{min})}} - \frac{1}{1+e^{-(v_t - v_{max})}} + 0.5 \right] - C_2 \left[\max\left(0, \frac{|v_{t+1} - v_t|}{v_t} - 0.2\right) \right],$$

Here, values v_t are the measurements of those vitals v believed to be indicative of physiological stability at time t, with desired ranges $[v_{min}, v_{max}]$. The penalty for exceeding these ranges at each time step is given by a truncated sigmoid function. The system also receives negative feedback when consecutive measurements see a sharp change.

Remark 9. *There are definitely improvements in shaping the reward space. For example, in medical situation, the definition of the normal range of a variable sometime depends demographic characterization. Also, sharp changes in a favorable direction should be rewarded.*

Appendix B Proofs: Oracle Estimators

The solution of (21) can be obtained by solving the estimating equation:

$$\widetilde{\boldsymbol{G}}(\pi, \boldsymbol{\theta}^{\pi}) = \frac{1}{NTJ} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widetilde{\boldsymbol{\Lambda}}_i^{\top} \boldsymbol{z}_{i,t} r_{i,t} - \widetilde{\boldsymbol{\Lambda}}_i^{\top} \boldsymbol{z}_{i,t} \left(\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1} \right)^{\top} \widetilde{\boldsymbol{\Lambda}}_i \boldsymbol{\theta}^{\pi} \right) = 0. \tag{26}$$

The resulting oracle estimator of the group coefficients obtained by (26) has the following decomposition:

$$\left(\widetilde{\boldsymbol{\theta}}^{\pi} - \mathring{\boldsymbol{\theta}}^{\pi} \right) = \widetilde{\Sigma}^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widetilde{\boldsymbol{\Lambda}}_i^{\top} \boldsymbol{z}_{i,t} (\varepsilon_{1,i,t} + \varepsilon_{2,i,t}) \right) = \widetilde{\Sigma}^{-1} \widetilde{\zeta}_1 + \widetilde{\Sigma}^{-1} \widetilde{\zeta}_2, \tag{27}$$

where

$$\widetilde{\Sigma} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\boldsymbol{\Lambda}}_i^\top \boldsymbol{z}_{i,t}(\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1})^\top \widetilde{\boldsymbol{\Lambda}}_i, \tag{28}$$

$$\widetilde{\zeta}_1 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\boldsymbol{\Lambda}}_i^\top \boldsymbol{z}_{i,t}\varepsilon_{1,i,t}, \tag{29}$$

$$\widetilde{\zeta}_2 = \sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\boldsymbol{\Lambda}}_i^\top \boldsymbol{z}_{i,t}\varepsilon_{2,i,t}, \tag{30}$$

$$\varepsilon_{1,i,t} = r_{i,t} + \gamma \cdot \sum_{a\in\mathcal{A}} Q(\pi, X_{i,t+1}, a)\pi(a \mid \boldsymbol{X}_{i,t+1}) - Q(\pi, X_{i,t}, A_{i,t}) \tag{31}$$

$$\varepsilon_{2,i,t} = \gamma \cdot \sum_{a\in\mathcal{A}} e(\pi, \boldsymbol{X}_{i,t+1}, a)\pi(a \mid \boldsymbol{X}_{i,t+1}) - \sum_{a\in\mathcal{A}} e(\pi, \boldsymbol{X}_{i,t}, a)\mathbb{1}(A_{i,t} = a), \tag{32}$$

$$e(\pi, \boldsymbol{x}, a) = Q(\pi, \boldsymbol{x}, a) - \boldsymbol{\phi}(\boldsymbol{x})^\top \mathring{\boldsymbol{\beta}}_{a,i}^{\pi}. \tag{33}$$

$$\widetilde{\boldsymbol{\beta}}^{\pi} = (\boldsymbol{W}\otimes \boldsymbol{I}_{JM})\widetilde{\boldsymbol{\theta}}^{\pi}. \tag{34}$$

Proposition B.1 (Oracle estimator ℓ_2 convergence)**.** *Suppose Assumption 4.2 – 4.6 hold.* $N_{min} = \min_{1\le k\le K} N_k$ *and* $N_{max} = \max_{1\le k\le K} N_k$*. If* $K = \mathrm{o}(N_{\min}T)$*,* $J \ll \sqrt{N_{\min}T}/\log(N_{\min}T)$*,* $J^{\kappa/p} \gg \sqrt{N_{\max}T}$*, as either* $N_{\min}\to\infty$ *or* $T\to\infty$*,*

$$\left\|\widetilde{\boldsymbol{\theta}}^{\pi} - \mathring{\boldsymbol{\theta}}^{\pi}\right\|_2 = \mathcal{O}_p\left(J^{1/2}(TN_{\min})^{-1/2}\right) + \mathcal{O}_p\left(J^{-\kappa/p}\right).$$

If we have balanced groups, that is, $N_{\min} \asymp N_{\max} \asymp N$*, then*

$$\left\|\widetilde{\boldsymbol{\theta}}^{\pi} - \mathring{\boldsymbol{\theta}}^{\pi}\right\|_2 = \mathcal{O}_p\left(J^{1/2}(TN)^{-1/2}\right) + \mathcal{O}_p\left(J^{-\kappa/p}\right).$$

Proof. Recall that $\widetilde{\boldsymbol{\Lambda}}_i = \boldsymbol{\Lambda}_i(\boldsymbol{W}\otimes \boldsymbol{I}_{JM})$. We have

$$\widetilde{\Sigma} = \mathrm{bdiag}\left(P_k\widetilde{\Sigma}_k\right)_{1\le k\le K}, \quad \text{and} \quad \widetilde{\zeta}_i = \begin{bmatrix} P_1\widetilde{\zeta}_{i1}^\top & \cdots & P_k\widetilde{\zeta}_{iK}^\top \end{bmatrix}^\top, \quad \text{for } i = 1, 2. \tag{35}$$

where $P_k = N_k/N$, $\widetilde{\Sigma}_k$ and $\widetilde{\zeta}_{ik}$ are defined in (42) and (44), respectively. Then, by decomposition (27) and Proposition D.1, we have

$$\left\|\widetilde{\boldsymbol{\theta}}^{\pi} - \mathring{\boldsymbol{\theta}}^{\pi}\right\|_2 = \left(\sum_{k=1}^{K}\left\|\Sigma_k^{-1}\zeta_{1k} + \Sigma_k^{-1}\zeta_{2k}\right\|_2^2\right)^{1/2} = \mathcal{O}_p\left(\sqrt{\frac{KJ}{TN_{min}}}\right) + \mathcal{O}_p\left(KJ^{-\kappa/p}\right).$$

The desired result follows by noticing that K is fixed and P_k is bounded away from zero

under our assumption. □

Proof of Theorem 4.7

Proof. For any $\boldsymbol{\nu} = [\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_K] \in \mathbb{R}^{JMK}$, $\boldsymbol{\nu}_k \in \mathbb{R}^{JM}$ for $k \in [K]$, satisfying $J^{-\kappa/p} \ll \sqrt{N_{\min} T \left(1 + \|\boldsymbol{\nu}\|_2^{-2}\right)}$,

$$\begin{aligned}\sqrt{NT} \cdot \boldsymbol{\nu}^\top \left(\widetilde{\theta}^\pi - \mathring{\theta}^\pi\right) &= \sqrt{NT} \cdot \boldsymbol{\nu}^\top \widetilde{\Sigma}^{-1} \widetilde{\zeta}_1 + \mathrm{o}_{\mathrm{p}}(1) \\ &= \sum_{k \in [K]} \sqrt{N_k T / P_k} \cdot \boldsymbol{\nu}_k^\top \widetilde{\Sigma}_k^{-1} \widetilde{\zeta}_{1k} + \mathrm{o}_{\mathrm{p}}(1) \\ &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{k \in [K]} \boldsymbol{\nu}_k^\top \widetilde{\Sigma}_k^{-1} \widetilde{\boldsymbol{\Omega}}_k (\widetilde{\Sigma}_k^\top)^{-1} \boldsymbol{\nu}_k / P_k\right),\end{aligned}$$

where the second equality follows from (35) and the distribution is obtained by applying Theorem D.2. Applying (35) and $\widetilde{\boldsymbol{\Omega}} = \mathrm{bdiag}\left(P_k \widetilde{\boldsymbol{\Omega}}_k\right)$, we have for any $\boldsymbol{\nu} \in \mathbb{R}^{JMK}$ satisfying $\|\boldsymbol{\nu}\|_2^{-1} = \mathcal{O}(1)$,

$$\frac{\sqrt{NT} \boldsymbol{\nu}^\top \left(\widetilde{\theta}^\pi - \mathring{\theta}^\pi\right)}{\sqrt{\boldsymbol{\nu}^\top \widetilde{\Sigma}^{-1} \widetilde{\boldsymbol{\Omega}} (\widetilde{\Sigma}^\top)^{-1} \boldsymbol{\nu}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

□

Proof of Corollary 4.8.

Proof. The proof uses the result in Corollary D.3 and is similar to that of Theorem 4.7. □

Proof of Theorem 4.9.

Proof. By decomposition (27), we have

$$\left\|\widetilde{\theta}^\pi - \mathring{\theta}^\pi\right\|_\infty \leq \left\|\left(\widetilde{\Sigma}\right)^{-1}\right\|_\infty \left\|\widetilde{\zeta}_1\right\|_\infty + \left\|\left(\widetilde{\Sigma}\right)^{-1}\right\|_\infty \left\|\widetilde{\zeta}_1\right\|_\infty$$

We bound each term on the right hand side as follows.

Using (35) and Lemma D.11 (iii), we have

$$\left\|(\widetilde{\Sigma})^{-1}\right\|_\infty = \left\|\mathrm{bdiag}\left((\widetilde{\Sigma})^{-1}\right)\right\|_\infty \leq \max_k \left\|(P_k \widetilde{\Sigma}_k)^{-1}\right\|_\infty = 6C^{-1}\sqrt{J}(N/N_{\min}),$$

with probability at least $1 - \mathcal{O}\left((N_{\min} T)^{-2}\right)$.

By Lemma D.13 (iii), we have

$$\Pr\left(\|\zeta_{1k}\|_\infty > c\sqrt{2\log(N_k T)/(N_k T)}\right) \leq 2JM(N_k T)^{-2}.$$

for some positive constant c. Lemma D.13 (iv) shows that $\|\zeta_{2k}\|_\infty \le c_1 J^{-\kappa/p} \ll c_1 (N_k T)^{-1/2}$ almost surely.

By union bound and Lemma D.13 (iii), we have

$$\begin{aligned}
\Pr\left(\left\|\widetilde{\zeta}_1\right\|_\infty > c(N_{\max}/N)\sqrt{2\log(N_{\max}T)/N_{\max}T}\right) &\le \sum_{k=1}^{K} \Pr\left(\left\|P_k\widetilde{\zeta}_{1k}\right\|_\infty > (N_{\max}/N)\sqrt{2\log(N_{\max}T)/N_{\max}T}\right) \\
&\le \sum_{k=1}^{K} \Pr\left(\left\|P_k\widetilde{\zeta}_{1k}\right\|_\infty > P_k\sqrt{2\log(N_k T)/N_k T}\right) \\
&= \sum_{k=1}^{K} 2JM(N_k T)^{-2} \\
&\le 2JMK(N_{\min}T)^{-2}
\end{aligned}$$

for some positive constant c. Similarly, using Lemma D.13 (iv), we have

$$\left\|\widetilde{\zeta}_2\right\|_\infty = \max_{1\le k\le K} \|P_k\zeta_{2k}\|_\infty \le c_1 J^{-\kappa/p} N_{max}/N \ll c_1 N_{max}/N (N_{\max}T)^{-1/2}, \quad \text{almost surely.}$$

Thus, by union bounds, we have

$$\left\|\widetilde{\theta}^\pi - \mathring{\theta}^\pi\right\|_\infty \le 6cC^{-1}(N_{\max}/N_{\min})\sqrt{2J\log(N_{\max}T)/N_{\max}T}$$

holds with probability at least $1 - 2JMK(N_{\min}T)^{-2} - \mathcal{O}\left((N_{\min}T)^{-2}\right)$. □

Appendix C Feasible estimator

Proof of Theorem 4.12.

Proof. Note that all coefficients are with respect to a given policy π, we drop superscript π for brevity. We use $\boldsymbol{B} = \left[\beta_1, \cdots, \beta_N\right]$ and $\beta = \text{vec}(\boldsymbol{B})$ (also $\boldsymbol{\Theta} = [\theta_1, \cdots, \theta_K]$ and $\theta = \text{vec}(\boldsymbol{\Theta})$) interchangeably. It is easy to see that $\boldsymbol{B} = \boldsymbol{\Theta}\boldsymbol{W}^\top$, $\beta_i = \boldsymbol{\Theta}\boldsymbol{w}_{i\cdot} = \boldsymbol{\Lambda}_i\beta$, and $\beta = (\boldsymbol{W}\otimes\boldsymbol{I}_p)\theta^\pi$. Define

$$\begin{aligned}
&L(\boldsymbol{B}) = \boldsymbol{G}(\pi,\boldsymbol{B})^\top \boldsymbol{G}(\pi,\boldsymbol{B}), \quad P(\boldsymbol{B},\lambda) = \frac{1}{N^2}\sum_{1\le i<j\le N} \mathcal{P}\left((JM)^{-1/2}\left\|\beta_i - \beta_j\right\|_2, \lambda\right), \\
&\widetilde{L}(\boldsymbol{\Theta}) = \widetilde{\boldsymbol{G}}(\pi,\boldsymbol{\Theta})^\top \widetilde{\boldsymbol{G}}(\pi,\boldsymbol{\Theta}), \quad \widetilde{P}(\boldsymbol{\Theta},\lambda) = P\left(\boldsymbol{\Theta}\boldsymbol{W}^\top, \lambda\right).
\end{aligned} \tag{36}$$

where

$$G(\pi, \boldsymbol{B}) = \frac{1}{NTJ} \sum_{i=1}^{N} \sum_{t=1}^{T} \boldsymbol{\Lambda}_i^\top \boldsymbol{Z}_{i,t} Y_{i,t} - \boldsymbol{\Lambda}_i^\top \boldsymbol{Z}_{i,t} (\boldsymbol{Z}_{i,t} - \gamma \boldsymbol{U}_{\pi,i,t+1})^\top \boldsymbol{\Lambda}_i \boldsymbol{\beta},$$

$$\widetilde{G}(\pi, \boldsymbol{\Theta}) = \frac{1}{NTJ} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{\boldsymbol{\Lambda}}_i^\top \boldsymbol{Z}_{i,t} Y_{i,t} - \widetilde{\boldsymbol{\Lambda}}_i^\top \boldsymbol{Z}_{i,t} (\boldsymbol{Z}_{i,t} - \gamma \boldsymbol{U}_{\pi,i,t+1})^\top \widetilde{\boldsymbol{\Lambda}}_i \boldsymbol{\theta}.$$

Let $\mathcal{M}_{\mathcal{G}}$ be the subspace of $\mathbb{R}^{JM \times N}$, defined as

$$\mathcal{M}_{\mathcal{G}} = \left\{ \boldsymbol{B} \in \mathbb{R}^{JM \times N} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j, \text{ for any } i, j \in \mathcal{G}_k, 1 \le k \le K \right\}.$$

Recall that $\boldsymbol{W}$ is the true $N \times K$ *membership matrix*, then for each $\boldsymbol{B} \in \mathcal{M}_{\mathcal{G}}$, it can be written as $\boldsymbol{B} = \boldsymbol{\Theta} \boldsymbol{W}^\top$ for some $\boldsymbol{\Theta} \in \mathbb{R}^{JM \times K}$. Also by matrix calculation, we have $\boldsymbol{W}^\top \boldsymbol{W} = \operatorname{diag}\{N_1, \cdots, N_K\}$ where N_k denotes the number of elements in $\mathcal{G}_k$.

Let $\mathcal{T} : \mathcal{M}_G \to \mathbb{R}^{JM \times K}$ be the mapping such that $\mathcal{T}(\boldsymbol{B})$ is the $JM \times K$ matrix whose k-th column equals to the common value of $\boldsymbol{\beta}_i$ for $i \in \mathcal{G}_k$. Let $\overline{\mathcal{T}} : \mathbb{R}^{JM \times N} \to \mathbb{R}^{JM \times K}$ be the mapping such that $\overline{\mathcal{T}}(\boldsymbol{B}) = \boldsymbol{B} \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{W})^{-1}$. Clearly, when $\boldsymbol{B} \in \mathcal{M}_{\mathcal{G}}$, $\mathcal{T}(\boldsymbol{B}) = \overline{\mathcal{T}}(\boldsymbol{B})$.

By calculation, we have $P(\boldsymbol{B}, \lambda) = \widetilde{P}(\mathcal{T}(\boldsymbol{B}), \lambda) = 0$ for every $\boldsymbol{B} \in \mathcal{M}_{\mathcal{G}}$ and $P\left(\mathcal{T}^{-1}(\boldsymbol{\Theta}), \lambda\right) = \widetilde{P}(\boldsymbol{\Theta}, \lambda)$ for every $\boldsymbol{\Theta} \in \mathbb{R}^{JM \times K}$.

Consider the neighborhood of $\mathring{\boldsymbol{B}}$:

$$\mathring{\mathcal{N}} = \left\{ \boldsymbol{B} \in \mathbb{R}^{JM \times N}, \left\| \operatorname{vec}\left(\boldsymbol{B} - \mathring{\boldsymbol{B}}\right) \right\|_\infty \le \phi_{NT} \right\}.$$

According to Theorem 4.9, there is an event E_1 such that on the event E_1,

$$\left\| \operatorname{vec}\left(\widetilde{\boldsymbol{B}} - \mathring{\boldsymbol{B}}\right) \right\|_\infty \le \phi_{NT},$$

and $\Pr(E_1^C) \le 2JMK(N_{min}T)^{-2} + \mathcal{O}\left((N_{\min}T)^{-2}\right)$. Hence, $\widetilde{\boldsymbol{B}} \in \mathring{\mathcal{N}}$ on the event E_1.

For any $\boldsymbol{B} \in \mathbb{R}^{JM \times N}$, let $\overline{\boldsymbol{\Theta}} = \overline{\mathcal{T}}(\boldsymbol{B})$, $\overline{\boldsymbol{B}} = \mathcal{T}^{-1}\left(\overline{\boldsymbol{\Theta}}\right)$. Lemma C.1 shows that, on event E_1, $\mathcal{L}\left(\widetilde{\boldsymbol{B}}\right) < \mathcal{L}\left(\overline{\boldsymbol{B}}\right)$ for any $\boldsymbol{B}$ whose $\overline{\boldsymbol{B}} \neq \widetilde{\boldsymbol{B}}$. Lemma C.2 shows that there is an event E_2 such that $\Pr\left(E_2^C\right) \le 2((NT)^{-1})$. On $E_1 \cap E_2$, there is a neighborhood of $\widetilde{\boldsymbol{B}}$, denote by $\widetilde{\mathcal{N}}$, such that $\mathcal{L}\left(\overline{\boldsymbol{B}}\right) < \mathcal{L}(\boldsymbol{B})$ for any $\boldsymbol{B} \in \mathring{\mathcal{N}} \cap \widetilde{\mathcal{N}}$ for sufficient large N or T. Therefore, we have $\mathcal{L}\left(\operatorname{vec}\left(\overline{\boldsymbol{B}}\right)\right) < \mathcal{L}(\operatorname{vec}(\boldsymbol{B}))$ for any $\boldsymbol{B} \in \mathring{\mathcal{N}} \cap \widetilde{\mathcal{N}}$ and $\boldsymbol{B} \neq \widetilde{\boldsymbol{B}}$, so that $\widetilde{\boldsymbol{B}}$ is a strict local minimizer of $\mathcal{L}(\operatorname{vec}(\boldsymbol{B}))$ given in (10) on the event $E_1 \cap E_2$ with $\Pr(E_1 \cap E_2) \ge 1 - 2(K + JM + 1)(NT)^{-1}$ for sufficient large N or T. □

Lemma C.1. *For any $\boldsymbol{B} \in \mathbb{R}^{JM \times N}$, let $\overline{\boldsymbol{B}} = \mathcal{T}^{-1}\left(\overline{\mathcal{T}}(\boldsymbol{B})\right)$. On event E_1, $\mathcal{L}\left(\widetilde{\boldsymbol{B}}\right) < \mathcal{L}\left(\overline{\boldsymbol{B}}\right)$ for any $\boldsymbol{B} \in \mathbb{R}^{JM \times N}$ whose $\overline{\boldsymbol{B}} \neq \widetilde{\boldsymbol{B}}$.*

Proof. Note that $\mathcal{L}(\cdot) = L(\cdot) + P(\cdot)$, we consider each term on the RHS. First, we have $L\left(\widetilde{\boldsymbol{B}}\right) < L\left(\overline{\boldsymbol{B}}\right)$ by the following argument. Since $\widetilde{\boldsymbol{\Theta}}$ is the unique global minimizer of $\widetilde{L}(\boldsymbol{\Theta})$, then $\widetilde{L}\left(\widetilde{\boldsymbol{\Theta}}\right) < \widetilde{L}\left(\overline{\mathcal{T}}(\boldsymbol{B})\right)$ for all $\overline{\mathcal{T}}(\boldsymbol{B}) \neq \widetilde{\boldsymbol{\Theta}}$. By definition we have $\widetilde{L}\left(\widetilde{\boldsymbol{\Theta}}\right) = L\left(\widetilde{\boldsymbol{B}}\right)$ and $\widetilde{L}\left(\overline{\mathcal{T}}(\boldsymbol{B})\right) = L\left(\overline{\boldsymbol{B}}\right)$ and thus the result.

Now we study the penalty term $P\left(\widetilde{\boldsymbol{B}}\right)$ and $P\left(\overline{\boldsymbol{B}}\right)$. The group-wise coefficients satisfy

$$\left\|\boldsymbol{\theta} - \mathring{\boldsymbol{\theta}}\right\|_\infty = \sup_k \left\| N_k^{-1} \sum_{i \in \mathcal{G}} \left(\boldsymbol{\beta}_i - \mathring{\boldsymbol{\beta}}_i\right) \right\|_\infty \leq \left\|\boldsymbol{\beta} - \mathring{\boldsymbol{\beta}}\right\|_\infty. \tag{37}$$

Then, by Assumption 4.11, we have for any pair of groups $k \neq l$

$$(JM)^{-1/2} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_l\|_2^2 \geq (JM)^{-1/2} \left\|\mathring{\boldsymbol{\theta}}_k - \mathring{\boldsymbol{\theta}}_l\right\|_2^2 - 2\left\|\boldsymbol{\theta} - \mathring{\boldsymbol{\theta}}\right\|_\infty \geq d_{NT} - 2\phi_{NT} > c\lambda.$$

By Assumption 4.10, we have $\widetilde{P}\left(\widetilde{\boldsymbol{\Theta}}\right) = \widetilde{P}\left(\overline{\boldsymbol{\Theta}}\right) = C_N$. Further we have $P\left(\widetilde{\boldsymbol{B}}\right) = P\left(\overline{\boldsymbol{B}}\right) = C_N$ since by definition we have $P\left(\widetilde{\boldsymbol{B}}\right) = \widetilde{P}\left(\widetilde{\boldsymbol{\Theta}}\right)$ and $P\left(\overline{\boldsymbol{B}}\right) = \widetilde{P}\left(\overline{\boldsymbol{\Theta}}\right)$ where $\overline{\boldsymbol{\Theta}} = \overline{\mathcal{T}}(\boldsymbol{B})$. As a result,

$$\mathcal{L}\left(\widetilde{\boldsymbol{B}}\right) = L\left(\widetilde{\boldsymbol{B}}\right) + P\left(\widetilde{\boldsymbol{B}}\right) < L\left(\overline{\boldsymbol{B}}\right) + P\left(\overline{\boldsymbol{B}}\right) = \mathcal{L}\left(\overline{\boldsymbol{B}}\right).$$

□

Lemma C.2. *For any $\boldsymbol{B} \in \mathbb{R}^{JM \times N}$, let $\overline{\boldsymbol{B}} = \mathcal{T}^{-1}\left(\overline{\mathcal{T}}(\boldsymbol{B})\right)$. There is an event E_2 such that $\Pr\left(E_2^C\right) \leq 2((NT)^{-1})$. On $E_1 \cap E_2$, there is a neighborhood $\widetilde{\mathcal{N}}$ of $\widetilde{\boldsymbol{B}}$ such that $\mathcal{L}\left(\overline{\boldsymbol{B}}\right) < \mathcal{L}(\boldsymbol{B})$ for any $\boldsymbol{B} \in \mathring{\mathcal{N}} \cap \widetilde{\mathcal{N}}$ for sufficient large N or T.*

Proof. For a positive sequence δ_{NT}, let $\widetilde{\mathcal{N}}_{NT} = \left\{\boldsymbol{\beta} : \left\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right\|_2 \leq \delta_{NT}\right\}$ be a neighborhood of $\widetilde{\boldsymbol{\beta}}$. Similar to (37), we have, for any $\boldsymbol{\beta} \in \widetilde{\mathcal{N}}_{NT} \cap \mathring{\mathcal{N}}$,

$$\left\|\overline{\boldsymbol{\theta}} - \mathring{\boldsymbol{\theta}}\right\|_\infty \leq \left\|\overline{\boldsymbol{\beta}} - \mathring{\boldsymbol{\beta}}\right\|_\infty \leq \phi_{NT},$$
$$\left\|\overline{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\right\|_\infty \leq \left\|\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right\|_\infty \leq \delta_{NT}.$$

thus $\overline{\boldsymbol{\beta}} \in \widetilde{\mathcal{N}}_{NT} \cap \mathring{\mathcal{N}}$. Let $\boldsymbol{\beta}^\iota = \iota\boldsymbol{\beta} + (1-\iota)\overline{\boldsymbol{\beta}}$ for some $\iota \in (0,1)$, we have

$$\left\|\boldsymbol{\beta}^\iota - \mathring{\boldsymbol{\beta}}\right\|_\infty \leq \left\|\boldsymbol{\beta} - \mathring{\boldsymbol{\beta}}\right\|_\infty \leq \phi_{NT},$$

By Taylor's expansion, we have

$$\begin{aligned}\mathcal{L}(\beta)-\mathcal{L}\left(\overline{\beta}\right) &= \boldsymbol{G}(\pi,\beta)^{\top}\boldsymbol{G}(\pi,\beta)-\boldsymbol{G}\left(\pi,\overline{\beta}\right)^{\top}\boldsymbol{G}\left(\pi,\overline{\beta}\right)\\ &\quad + P(\beta,\lambda)-P\left(\overline{\beta},\lambda\right)\\ &= \frac{1}{N^2}(I_1+I_2).\end{aligned}$$

where

$$I_1 = 2(TJ)^{-1}\boldsymbol{G}(\beta^{\iota})^{\top}\Sigma\left(\beta-\overline{\beta}\right),\quad \text{and}\quad I_2=\sum_{1\le i\le N}\left(\frac{\partial P(\beta,\lambda)}{\partial\beta_i}\Big|_{\beta=\beta^{\iota}}\right)^{\top}\left(\beta_i-\overline{\beta}_i\right),$$

for some $\beta^{\iota}=\iota\beta+(1-\iota)\overline{\beta}$ with $\iota\in(0,1)$.

Firstly, we consider I_2 which can be rewritten as

$$\begin{aligned}I_2 &= \sum_{1\le i\le N}\left(\frac{\partial P(\beta,\lambda)}{\partial\beta_i}\Big|_{\beta=\beta^{\iota}}\right)^{\top}\left(\beta_i-\overline{\beta}_i\right)\\ &= \lambda\sum_{i<j}\rho'\left((JM)^{-1/2}\left\|\beta_i^{\iota}-\beta_j^{\iota}\right\|_2\right)(JM)^{-1/2}\left\|\beta_i^{\iota}-\beta_j^{\iota}\right\|_2^{-1}\left(\beta_i^{\iota}-\beta_j^{\iota}\right)^{\top}\left(\beta_i-\overline{\beta}_i\right)\\ &\quad+\lambda\sum_{i>j}\rho'\left((JM)^{-1/2}\left\|\beta_i^{\iota}-\beta_j^{\iota}\right\|_2\right)(JM)^{-1/2}\left\|\beta_i^{\iota}-\beta_j^{\iota}\right\|_2^{-1}\left(\beta_i^{\iota}-\beta_j^{\iota}\right)^{\top}\left(\beta_i-\overline{\beta}_i\right)\\ &= \lambda\sum_{i<j}\rho'\left((JM)^{-1/2}\left\|\beta_i^{\iota}-\beta_j^{\iota}\right\|_2\right)(JM)^{-1/2}\left\|\beta_i^{\iota}-\beta_j^{\iota}\right\|_2^{-1}\left(\beta_i^{\iota}-\beta_j^{\iota}\right)^{\top}\left(\beta_i-\overline{\beta}_i\right)\\ &\quad+\lambda\sum_{j>i}\rho'\left((JM)^{-1/2}\left\|\beta_j^{\iota}-\beta_i^{\iota}\right\|_2\right)(JM)^{-1/2}\left\|\beta_j^{\iota}-\beta_i^{\iota}\right\|_2^{-1}\left(\beta_j^{\iota}-\beta_i^{\iota}\right)^{\top}\left(\beta_j-\overline{\beta}_j\right)\\ &= \lambda\sum_{i<j}\rho'\left((JM)^{-1/2}\left\|\beta_i^{\iota}-\beta_j^{\iota}\right\|_2\right)(JM)^{-1/2}\left\|\beta_i^{\iota}-\beta_j^{\iota}\right\|_2^{-1}\left(\beta_i^{\iota}-\beta_j^{\iota}\right)^{\top}\left(\left(\beta_i-\overline{\beta}_i\right)-\left(\beta_j-\overline{\beta}_j\right)\right).\end{aligned}$$

Note that, for any $i,j\in\mathcal{G}_k$, $\overline{\beta}_i=\overline{\beta}_j$ and $\beta_i^{\iota}-\beta_j^{\iota}=\iota\left(\beta_i-\beta_j\right)$, while, for any $i\in\mathcal{G}_k$, $j\in\mathcal{G}_l$, $k\neq l$,

$$(JM)^{-1/2}\left\|\beta_i^{\iota}-\beta_j^{\iota}\right\|_2\ge\min_{k\neq l}(JM)^{-1/2}\left\|\mathring{\theta}_k-\mathring{\theta}_l\right\|_2-2\left\|\beta-\mathring{\beta}\right\|_{\infty}\ge d_{NT}^2-2\phi_{NT}>a\lambda,$$

and thus $\rho'\left((JM)^{-1/2}\left\|\boldsymbol{\beta}_i^\iota-\boldsymbol{\beta}_j^\iota\right\|_2\right)=0$. As a result, we have

$$I_2=\lambda\sum_{k=1}^{K}\sum_{i,j\in\mathcal{G}_k,i<j}\rho'\left((JM)^{-1/2}\iota\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right\|_2\right)(JM)^{-1/2}\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right\|_2.$$

Note that, since $\overline{\boldsymbol{\beta}}\in\widetilde{\mathcal{N}}_{NT}\cap\mathring{\mathcal{N}}$, we have

$$(JM)^{-1}\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right\|_2^2\le 2\left\|\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right\|_\infty^2\le 2\left\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right\|_\infty^2+2\left\|\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\right\|_\infty^2\le 4\delta_{NT}.$$

Hence, by concavity of $\rho(\cdot)$, we have

$$I_2\ge 2\lambda\sum_{k=1}^{K}\sum_{i,j\in\mathcal{G}_k,i<j}\rho'(4\delta_{NT})(JM)^{-1/2}\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right\|_2.$$

Now we consider I_1 which can be rewritten as

$$\begin{aligned}
I_1&=\boldsymbol{G}(\boldsymbol{\beta}^\iota)^\top\Sigma\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)\\
&=\sum_{i=1}^{N}(\boldsymbol{G}_i^\iota)^\top\Sigma_i\left(\boldsymbol{\beta}_i-\overline{\boldsymbol{\beta}}_i\right)\\
&=\sum_{1\le k\le K}\sum_{i,j\in\mathcal{G}_k}(N_k)^{-1}\left(\Sigma_i^\top\boldsymbol{G}_i^\iota\right)^\top\left(\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right)\\
&=\sum_{1\le k\le K}\sum_{i,j\in\mathcal{G}_k,i<j}(N_k)^{-1}\left(\Sigma_i^\top\boldsymbol{G}_i^\iota-\Sigma_j^\top\boldsymbol{G}_j^\iota\right)^\top\left(\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right).
\end{aligned}$$

where the third equation follows from the fact that $\sum_{j\in\mathcal{G}_k}\left(\boldsymbol{\beta}_j-\overline{\boldsymbol{\beta}}_j\right)=0$ and

$$\Sigma_i=(TJ)^{-1}\sum_{t=1}^{T}\boldsymbol{Z}_{i,t}(\boldsymbol{Z}_{i,t}-\gamma\boldsymbol{U}_{\pi,i,t+1})^\top,\quad\text{and}\quad\boldsymbol{G}_i^\iota=(TJ)^{-1}\sum_{t=1}^{T}\boldsymbol{Z}_{i,t}\left(\varepsilon_{1,i,t}+\varepsilon_{2,i,t}^\iota\right).$$

By Lemma D.10, we have $\max_{1\le i\le N}\|\Sigma_i\|_2\le 4c_1^2TJ$. We have

$$\begin{aligned}
\max_{i,j}\left\|\Sigma_i\boldsymbol{G}_i^\iota-\Sigma_j\boldsymbol{G}_j^\iota\right\|_2&\le 2\max_{1\le i\le N}\left\|\Sigma_i\boldsymbol{G}_i^\iota\right\|_2\\
&\le 2\max_{1\le i\le N}\|\Sigma_i\|_2\max_{1\le i\le N}\left(\left\|(TJ)^{-1}\sum_{t=1}^{T}\boldsymbol{Z}_{i,t}\varepsilon_{1,i,t}\right\|_2+\left\|(TJ)^{-1}\sum_{t=1}^{T}\boldsymbol{Z}_{i,t}r_{i,t}^\iota\right\|_2\right)\\
&\le 8c\cdot\left((c_1+2c_2+2c_3J^{-\kappa/p})+c_4\sqrt{J}\phi_{NT}\right).
\end{aligned}$$

for some positive constant c_j, $j = 1,2,3,4$. Hence, we have

$$\begin{aligned}
|I_1| &= \sum_{1\le k\le K}\sum_{i,j\in\mathcal{G}_k, i<j} (N_k)^{-1}\left(\Sigma_i \boldsymbol{G}_i^{\iota} - \Sigma_j \boldsymbol{G}_j^{\iota}\right)^{\top}\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right) \\
&\le \sum_{1\le k\le K} (N_k)^{-1} \max_{i,j}\left\|\Sigma_i \boldsymbol{G}_i^{\iota} - \Sigma_j \boldsymbol{G}_j^{\iota}\right\|_2 \sum_{i,j\in\mathcal{G}_k, i<j}\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|_2 \\
&\le (N_{\min})^{-1} \max_{i,j}\left\|\Sigma_i \boldsymbol{G}_i^{\iota} - \Sigma_j \boldsymbol{G}_j^{\iota}\right\|_2 \sum_{1\le k\le K}\sum_{i,j\in\mathcal{G}_k, i<j}\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|_2.
\end{aligned}$$

Since $\lambda \gg \max\left(N_{\min}^{-1}\sqrt{J}, \phi_{NT}\right)$, $J \ll \sqrt{N_{\min}T}/\log(N_{\min}T)$, and $J^{\kappa/p} \gg \sqrt{N_{\max}T}$, we have

$$\lambda \gg N_{\min}^{-1}\sqrt{J} \gg N_{\min}^{-1} J^{1/2-\kappa/p}, \quad \text{and} \quad \lambda \gg N_{\min}^{-1} J\phi_{NT}.$$

Let $\delta_{NT} = \mathrm{o}(1)$, then $\rho'(2\iota\delta_{NT}) \to 1$. Therefore, we have

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{B}, \overline{\boldsymbol{\Theta}}\right) - \mathcal{L}\left(\overline{\boldsymbol{\beta}}\right) &\ge \left(\lambda\rho'(4\delta_{NT})(JM)^{-1/2} - (N_{\min})^{-1}\left(8c\left((c_1 + 2c_2 + 2c_3 J^{-\kappa/p}) + c_4\sqrt{J}\phi_{NT}\right)\right)\right) \\
&\quad \cdot \sum_{1\le k\le K}\sum_{i,j\in\mathcal{G}_k, i<j}\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|_2 \\
&\ge 0,
\end{aligned}$$

for sufficiently large $N_{\min}$. □

Proof of Corollary 4.13.

Proof. Let $\boldsymbol{u} = (\widehat{\boldsymbol{w}}_{i\cdot}^{\top} \otimes \boldsymbol{I}_{JM})\boldsymbol{\nu}$

$$\widehat{\sigma}_{\beta_i}^{-1}\boldsymbol{\nu}^{\top}\left(\widehat{\boldsymbol{\beta}}_i^{\pi} - \mathring{\boldsymbol{\beta}}_i^{\pi}\right) = \frac{\boldsymbol{u}^{\top}\left(\boldsymbol{\theta}^{\pi} - \mathring{\boldsymbol{\theta}}^{\pi}\right)}{\boldsymbol{u}^{\top}\widehat{\Sigma}^{-1}\widehat{\boldsymbol{\Omega}}(\widehat{\Sigma}^{\top})^{-1}\boldsymbol{u}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

□

Proof of Theorem 4.17.

Proof. $\widehat{V}_{\mathcal{P}}(\pi(\boldsymbol{\theta}_k)) = \int \widehat{V}(\pi(\boldsymbol{\theta}_k), \boldsymbol{x}) d\mathcal{R}(\boldsymbol{x}) = \int \boldsymbol{U}(\pi_{\boldsymbol{\theta}_k}, \boldsymbol{x})^{\top}\widehat{\boldsymbol{\beta}}^{\pi_{\boldsymbol{\theta}_k}} d\mathcal{P}(\boldsymbol{x})$.

$\boldsymbol{U}(\pi, \boldsymbol{x}) = \boldsymbol{U}_{\pi,i,t} = [\boldsymbol{\phi}(\boldsymbol{X}_{i,t})^{\top}\pi(1|\boldsymbol{X}_{i,t}), \cdots, \boldsymbol{\phi}(\boldsymbol{X}_{i,t})^{\top}\pi(M|\boldsymbol{X}_{i,t})]^{\top} \in \mathbb{R}^{JM}$

$$\begin{aligned}
\mathring{V}(\pi(\boldsymbol{\theta})) &= \sum_{j=1}^{M-1} \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)}{1+\sum_{j=1}^{M-1}\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)} \boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top \mathring{\boldsymbol{\beta}}_j^{\pi(\boldsymbol{\theta})} + \frac{1}{1+\sum_{j=1}^{M-1}\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)} \boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top \mathring{\boldsymbol{\beta}}_M^{\pi(\boldsymbol{\theta})} \\
\widehat{V}(\pi(\boldsymbol{\theta})) &= \sum_{j=1}^{M-1} \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)}{1+\sum_{j=1}^{M-1}\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)} \boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top \widehat{\boldsymbol{\beta}}_j^{\pi(\boldsymbol{\theta})} + \frac{1}{1+\sum_{j=1}^{M-1}\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)} \boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top \widehat{\boldsymbol{\beta}}_M^{\pi(\boldsymbol{\theta})}.
\end{aligned} \tag{38}$$

$$\begin{aligned}
\widehat{V}(\pi(\boldsymbol{\theta})) - \mathring{V}(\pi(\boldsymbol{\theta})) = &\sum_{j=1}^{M-1} \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)}{1+\sum_{j=1}^{M-1}\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)} \boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top \left(\widehat{\boldsymbol{\beta}}_j^{\pi(\boldsymbol{\theta})} - \mathring{\boldsymbol{\beta}}_j^{\pi(\boldsymbol{\theta})}\right) \\
&+ \frac{1}{1+\sum_{j=1}^{M-1}\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)} \boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top \left(\widehat{\boldsymbol{\beta}}_M^{\pi(\boldsymbol{\theta})} - \mathring{\boldsymbol{\beta}}_M^{\pi(\boldsymbol{\theta})}\right)
\end{aligned} \tag{39}$$

$$\begin{aligned}
\sup_{\boldsymbol{\theta}} \left|\widehat{V}(\pi(\boldsymbol{\theta})) - \mathring{V}(\pi(\boldsymbol{\theta}))\right| = &\sup_{\boldsymbol{\theta}} \sum_{j=1}^{M-1} \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)}{1+\sum_{j=1}^{M-1}\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)} \left\|\boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top\right\|_2 \sup_{\boldsymbol{\theta}} \left\|\widehat{\boldsymbol{\beta}}_j^{\pi(\boldsymbol{\theta})} - \mathring{\boldsymbol{\beta}}_j^{\pi(\boldsymbol{\theta})}\right\|_2 \\
&+ \sup_{\boldsymbol{\theta}} \frac{1}{1+\sum_{j=1}^{M-1}\exp(\boldsymbol{x}^\top \boldsymbol{\theta}_j)} \left\|\boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top\right\|_2 \sup_{\boldsymbol{\theta}} \left\|\widehat{\boldsymbol{\beta}}_M^{\pi(\boldsymbol{\theta})} - \mathring{\boldsymbol{\beta}}_M^{\pi(\boldsymbol{\theta})}\right\|_2 \\
&\to 0.
\end{aligned} \tag{40}$$

□

Appendix D Sub-homogeneous MDP

With the oracle knowledge of the true subgroups, the solution $\boldsymbol{\theta}_k^\pi$ for group k of (21) is equivalent to that of the quasi-likelihood estimating equation:

$$\boldsymbol{G}_k\left(\boldsymbol{\theta}_k^\pi\right) = \sum_{i\in\mathcal{G}_k}\sum_{t\in[T]} \boldsymbol{z}_{i,t}\left(r_{i,t} - (\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1})^\top \boldsymbol{\theta}_k^\pi\right) = 0. \tag{41}$$

We have $\mathbb{E}\left[\boldsymbol{G}_k\left(\mathring{\boldsymbol{\theta}}_k^\pi\right)\right] = 0$. Let

$$\widetilde{\Sigma}_k = \frac{1}{N_k T}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]} \boldsymbol{z}_{i,t}(\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1})^\top, \tag{42}$$

The oracle estimator $\widetilde{\theta}_k^\pi$ for each group k have the following decomposition:

$$\left(\widetilde{\theta}_k^\pi - \mathring{\theta}_k^\pi\right) = \widetilde{\Sigma}_k^{-1} G_k\left(\theta_k^\pi\right) = \widetilde{\Sigma}_k^{-1}\left(\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\left(\boldsymbol{z}_{i,t}(\varepsilon_{1,i,t}+\varepsilon_{2,i,t})\right)\right) = \widetilde{\Sigma}_k^{-1}\widetilde{\zeta}_{1k} + \widetilde{\Sigma}_k^{-1}\widetilde{\zeta}_{2k}, \tag{43}$$

where

$$\widetilde{\zeta}_{1k} = \frac{1}{N_k T}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{z}_{i,t}\varepsilon_{1,i,t}, \qquad \widetilde{\zeta}_{2k} = \frac{1}{N_k T}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{z}_{i,t}\varepsilon_{2,i,t}, \tag{44}$$

$$\varepsilon_{1,i,t} = r_{i,t} + \gamma\cdot\sum_{a\in\mathcal{A}} Q(\pi, X_{i,t+1}, a)\pi(a\mid \boldsymbol{X}_{i,t+1}) - Q(\pi, X_{i,t}, A_{i,t}), \tag{45}$$

$$\varepsilon_{2,i,t} = \gamma\cdot\sum_{a\in\mathcal{A}} e(\pi, \boldsymbol{X}_{i,t+1}, a)\pi(a\mid \boldsymbol{X}_{i,t+1}) - \sum_{a\in\mathcal{A}} e(\pi, \boldsymbol{X}_{i,t}, a)\mathbb{1}(A_{i,t}=a), \tag{46}$$

$$e(\pi, \boldsymbol{x}, a) = Q(\pi, \boldsymbol{x}, a) - \boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{\beta}_{a,i}^\pi. \tag{47}$$

The one step TD-error is the term $\varepsilon_{1,i,t}+\varepsilon_{2,i,t}$, of which $\varepsilon_{1,i,t} = r_{i,t} - \varepsilon_{2,i,t}$ is the random noise of the immediate rewards and $\varepsilon_{2,i,t}$ is the biases from approximating $Q(\pi, \boldsymbol{x}, a)$.

To derive the convergence rate of $\left(\widetilde{\theta}_k^\pi - \mathring{\theta}_k^\pi\right)$, we study the properties of $\widetilde{\Sigma}_k$, $\widetilde{\zeta}_{1k}$, and $\widetilde{\zeta}_{2k}$, respectively. Since data is not independent over T, we specifically consider two setting as follows.

(I) T is fixed and N_k goes to infinity. When T is fixed and N_k goes to infinity, we consider the concentration property of a random matrix $T^{-1}\sum_{t=0}^T \boldsymbol{z}_{i,t}\left(\boldsymbol{z}_{i,t} - \gamma\boldsymbol{u}_{\pi,i,t+1}\right)^\top$.

(II) T goes to infinity and N_k can be either fixed or go to infinity. When T goes to infinity, we consider $\boldsymbol{z}_{i,t}\left(\boldsymbol{z}_{i,t} - \gamma\boldsymbol{u}_{\pi,i,t+1}\right)^\top$ and apply the martingale concentration inequality under Assumption 4.6.

Although the techniques of proofs are different for these two settings, we obtain the same ℓ_2 and ℓ_∞ bounds as summarized in Proposition D.1, Theorem D.2, Corollary D.3 and Theorem D.4. Proofs of the results are presented in subsequent sections.

Proposition D.1 (ℓ_2 convergence of $\widetilde{\theta}_k^\pi$). *Suppose Assumption 4.2 – 4.6 hold. If $J^{\kappa/p} \gg \sqrt{N_k T}$ and $J \ll \sqrt{N_k T}/\log(N_k T)$, we have, as either $N\to\infty$ or $T\to\infty$,*

$$\left\|\widetilde{\theta}_k^\pi - \mathring{\theta}_k^\pi\right\|_2 = \mathcal{O}_p\left(J^{-\kappa/p}\right) + \mathcal{O}_p\left(J^{1/2}(N_k T)^{-1/2}\right).$$

Proof. Applying decomposition (43), Lemma (D.11) and (D.13), we have

$$\begin{aligned}\left\|\widetilde{\theta}_k^{\pi}-\mathring{\theta}_k^{\pi}\right\|_2 &= \left\|\widetilde{\Sigma}_k^{-1}\right\|_2\left\|\widetilde{\zeta}_{1k}\right\|_2+\left\|\widetilde{\Sigma}_k^{-1}\right\|_2\left\|\widetilde{\zeta}_{2k}\right\|_2\\ &= \mathcal{O}_p\left(J^{1/2}(N_kT)^{-1/2}\right)+\mathcal{O}_p\left(J^{-\kappa/p}\right).\end{aligned}$$

□

Theorem D.2 (Bidirectional asymptotic of $\widetilde{\theta}_k^{\pi}$)**.** *Suppose Assumption 4.2 – 4.6 hold. If $J \ll \sqrt{N_kT}/\log(N_kT)$, $J^{\kappa/p} \gg \sqrt{N_kT}$, as either $N_k \to \infty$ or $T \to \infty$, we have for any vector $\boldsymbol{\nu} \in \mathbb{R}^{JM}$*

$$\sqrt{N_kT}\cdot\widetilde{\sigma}_k^{-1}(\pi,\boldsymbol{\nu})\cdot\boldsymbol{\nu}^{\top}\left(\widetilde{\theta}_k^{\pi}-\mathring{\theta}_k^{\pi}\right)\xrightarrow{\mathcal{D}}\mathcal{N}(0,1),$$

where $\widetilde{\sigma}_k(\pi,\boldsymbol{\nu})=\boldsymbol{\nu}^{\top}\widetilde{\Sigma}_k\widetilde{\boldsymbol{\Omega}}_k(\widetilde{\Sigma}_k^{\top})^{-1}\boldsymbol{\nu}$, $\widetilde{\Sigma}_k$ is given in (42) and

$$\widetilde{\boldsymbol{\Omega}}_k=(N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^{\top}\left(r_{i,t}-(\boldsymbol{z}_{i,t}-\gamma\boldsymbol{u}_{\pi,i,t+1})^{\top}\widetilde{\theta}_k^{\pi}\right)^2. \tag{48}$$

Proof.

Step I. Let $\boldsymbol{\nu}_k \in \mathbb{R}^{JM}$ be any vector with $\|\boldsymbol{\nu}_k\|_2$ bounded away from zero, we have

$$\begin{aligned}\left|\boldsymbol{\nu}^{\top}\left(\widetilde{\theta}_k^{\pi}-\mathring{\theta}_k^{\pi}\right)-\boldsymbol{\nu}_k^{\top}\Sigma_k^{-1}\widetilde{\zeta}_{k1}\right| &\le \left|\boldsymbol{\nu}_k^{\top}\left(\widetilde{\theta}_k^{\pi}-\mathring{\theta}_k^{\pi}-\Sigma_k^{-1}\widetilde{\zeta}_{k1}\right)\right|\\ &\le \|\boldsymbol{\nu}_k\|_2\left\|\widetilde{\theta}_k^{\pi}-\mathring{\theta}_k^{\pi}-\Sigma_k^{-1}\widetilde{\zeta}_{k1}\right\|_2.\end{aligned}$$

Let

$$\sigma_k^2(\pi,\boldsymbol{\nu})=\boldsymbol{\nu}^{\top}\Sigma_k^{-1}\boldsymbol{\Omega}_k\Sigma_k^{\top-1}\boldsymbol{\nu}.$$

By Lemma D.11 and D.13, thus we have

$$\sigma_k^2(\pi,\boldsymbol{\nu})\ge 3^{-1}c_0^{-1}c_1C\|\boldsymbol{\nu}\|_2^2.$$

Further, it holds that

$$\begin{aligned}\sigma_k^{-1}(\pi,\boldsymbol{\nu})\left|\boldsymbol{\nu}^{\top}\left(\widetilde{\theta}_k^{\pi}-\mathring{\theta}_k^{\pi}\right)-\boldsymbol{\nu}_k^{\top}\Sigma_k^{-1}\widetilde{\zeta}_{k1}\right| &\le \sqrt{3c_0c_1^{-1}C^{-1}}\left\|\widetilde{\theta}_k^{\pi}-\mathring{\theta}_k^{\pi}-\Sigma_k^{-1}\widetilde{\zeta}_{k1}\right\|_2\\ &= \mathcal{O}_p\left(J^{-\kappa/p}\right)+\mathcal{O}_p\left(J(N_kT)^{-1}\log(N_kT)\right).\end{aligned}$$

Under the condition that $J \ll \sqrt{N_kT}\log(N_kT)^{-1}$ and $J^{\kappa/p} \gg \sqrt{N_kT}$, we have

$$\sqrt{N_kT}\cdot\sigma_k^{-1}(\pi,\boldsymbol{\nu})\,\boldsymbol{\nu}^{\top}\left(\widetilde{\theta}_k^{\pi}-\mathring{\theta}_k^{\pi}\right)=\sqrt{N_kT}\cdot\sigma_k^{-1}(\pi)\boldsymbol{\nu}^{\top}\Sigma_k^{-1}\widetilde{\zeta}_{k1}+\mathrm{o}_{\mathrm{p}}(1). \tag{49}$$

Step II. We have

$$\sqrt{N_k T}\cdot\sigma_k^{-1}(\pi,\boldsymbol{\nu})\boldsymbol{\nu}^\top\Sigma_k^{-1}\widetilde{\zeta}_{k1}=\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\frac{\boldsymbol{\nu}^\top\Sigma_k^{-1}\boldsymbol{z}_{i,t}\varepsilon_{1,i,t}}{\sqrt{N_k T}\sigma_k(\pi,\boldsymbol{\nu})}\xrightarrow{\mathcal{D}}\mathcal{N}(0,1). \tag{50}$$

This can be shown by applying a martingale central limit theory for triangular arrays (Corollary 2.8 of McLeish, 1974). The construction of the martingale is the same as that in Shi et al. (2020). The following two conditions

(a) $\max_{i\in\mathcal{G}_k,t\in[T]}\left|\frac{\boldsymbol{\nu}^\top\Sigma_k^{-1}\boldsymbol{z}_{i,t}\varepsilon_{1,i,t}}{\sqrt{N_k T}\sigma_k(\pi,\boldsymbol{\nu})}\right|\xrightarrow{\mathcal{P}}0.$

(b) $\frac{\sigma_k(\pi,\boldsymbol{\nu})^{-2}}{\sqrt{N_k T}}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{\nu}^\top\Sigma_k^{-1}\boldsymbol{z}_{i,t}\varepsilon_{1,i,t}\xrightarrow{\mathcal{P}}1.$

can be checked by applying Lemma D.11 and D.14, the argument is thus omitted here.

□

Corollary D.3 (Bidirectional asymptotic of $\widetilde{V}_k(\pi,\boldsymbol{x})$)**.** *Suppose Assumption 4.2 – 4.6 hold. If* $J\ll\sqrt{N_k T}/\log(N_k T)$, $J^{-\kappa/p}\ll\left(N_k T\left(1+\|\boldsymbol{U}(\pi,\boldsymbol{x})\|_2^2\right)\right)^{-1/2}$, *as either* $N\to\infty$ *or* $T\to\infty$, *we have*

$$\sqrt{N_k T}\cdot\widetilde{\sigma}_k^{-1}(\pi)\left(\widetilde{V}_k(\pi,\boldsymbol{x})-V_k(\pi,\boldsymbol{x})\right)\xrightarrow{\mathcal{D}}\mathcal{N}(0,1).$$

Here $\widetilde{\sigma}_k^{-1}(\pi,\boldsymbol{x})=\boldsymbol{U}(\pi,\boldsymbol{x})^\top\widetilde{\Sigma}_k\widetilde{\boldsymbol{\Omega}}_k(\widetilde{\Sigma}_k^\top)^{-1}\boldsymbol{U}(\pi,\boldsymbol{x})$, $\widetilde{\Sigma}_k$ *and* $\widetilde{\Sigma}_k$ *are given in* (42) *and* (48).

Proof. First note that

$$\begin{aligned}\left|\widetilde{V}_k(\pi,\boldsymbol{x})-V_k(\pi,\boldsymbol{x})-\boldsymbol{\nu}_k^\top\Sigma_k^{-1}\widetilde{\zeta}_{k1}\right|&\le\left|\boldsymbol{U}(\pi,\boldsymbol{x})^\top\left(\widetilde{\boldsymbol{\theta}}_k^\pi-\boldsymbol{\theta}_k^\pi-\Sigma_k^{-1}\widetilde{\zeta}_{k1}\right)\right|\\&\quad+\left|\boldsymbol{U}(\pi,\boldsymbol{x})^\top\boldsymbol{\theta}_k^\pi-V_k(\pi,\boldsymbol{x})\right|\\&\le\left\|\boldsymbol{U}(\pi,\boldsymbol{x})^\top\right\|_2\left\|\widetilde{\boldsymbol{\theta}}_k^\pi-\boldsymbol{\theta}_k^\pi-\Sigma_k^{-1}\widetilde{\zeta}_{k1}\right\|_2+CJ^{-\kappa/p}.\end{aligned}$$

Under the condition that $J\ll\sqrt{N_k T}\log(N_k T)^{-1}$, $J^{-\kappa/p}\ll\sqrt{N_k T\left(1+\|\boldsymbol{U}(\pi,\boldsymbol{x})\|_2^2\right)}$, we obtain

$$\sqrt{N_k T}\cdot\sigma_k^{-1}(\pi)\left(\widetilde{V}_k(\pi,\boldsymbol{x})-V_k(\pi,\boldsymbol{x})\right)=\sqrt{N_k T}\cdot\sigma_k^{-1}(\pi)\boldsymbol{U}(\pi,\boldsymbol{x})^\top\Sigma_k^{-1}\widetilde{\zeta}_{k1}+\mathrm{o}_\mathrm{p}(1).$$

Applying Theorem D.2 by setting $\boldsymbol{\nu}_k=\boldsymbol{U}(\pi,\boldsymbol{x})$, we obtain the desired result. □

Theorem D.4 (Uniform convergence of $\widetilde{\boldsymbol{\theta}}_k$)**.** *Suppose Assumption 4.2 – 4.6 hold. If* $J\ll\sqrt{N_k T}/\log(N_k T)$, $J^{\kappa/p}\gg\sqrt{N_k T}$, *There exists some positive constant* C *and* C_1 *that*

$$\left\|\widetilde{\boldsymbol{\theta}}_k^\pi-\mathring{\boldsymbol{\theta}}_k^\pi\right\|_\infty\le C_1C^{-1}\sqrt{J\cdot\frac{\log(N_k T)}{N_k T}},$$

holds with probability at least $1 - 2JMK(N_kT)^{-2}$.

Proof. By decomposition (43), we have

$$\left\|\widetilde{\theta}_k^{\pi} - \mathring{\theta}_k^{\pi}\right\|_\infty \le \left\|\left(\widetilde{\Sigma}_k\right)^{-1}\right\|_\infty \left\|\widetilde{\zeta}_{1k}\right\|_\infty + \left\|\left(\widetilde{\Sigma}_k\right)^{-1}\right\|_\infty \left\|\widetilde{\zeta}_{2k}\right\|_\infty$$

We bound each term in the right hand side as follows. By Lemma D.11 (iii), we have with probability at least $1 - \mathcal{O}\left((N_kT)^{-2}\right)$,

$$\left\|\widetilde{\Sigma}_k^{-1}\right\|_\infty \le \sqrt{J}\left\|\widetilde{\Sigma}_k^{-1}\right\|_2 \le 6C^{-1}\sqrt{J}.$$

By Lemma D.13 (iii), we have

$$\Pr\left(\|\zeta_{1k}\|_\infty > c\sqrt{2\log(N_kT)/(N_kT)}\right) \le 2JM(N_kT)^{-2}.$$

for some positive constant c. Lemma D.13 (iv) shows that $\|\zeta_{2k}\|_\infty \le c_1 J^{-\kappa/p} \ll c_1 (N_kT)^{-1/2}$ almost surely.

Thus, by union bounds, we have

$$\left\|\widetilde{\theta}_k^{\pi} - \mathring{\theta}_k^{\pi}\right\|_\infty \le C_1 C^{-1}\sqrt{J \cdot \frac{\log(N_kT)}{N_kT}},$$

holds with probability at least $1 - 2JMK(N_kT)^{-2}$. □

D.1 Bounds on terms in TD error

The results in this section applies to the setting as either $N \to \infty$ or $T \to \infty$.

Lemma D.5. *User Assumption 4.2, there exist some $c > 0$ such that, for any given π and $a \in \mathcal{A}$, $Q(\pi; \boldsymbol{x}, a)$ as a function of $\boldsymbol{x}$ belongs to $\mathcal{H}(\kappa, c)$.*

Proof. See the proof of Lemma 1 in Shi et al. (2020). □

Lemma D.6. *Under Assumption 4.3, there exists some constant $c \ge 1$ such that*

$$c^{-1} \le \lambda_{\min}\left(\int_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{\phi}_J(\boldsymbol{x})\boldsymbol{\phi}_J(\boldsymbol{x})^\top d\boldsymbol{x}\right) \le \lambda_{\max}\left(\int_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{\phi}_J(\boldsymbol{x})\boldsymbol{\phi}_J(\boldsymbol{x})^\top d\boldsymbol{x}\right) \le c,$$

and $\sup_{\boldsymbol{x}\in\mathcal{X}} \left\|\phi_J(\boldsymbol{x})\right\|_2 \le c\sqrt{J}$.

Proof. See Shi et al. (2020) Lemma 2. For the B-spline basis, the first assertion follows from the arguments used in the proof of Theorem 3.3 of Burman et al. (1989). For wavelet basis, the first assertion follows from the arguments used in the proof of Theorem 5.1 of Chen and Christensen (2015).

For either B-spline or wavelet sieve and any $J \geq 1$, $\boldsymbol{x} \in \mathcal{X}$, the number of nonzero elements in the vector $\boldsymbol{\phi}(\boldsymbol{x})$ is bounded by some constant. Moreover, each of the basis function is uniformly bounded by $\mathcal{O}\left(\sqrt{J}\right)$. This proves the second assertion. □

Lemma D.7. *Under Assumption 4.3 and 4.4, we have*

(i) $\max_{i\in[N],t\in[T]} \left\|\boldsymbol{z}_{i,t}\right\|_2 \leq c\sqrt{J}$

(ii) $\max_{i\in[N],t\in[T]} \left\|\boldsymbol{U}_{i,t}\right\|_2 \leq c\sqrt{J}$

(iii) $\left\|\mathbb{E}\left[\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\right]\right\|_2^2 = \mathcal{O}(1)$ *and* $\left\|\mathbb{E}\left[\boldsymbol{z}_{i,t}^\top\boldsymbol{z}_{i,t}\right]\right\|_2^2 = \mathcal{O}(1)$

Proof. Recall that

$$\begin{aligned}
\boldsymbol{z}_{i,t} &= \left[\boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top \mathbb{1}(A_{i,t}=1), \cdots, \boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top \mathbb{1}(A_{i,t}=M)\right]^\top \in \mathbb{R}^{JM}, \\
\boldsymbol{U}_{\pi,i,t+1} &= \left[\boldsymbol{\phi}(\boldsymbol{X}_{i,t+1})^\top \pi(1|\boldsymbol{X}_{i,t+1}), \cdots, \boldsymbol{\phi}(\boldsymbol{X}_{i,t+1})^\top \pi(M|\boldsymbol{X}_{i,t+1})\right]^\top \in \mathbb{R}^{JM}.
\end{aligned}$$

Then, we have

(i) $\max_{i\in[N],t\in[T]} \left\|\boldsymbol{z}_{i,t}\right\|_2 \leq \sup_{\boldsymbol{x}} \left\|\boldsymbol{\phi}_J(\boldsymbol{x})\right\|_2 \leq c\sqrt{J}$

(ii) $\max_{i\in[N],t\in[T]} \left\|\boldsymbol{U}_{i,t}\right\|_2 \leq \sup_{\boldsymbol{x}} \left\|\boldsymbol{\phi}_J(\boldsymbol{x})\right\|_2 \leq c\sqrt{J}$

(iii) Let $\boldsymbol{\nu} = \left[\boldsymbol{\nu}_1^\top, \cdots, \boldsymbol{\nu}_M^\top\right]^\top$ where sub-vector $\boldsymbol{\nu}_m \in \mathbb{R}^J$ for $1 \leq m \leq M$.

$$\begin{aligned}
\boldsymbol{\nu}^\top \mathbb{E}\left[\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\right]\boldsymbol{\nu} &= \sum_{m=1}^M \boldsymbol{\nu}_m^\top \mathbb{E}\left[\boldsymbol{\phi}(\boldsymbol{X}_{i,t})\boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top \mathbb{1}(A_{i,t}=1)\right]\boldsymbol{\nu}_m \\
&= \sum_{m=1}^M \boldsymbol{\nu}_m^\top \mathbb{E}\left[\boldsymbol{\phi}(\boldsymbol{X}_{i,t})\boldsymbol{\phi}(\boldsymbol{X}_{i,t})^\top\right]\boldsymbol{\nu}_m b(m|\boldsymbol{X}_{i,t}) \\
&\leq \lambda_{\max}\left(\int_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top \mu_0(\boldsymbol{x})d\boldsymbol{x}\right)\|\boldsymbol{\nu}\|_2^2
\end{aligned}$$

By Assumption 4.4 and Lemma D.6, we obtain

$$\lambda_{\max}\left(\int_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top \mu_0(\boldsymbol{x})d\boldsymbol{x}\right) \leq \sup_{\boldsymbol{x}\in\mathcal{X}}\mu_0(\boldsymbol{x})\ \lambda_{\max}\left(\int_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top d\boldsymbol{x}\right) = \mathcal{O}(1).$$

Thus, we have

$$\left\|\mathbb{E}\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\right\|_2^2 = \mathcal{O}(1)$$

Using similar arguments, we have $\left\|\mathbb{E}\boldsymbol{z}_{i,t}^\top\boldsymbol{z}_{i,t}\right\|_2^2 = \mathcal{O}(1)$.

□

Lemma D.8. *Under Assumption 4.2 and 4.3, there exists some positive constants c_1, c_2, c_3, such that*

(i) $\max_{i\in\mathcal{G}_k,1\le t\le T}\left|\varepsilon_{1,i,t}\right| \le c_1 + 2c_2$

(ii) $\max_{i\in\mathcal{G}_k,1\le t\le T}\left|\varepsilon_{2,i,t}\right| \le 2c_3 J^{-\kappa/p}$

Proof. Recall that we define, for $\forall i \in \mathcal{G}_k$, $1 \le k \le K$,

$$\begin{aligned}
\varepsilon_{1,i,t} &= r_{i,t} + \gamma\cdot\sum_{a\in\mathcal{A}} Q_k(\pi, X_{i,t+1}, a)\pi(a\mid \boldsymbol{X}_{i,t+1}) - Q_k(\pi, X_{i,t}, A_{i,t})\\
\varepsilon_{2,i,t} &= \gamma\cdot\sum_{a\in\mathcal{A}} e_k(\pi, \boldsymbol{X}_{i,t+1}, a)\pi(a\mid \boldsymbol{X}_{i,t+1}) - \sum_{a\in\mathcal{A}} e_k(\pi, \boldsymbol{X}_{i,t}, a)\mathbb{1}(A_{i,t}=a),\\
e_k(\pi,\boldsymbol{x},a) &= Q_k(\pi,\boldsymbol{x},a) - \boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{\theta}_{a,k}^\pi.
\end{aligned}$$

(i) The condition that $\Pr\left(\max_{0\le t\le T}\left|r_{i,t}\right|\le c_1\right) = 1$ implies that $\left|r_{i,t}\right| \le c_1$ for $\forall i,t$, almost surely. By Lemma D.5 and the definition of k-smooth function, we obtain that $|Q(\pi,\boldsymbol{x},a)| \le c_2$ for any $\pi,\boldsymbol{x},a$. Thus, we have

$$\max_{i\in\mathcal{G}_k,1\le t\le T}\left|\varepsilon_{1,i,t}\right| \le c_1 + (1+\gamma)c_2 \le c_1 + 2c_2, \quad \text{almost surely.} \tag{51}$$

(ii) By Lemma D.5 and Assumption 4.2, there exist a set of vector $\boldsymbol{\theta}_{a,k}^\pi$ that satisfy (see Section 2.2 of Huang et al. (1998) for details)

$$\sup_{\boldsymbol{x}\in\mathcal{X},a\in\mathcal{A}}|e_k(\pi,\boldsymbol{x},a)| = \sup_{\boldsymbol{x}\in\mathcal{X},a\in\mathcal{A}}\left|Q_k(\pi,\boldsymbol{x},a) - \boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{\theta}_{a,k}^\pi\right| \le c_3 J^{-\kappa/p},$$

for some positive constant c_3. Thus, we have

$$\begin{aligned}
\sup_{i\in\mathcal{G}_k,1\le t\le T}\left|\varepsilon_{2,i,t}\right| &= \sup_{1\le i\le N,1\le t\le T}\left|\gamma\cdot\sum_{a\in\mathcal{A}} e_k(\pi,\boldsymbol{X}_{i,t+1},a)\pi(a\mid\boldsymbol{X}_{i,t+1}) - \sum_{a\in\mathcal{A}} e_k(\pi,\boldsymbol{X}_{i,t},a)\mathbb{1}(A_{i,t}=a)\right|,\\
&\le (1+\gamma)\sup_{\boldsymbol{x}\in\mathcal{X},a\in\mathcal{A}}|e_k(\pi,\boldsymbol{x},a)| = 2c_3J^{-\kappa/p}.
\end{aligned}$$

□

Define $\widetilde{e}(\pi, \boldsymbol{x}, a) = Q(\pi, \boldsymbol{x}, a) - \boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\beta}}_{a,i}^{\pi}$ and

$$\widetilde{\varepsilon}_{2,i,t} = \gamma \cdot \sum_{a \in \mathcal{A}} \widetilde{e}(\pi, \boldsymbol{X}_{i,t+1}, a) \pi(a \mid \boldsymbol{X}_{i,t+1}) - \sum_{a \in \mathcal{A}} \widetilde{e}(\pi, \boldsymbol{X}_{i,t}, a) \mathbb{1}(A_{i,t} = a).$$

The following lemma is needed for the proof of asymptotic normality, that is, Theorem D.2 and Corollary D.3. Its proof uses results in Proposition D.1.

Lemma D.9. *Under Assumption 4.3, we have*

(i) $\max_{i \in \mathcal{G}_k, 1 \le t \le T} \left| 2\varepsilon_{1,i,t} + \widetilde{\varepsilon}_{2,i,t} \right| = \mathcal{O}_p(1)$

(ii) $\max_{i \in \mathcal{G}_k, 1 \le t \le T} \left| \widetilde{\varepsilon}_{2,i,t} \right| = \mathrm{o}_p(1)$

(iii) $\max_{i \in \mathcal{G}_k, 1 \le t \le T} \left| \varepsilon_{1,i,t}^2 - (\varepsilon_{1,i,t} + \widetilde{\varepsilon}_{2,i,t})^2 \right| = \mathrm{o}_p(1)$

Proof. (i) The result follows trivially from Lemma D.8.

(ii) First, we have

$$\begin{aligned}
\sup_{\boldsymbol{x} \in \mathcal{X}, a \in \mathcal{A}} |\widetilde{e}(\pi, \boldsymbol{x}, a)| &\le \sup_{\boldsymbol{x} \in \mathcal{X}, a \in \mathcal{A}} \left| Q(\pi, \boldsymbol{x}, a) - \boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\beta}}_{a,i}^{\pi} \right| \\
&\le \sup_{\boldsymbol{x} \in \mathcal{X}, a \in \mathcal{A}} \left| Q(\pi, \boldsymbol{x}, a) - \boldsymbol{\phi}(\boldsymbol{x})^\top \mathring{\boldsymbol{\beta}}_{a,i}^{\pi} \right| + \sup_{\boldsymbol{x} \in \mathcal{X}, a \in \mathcal{A}} \left| \boldsymbol{\phi}(\boldsymbol{x})^\top \mathring{\boldsymbol{\beta}}_{a,i}^{\pi} - \boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\beta}}_{a,i}^{\pi} \right| \\
&\le C J^{-\kappa/p} + \sup_{\boldsymbol{x} \in \mathcal{X}, a \in \mathcal{A}} \left\| \boldsymbol{\phi}(\boldsymbol{x}) \right\|_2 \sup_{a \in \mathcal{A}} \left\| \mathring{\boldsymbol{\beta}}_{a,i}^{\pi} - \widetilde{\boldsymbol{\beta}}_{a,i}^{\pi} \right\|_2 \\
&= \mathcal{O}_p\left(J^{1/2-\kappa/p} \right) + \mathcal{O}_p\left(J (N_k T)^{-1/2} \right).
\end{aligned}$$

Thus, we obtain that

$$\begin{aligned}
\max_{i \in \mathcal{G}_k, 1 \le t \le T} \left| \widetilde{\varepsilon}_{2,i,t} \right| &\le \max_{i \in \mathcal{G}_k, 1 \le t \le T} \left| \gamma \cdot \sum_{a \in \mathcal{A}} \widetilde{e}(\pi, \boldsymbol{X}_{i,t+1}, a) \pi(a \mid \boldsymbol{X}_{i,t+1}) - \sum_{a \in \mathcal{A}} \widetilde{e}(\pi, \boldsymbol{X}_{i,t}, a) \mathbb{1}(A_{i,t} = a) \right| \\
&\le (1 + \gamma) \sup_{\boldsymbol{x} \in \mathcal{X}, a \in \mathcal{A}} |\widetilde{e}(\pi, \boldsymbol{x}, a)| \\
&= \mathcal{O}_p\left(J^{1/2-\kappa/p} \right) + \mathcal{O}_p\left(J (N_k T)^{-1/2} \right).
\end{aligned}$$

Under the condition that $J^{-\kappa/p} \ll (N_k T)^{-1/2}$ and $J \ll (N_k T)^{1/2} \log(N_k T)^{-1}$, we have $\mathcal{O}_p\left(J^{1/2-\kappa/p} \right) = \mathrm{o}_p(1)$ and $\mathcal{O}_p\left(J (N_k T)^{-1/2} \right) = \mathrm{o}_p(1)$. Therefore, we have $\max_{i \in \mathcal{G}_k, 1 \le t \le T} \left| \widetilde{\varepsilon}_{2,i,t} \right| = \mathrm{o}_p(1)$.

(iii) $\max_{i\in\mathcal{G}_k,1\le t\le T}\left|\varepsilon_{1,i,t}^2-(\varepsilon_{1,i,t}+\widetilde{\varepsilon}_{2,i,t})^2\right|\le\max_{i\in\mathcal{G}_k,1\le t\le T}\left|2\varepsilon_{1,i,t}+\widetilde{\varepsilon}_{2,i,t}\right|\max_{i\in\mathcal{G}_k,1\le t\le T}\left|\widetilde{\varepsilon}_{2,i,t}\right|=o_p(1)$

□

D.2 Setting I: T is bounded and $N_k\to\infty$

In this section, we study the properties of $\widetilde{\Sigma}_k$, $\widetilde{\Sigma}_k$, $\widetilde{\zeta}_{1k}$, and $\widetilde{\zeta}_{2k}$ under the setting that T is bounded and N_k goes to infinity. We apply concentration inequality over N on an average of T random quantities.

D.2.1 Properties of $\widetilde{\Sigma}_k$

Since T is bounded, we define the random matrix

$$\boldsymbol{M}_i=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}_{i,t}(\boldsymbol{z}_{i,t}-\gamma\boldsymbol{u}_{\pi,i,t+1})^\top. \tag{52}$$

Then, we have

$$\widetilde{\Sigma}_k=\frac{1}{N_k}\sum_{i\in\mathcal{G}_k}\boldsymbol{M}_i,\quad \mathbb{E}\left[\widetilde{\Sigma}_k\right]=\mathbb{E}\widetilde{\Sigma}_k=\mathbb{E}\boldsymbol{M}_i,\quad\text{and}\quad \widetilde{\Sigma}_k-\mathbb{E}\left[\widetilde{\Sigma}_k\right]=\frac{1}{N_k}\sum_{i\in\mathcal{G}_k}(\boldsymbol{M}_i-\mathbb{E}\boldsymbol{M}_i).$$

Lemma D.10. *Denote*

$$\sigma_N^2=\max\left(\left\|\sum_{i=1}^{N_k}\mathbb{E}\left[(\boldsymbol{M}_i-\mathbb{E}\boldsymbol{M}_i)(\boldsymbol{M}_i-\mathbb{E}\boldsymbol{M}_i)^\top\right]\right\|_2,\left\|\sum_{i=1}^{N_k}\mathbb{E}\left[(\boldsymbol{M}_i-\mathbb{E}\boldsymbol{M}_i)^\top(\boldsymbol{M}_i-\mathbb{E}\boldsymbol{M}_i)\right]\right\|_2\right)$$

There exists some constants c_1 and c_2 such that,

(i) $\max_{1\le i\le N}\|\boldsymbol{M}_i-\mathbb{E}\boldsymbol{M}_i\|_2\le 4c_1^2J$

(ii) $\sigma_N^2\le c_2JN$

Proof. (i) It follows from Lemma D.7 that

$$\begin{aligned}\max_{1\le i\le N}\|\boldsymbol{M}_i-\mathbb{E}\boldsymbol{M}_i\|_2&\le 2\max_{1\le i\le N}\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}_{i,t}(\boldsymbol{z}_{i,t}-\gamma\boldsymbol{u}_{\pi,i,t+1})^\top\right\|_2\\&\le\frac{2}{T}\sum_{t=1}^{T}c_1\sqrt{J}(c_1\sqrt{J}+\gamma c_1\sqrt{J})\\&\le 2c_1^2(1+\gamma)J\le 4c_1^2J\end{aligned}$$

(ii) For any $\boldsymbol{\nu} \in \mathbb{R}^{JM}$, we have

$$\boldsymbol{\nu}^\top \mathbb{E}\Big[(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)^\top\Big]\boldsymbol{\nu} = \boldsymbol{\nu}^\top \mathbb{E}\Big[\boldsymbol{M}_i \boldsymbol{M}_i^\top\Big]\boldsymbol{\nu} - \boldsymbol{\nu}^\top \mathbb{E}\Big[\mathbb{E}\boldsymbol{M}_i \mathbb{E}\boldsymbol{M}_i^\top\Big]\boldsymbol{\nu} \le \boldsymbol{\nu}^\top \mathbb{E}\Big[\boldsymbol{M}_i \boldsymbol{M}_i^\top\Big]\boldsymbol{\nu}.$$

We have

$$\begin{aligned}
\boldsymbol{\nu}^\top \mathbb{E}\Big[\boldsymbol{M}_i \boldsymbol{M}_i^\top\Big]\boldsymbol{\nu} = \mathbb{E}\big\|\boldsymbol{\nu}^\top \boldsymbol{M}_i\big\|_2^2 &= \mathbb{E}\left\|\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\nu}^\top \boldsymbol{z}_{i,t}\left(\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1}\right)^\top\right\|_2^2 \\
&\le \max_{i\in[N],t\in[T]} \big\|\boldsymbol{z}_{i,t} - \gamma \boldsymbol{U}_{\pi,i,t+1}\big\|_2^2 \, \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\nu}^\top \boldsymbol{z}_{i,t}\right)^2 \\
&\le 4c_1^2 J \cdot \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left(\boldsymbol{\nu}^\top \boldsymbol{z}_{i,t}\right)^2 \\
&\le 4c_1^2 J \cdot \|\boldsymbol{\nu}\|_2^2 \left\|\mathbb{E}\Big[\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\Big]\right\|_2,
\end{aligned}$$

where the second inequality uses Lemma D.7 (i) and (ii). Similarly, we can show that

$$\boldsymbol{\nu}^\top \mathbb{E}\Big[\boldsymbol{M}_i^\top \boldsymbol{M}_i\Big]\boldsymbol{\nu} \le 4c_1^2 J \cdot \|\boldsymbol{\nu}\|_2^2 \left\|\mathbb{E}\Big[\boldsymbol{z}_{i,t}^\top \boldsymbol{z}_{i,t}\Big]\right\|_2,$$

Lemma D.7 (iii) establishes that $\left\|\mathbb{E}\Big[\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\Big]\right\|_2^2 = \mathcal{O}(1)$ and $\left\|\mathbb{E}\Big[\boldsymbol{z}_{i,t}^\top \boldsymbol{z}_{i,t}\Big]\right\|_2^2 = \mathcal{O}(1)$, thus we obtain

$$\begin{aligned}
\sigma_N^2 &\le N_k \max\left(\left\|\mathbb{E}\Big[(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)^\top\Big]\right\|_2, \left\|\mathbb{E}\Big[(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)^\top(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)\Big]\right\|_2\right) \\
&\le c_2 J N_k,
\end{aligned}$$

for some constant $c_2 > 0$.

□

Lemma D.11. *Under Assumption 4.2, 4.3, 4.4 and 4.5, $J \ll (N_k T)/\log(N_k T)$, for any $1 \le k \le K$, we have as T fixed and $N_k \to \infty$,*

(i) $\left\|\mathbb{E}\Big[\widetilde{\Sigma}_k\Big]\right\|_2 \ge 2^{-1}C_1$ *and* $\left\|\mathbb{E}\Big[\widetilde{\Sigma}_k\Big]^{-1}\right\|_2 \le 2C_1^{-1}$,

(ii) $\left\|\widetilde{\Sigma}_k - \mathbb{E}\Big[\widetilde{\Sigma}_k\Big]\right\|_2 \le c_2\sqrt{J(N_k T)^{-1}\log(N_k T)}$ *with probability at least* $1 - \mathcal{O}\Big((N_k T)^{-2}\Big)$,

(iii) $\left\|\widetilde{\Sigma}_k\right\|_2 \le 6C_1^{-1}$ *with probability at least* $1 - \mathcal{O}\Big((N_k T)^{-2}\Big)$,

(iv) $\left\|\widetilde{\Sigma}_k^{-1} - \mathbb{E}\Big[\widetilde{\Sigma}_k\Big]^{-1}\right\| = \mathcal{O}_p\Big(\sqrt{J(N_k T)^{-1}\log(N_k T)}\Big)$,

where C_1 is specified in Assumption 4.5 and c_1 and c_2 are some positive constant.

Proof. Recall that

$$\widetilde{\Sigma}_k = (N_k T)^{-1} \sum_{i \in \mathcal{G}_k} \sum_{t=1}^{T} \boldsymbol{z}_{i,t} (\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1})^\top,$$

$$\mathbb{E}\left[\widetilde{\Sigma}_k\right] = \mathbb{E}\left[\boldsymbol{M}_i\right] = \mathbb{E}\left[T^{-1} \sum_{t=1}^{T} \boldsymbol{z}_{i,t} (\boldsymbol{z}_{i,t} - \gamma \boldsymbol{u}_{\pi,i,t+1})^\top\right].$$

(i) We first show that, under Assumption 4.5, for any $\boldsymbol{\nu} \in \mathbb{R}^{JM}$, $\boldsymbol{\nu}^\top \mathbb{E}\left[\widetilde{\Sigma}_k\right] \boldsymbol{\nu} \geq 2^{-1} C_1 \|\boldsymbol{\nu}\|_2^2$ and $\left\| \mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1} \right\|_2 \leq 2 C_1^{-1}$.

Let $\boldsymbol{\nu} \in \mathbb{R}^{JM}$ be arbitrary and note that, by Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\boldsymbol{\nu}^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_{i,t} \boldsymbol{U}_{\pi,i,t+1}^\top\right) \boldsymbol{\nu}\right] \leq \sqrt{\mathbb{E}\left[\boldsymbol{\nu}^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_{i,t} \boldsymbol{z}_{i,t}^\top\right) \boldsymbol{\nu}\right]} \cdot \sqrt{\mathbb{E}\left[\boldsymbol{\nu}^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{U}_{\pi,i,t+1} \boldsymbol{U}_{\pi,i,t+1}^\top\right) \boldsymbol{\nu}\right]}$$

This implies that

$$\begin{aligned}
\boldsymbol{\nu}^\top \mathbb{E}\left[\widetilde{\Sigma}_k\right] \boldsymbol{\nu} &\geq \mathbb{E}\left[\boldsymbol{\nu}^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_{i,t} \boldsymbol{z}_{i,t}^\top\right) \boldsymbol{\nu}\right] \\
&\quad - \gamma \cdot \sqrt{\mathbb{E}\left[\boldsymbol{\nu}^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_{i,t} \boldsymbol{z}_{i,t}^\top\right) \boldsymbol{\nu}\right]} \cdot \sqrt{\mathbb{E}\left[\boldsymbol{\nu}^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{U}_{\pi,i,t+1} \boldsymbol{U}_{\pi,i,t+1}^\top\right) \boldsymbol{\nu}\right]} \\
&= \boldsymbol{A}^{1/2} (\boldsymbol{A}^{1/2} - \gamma \boldsymbol{B}^{1/2}) \\
&= \frac{\boldsymbol{A}^{1/2} (\boldsymbol{A} - \gamma^2 \boldsymbol{B})}{(\boldsymbol{A}^{1/2} + \gamma \boldsymbol{B}^{1/2})},
\end{aligned}$$

where $\boldsymbol{A} = \mathbb{E}\left[\boldsymbol{\nu}^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_{i,t} \boldsymbol{z}_{i,t}^\top\right) \boldsymbol{\nu}\right]$ and $\boldsymbol{B} = \mathbb{E}\left[\boldsymbol{\nu}^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{U}_{\pi,i,t+1} \boldsymbol{U}_{\pi,i,t+1}^\top\right) \boldsymbol{\nu}\right]$. By Lemma D.6, we have $\mathbb{E}\left[\left\|\boldsymbol{U}_{i,t+1}\right\|_2^2\right] \leq c_2$ for some positive constant c_2. Thus,

$$\begin{aligned}
\boldsymbol{A}^{1/2} + \gamma \boldsymbol{B}^{1/2} &\leq c_1^{-1/2} \|\boldsymbol{\nu}\|_2 \, \mathbb{E}\left[\left\|\boldsymbol{z}_{i,t}\right\|_2^2\right]^{1/2} + \gamma^2 c_1^{-1/2} \|\boldsymbol{\nu}\|_2 \, \mathbb{E}\left[\left\|\boldsymbol{U}_{i,t+1}\right\|_2^2\right]^{1/2} \\
&\leq 2 c_1^{-1/2} c_2^{1/2} \|\boldsymbol{\nu}\|_2.
\end{aligned}$$

By Assumption 4.5 and $\boldsymbol{B} \geq 0$, we have $\boldsymbol{A} \geq \boldsymbol{A} - \gamma^2 \boldsymbol{B} \geq c_3 \|\boldsymbol{\nu}\|_2^2$, $\forall \boldsymbol{\nu} \in \mathbb{R}^{JM}$. Thus,

$$\boldsymbol{\nu}^\top \mathbb{E}\left[\widetilde{\Sigma}_k\right] \boldsymbol{\nu} \geq 2^{-1} c_1^{1/2} c_2^{-1/2} c_3^{3/2} \|\boldsymbol{\nu}\|_2^2.$$

It follows that $\left\|\mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1}\right\|_2 \leq 2C_1^{-1}$.

(ii) To bound $\left\|\widetilde{\Sigma}_k - \mathbb{E}\left[\widetilde{\Sigma}_k\right]\right\|$, we first bound $\left\|\sum_{i=1}^{N_k}(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)\right\|$ by applying the matrix concentration inequality in (Tropp, 2012, Theorem 1.6) and using results in Lemma D.10:

$$\Pr\left(\left\|\sum_{i=1}^{N_k}(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)\right\|_2 \geq \delta\right) \leq 2JM \cdot \exp\left(\frac{-\delta^2/2}{c_2 J N_k + 4c_1^2 J\delta/3}\right).$$

Set $\delta = 3\sqrt{CJN_k \log N_k}$. We can show that the following event occurs with probability at least $1 - \mathcal{O}\left((N_k T)^{-2}\right)$,

$$\left\|\widetilde{\Sigma}_k - \mathbb{E}\left[\widetilde{\Sigma}_k\right]\right\|_2 = \left\|\frac{1}{N_k}\sum_{i=1}^{N_k}(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)\right\|_2 \leq C\sqrt{J(N_k T)^{-1}\log(N_k T)},$$

for some constant C.

(iii) Since $J(N_k T)^{-1}\log(N_k T) = \mathrm{o}(1)$, with probability at least $1 - \mathcal{O}\left((N_k T)^{-2}\right)$

$$\left\|\widetilde{\Sigma}_k\right\|_2^2 \geq \left\|\mathbb{E}\left[\widetilde{\Sigma}_k\right]\right\|_2 - \left\|\Sigma_k - \mathbb{E}\left[\widetilde{\Sigma}_k\right]\right\|_2 \geq C_1/2 - \sqrt{J(N_k T)^{-1}\log(N_k T)} \geq C_1/6.$$

This implies that $\widetilde{\Sigma}_k$ is invertible and $\|\Sigma_k\|_2^2 \leq 6C_1^{-1}$ with probability at least $1 - \mathcal{O}\left((N_k T)^{-2}\right)$.

(iv) Applying results in (i), (ii) and (iii), we have

$$\begin{aligned}
\left\|\widetilde{\Sigma}_k^{-1} - \mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1}\right\|_2 &= \left\|\widetilde{\Sigma}_k^{-1}\left(\widetilde{\Sigma}_k - \mathbb{E}\left[\widetilde{\Sigma}_k\right]\right)\mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1}\right\|_2 \\
&\leq \left\|\widetilde{\Sigma}_k^{-1}\right\|_2 \cdot \left\|\widetilde{\Sigma}_k - \mathbb{E}\left[\widetilde{\Sigma}_k\right]\right\|_2 \cdot \left\|\mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1}\right\|_2 \\
&\leq 12C_1^{-2}\left\|\widetilde{\Sigma}_k - \mathbb{E}\left[\widetilde{\Sigma}_k\right]\right\|_2 \\
&\leq 12C_1^{-2}C\sqrt{J(N_k T)^{-1}\log(N_k T)},
\end{aligned}$$

with probability at least $1 - \mathcal{O}\left((N_k T)^{-2}\right)$.

□

D.2.2 Properties of $\widetilde{\zeta}_1$ and $\widetilde{\zeta}_2$

Lemma D.12. *Under Assumption 4.3, as either $N \to \infty$ or $T \to \infty$, we have*

(i) $\lambda_{\max}\left(T^{-1}\sum_{t=0}^{T-1} \boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\right) = \mathcal{O}_p(1)$

(ii) $\lambda_{\max}\left((N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t=0}^{T-1} \boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\right) = \mathcal{O}_p(1)$

(iii) $\lambda_{\min}\left(T^{-1}\sum_{t=0}^{T-1} \boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\right) \geq c/2$ *with probability approaching one.*

(iv) $\lambda_{\min}\left((N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t=0}^{T-1} \boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\right) \geq c/3$ *with probability approaching one.*

Proof. The proof is similar to that of Lemma D.11 and thus omitted here. □

Lemma D.13. *Under Assumption 4.3, there exists some positive constant c, c_1 such that*

(i) $\mathbb{E}[\zeta_{1k}] = 0$, $\mathbb{E}\left[\|\zeta_{1k}\|_2^2\right] \leq c_3 J(N_kT)^{-1}$ *and* $\zeta_{1k} = \mathcal{O}_p\left(\sqrt{J/(N_kT)}\right)$.

(ii) $\|\zeta_{2k}\|_2 = \mathcal{O}_p\left(J^{-\kappa/p}\right)$

(iii) $\Pr\left(\|\zeta_{1k}\|_\infty > c\sqrt{2\log(N_kT)/(N_kT)}\right) \leq 2JM(N_kT)^{-2}$.

(iv) $\|\zeta_{2k}\|_\infty \leq c_1 J^{-\kappa/p} \ll c_1 (N_kT)^{-1/2}$ *almost surely.*

Proof. Recall that

$$\zeta_{1k} = (N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T} \boldsymbol{z}_{i,t}\varepsilon_{1,i,t}, \quad \text{and} \quad \zeta_{2k} = (N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T} \boldsymbol{z}_{i,t}\varepsilon_{2,i,t}.$$

(i) By the Bellman first moment condition (7), MA and CMIA, we have $\mathbb{E}[\varepsilon_{1,i,t} \mid \mathcal{F}_{i,t}] = 0$. Thus

$$\mathbb{E}[\zeta_{1k}] = (N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{z}_{i,t}\varepsilon_{1,i,t} \mid \mathcal{F}_{i,t}\right]\right] = 0,$$

Also by the Bellman first moment condition (7), MA and CMIA, we have, for any $0 \leq t_1 < t_2 \leq T-1$,

$$\mathbb{E}\left[\varepsilon_{i,t_1}\varepsilon_{i,t_2}\boldsymbol{Z}_{i,t_1}\boldsymbol{Z}_{i,t_2}^\top\right] = \mathbb{E}\left[\varepsilon_{i,t_1}\boldsymbol{Z}_{i,t_1}\boldsymbol{Z}_{i,t_2}^\top\mathbb{E}\left[\varepsilon_{i,t_2} \mid \mathcal{F}_{t_2}\right]\right] = 0,$$

since $\boldsymbol{z}_{i,t}$ is a function of $\boldsymbol{X}_{i,t}$ and $A_{i,t}$ only. By the independence assumption, we have, for any $0 \leq t_1 \leq t_2 \leq T-1$ and $i_1 < i_2 \in \mathcal{G}_k$, that $\mathbb{E}\left[\varepsilon_{i_1,t_1}\varepsilon_{i_2,t_2}\boldsymbol{Z}_{i_1,t_1}\boldsymbol{Z}_{i_2,t_2}^\top\right] = 0$.

Applying Lemma D.8 (i), we have

$$\begin{aligned}\mathbb{E}\left[\|\zeta_{1k}\|_2^2\right] &= (N_k T)^{-2} \sum_{i=\mathcal{G}_k}^{} \sum_{t=1}^{T} \mathbb{E}\left[\varepsilon_{1,i,t}^2 \boldsymbol{z}_{i,t}^\top \boldsymbol{z}_{i,t}\right] \leq (c_1 + 2c_2)^2 (N_k T)^{-1} \mathbb{E}\left[\boldsymbol{z}_{i,t}^\top \boldsymbol{z}_{i,t}\right] \\ &\leq (N_k T)^{-1} (c_1 + 2c_2)^2 \sup_{\boldsymbol{x}\in\mathcal{X}} \left\|\boldsymbol{\phi}(\boldsymbol{x})\right\|_2^2 \\ &\leq c_3 J (N_k T)^{-1},\end{aligned}$$

for some constant c_3. By Markov's inequality, we obtain

$$\Pr\left(\|\zeta_{1k}\|_2 > \delta\right) \leq \delta^{-2} \mathbb{E}\left[\|\zeta_{1k}\|_2^2\right] \leq \delta^{-2} c_3 J (N_k T)^{-1}.$$

The result follows.

(ii) By Lemma D.8 and Cauchy-Schwarz inequality, we have, for any $\boldsymbol{\nu} \in \mathbb{R}^{JM}$,

$$\begin{aligned}\left|\boldsymbol{\nu}^\top \zeta_{2k} \zeta_{2k}^\top \boldsymbol{\nu}\right| = \left|\boldsymbol{\nu}^\top \zeta_{2k}\right|^2 &\leq \left((N_k T)^{-1} \sum_{i\in\mathcal{G}_k}^{} \sum_{t=1}^{T} \left|\boldsymbol{\nu}^\top \boldsymbol{z}_{i,t}\right| \left|\varepsilon_{2,i,t}\right| \right)^2 \\ &\leq \sup_{i\in\mathcal{G}_k, 1\leq t\leq T} \left|\varepsilon_{2,i,t}\right|^2 \cdot \left((N_k T)^{-1} \sum_{i\in\mathcal{G}_k}^{} \sum_{t=1}^{T} \left|\boldsymbol{\nu}^\top \boldsymbol{z}_{i,t}\right| \right)^2 \\ &\leq (c_4 J^{-\kappa/p})^2 \left((N_k T)^{-1} \sum_{i\in\mathcal{G}_k}^{} \sum_{t=1}^{T} \left|\boldsymbol{\nu}^\top \boldsymbol{z}_{i,t} \boldsymbol{z}_{i,t}^\top \boldsymbol{\nu}\right| \right)\end{aligned}$$

Thus we have,

$$\|\zeta_{2k}\|_2 = \max_{\|\boldsymbol{\nu}\|=1} \left(\left|\boldsymbol{\nu}^\top \zeta_{2k} \zeta_{2k}^\top \boldsymbol{\nu}\right|\right)^{1/2} \leq c_4 J^{-\kappa/p} \cdot \lambda_{\max}^{1/2} \left((N_k T)^{-1} \sum_{i\in\mathcal{G}_k}^{} \sum_{t=1}^{T} \boldsymbol{z}_{i,t} \boldsymbol{z}_{i,t}^\top \right)$$

Applying Lemma D.12, we have,

$$\|\zeta_{2k}\|_2 = \mathcal{O}_p\left(J^{-\kappa/p}\right)$$

(iii) By union bounds and Bernstein inequality for a.s. bounded $\varepsilon_{1,i,t}$ (51), we have

$$\begin{aligned}\Pr\left(\|\zeta_{1k}\|_\infty > \delta\right) &\le \sum_{1\le j\le J, 1\le m\le M} \Pr\left(\left|\phi_j(\boldsymbol{X}_{i,t})\mathbf{1}(A_{i,t}=m)(N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\varepsilon_{1,i,t}\right| > \delta\right)\\ &\le JM\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\phi_j(\boldsymbol{x})\right|\cdot\left|(N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\varepsilon_{1,i,t}\right| > \delta\right)\\ &\le 2JM\exp\left(-c^{-2}N_kT\delta^2\right).\end{aligned}$$

Setting $\delta = c\sqrt{2\log(N_kT)/(N_kT)}$, we have

$$\Pr\left(\|\zeta_{1k}\|_\infty > c\sqrt{2\log(N_kT)/(N_kT)}\right) \le 2JM(N_kT)^{-2}.$$

(iv) Under the condition that $J^{\kappa/p} \gg \sqrt{N_kT}$ and $J \ll \sqrt{N_kT}/\log(N_kT)$, we have,

$$\|\zeta_{2k}\|_\infty \le \sup_{\boldsymbol{x}\in\mathcal{X}}\left|\phi_j(\boldsymbol{x})\right|\cdot\left|(N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\varepsilon_{2,i,t}\right| \le c_1 J^{-\kappa/p} \ll c_1\left(N_kT\right)^{-1/2}, \quad \text{almost surely.}$$

□

D.2.3 Properties of variance estimator

Define

$$\begin{aligned}\widetilde{\boldsymbol{\Omega}}_k &= (N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top(\varepsilon_{1,i,t}+\varepsilon_{2,i,t})^2\\ \boldsymbol{\Omega}_k &= (N_kT)^{-1}\mathbb{E}\left[\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top(\varepsilon_{1,i,t}+\varepsilon_{2,i,t})^2\right]\\ \widetilde{\boldsymbol{H}}_k &= (N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\varepsilon_{1,i,t}^2\\ \boldsymbol{H}_k &= \mathbb{E}\left[(N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^\top\varepsilon_{1,i,t}^2\right].\end{aligned}$$

For any $\boldsymbol{x}\in\mathcal{X}$, $a\in\mathcal{A}$, define

$$w^\pi(\boldsymbol{x},a) = \mathbb{E}\left[\varepsilon_{1,i,t}^2\middle|\boldsymbol{X}_{i,t}=\boldsymbol{x}, A_{i,t}=a\right].$$

Then $\boldsymbol{H}_k = \mathbb{E}\Big[T^{-1}\sum_{t\in[T]} w^{\pi}(\boldsymbol{X}_{i,t}, A_{i,t})\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^{\top}\Big]$.

Lemma D.14. *Under the conditions in Theorem D.2, we have*

(i) $\lambda_{\min}(\boldsymbol{H}_k) \geq 3^{-1}c_0^{-1}c$

(ii) $\lambda_{\max}(\boldsymbol{H}_k) = \mathcal{O}(1)$ *or* $\|\boldsymbol{H}_k\|_2 = \mathcal{O}(1)$

(iii) $\big\|\widetilde{\boldsymbol{\Omega}}_k - \widetilde{\boldsymbol{H}}_k\big\|_2 = \mathrm{o}_{\mathrm{p}}(1)$

(iv) $\big\|\widetilde{\boldsymbol{H}}_k - \boldsymbol{H}_k\big\|_2 = \mathrm{o}_{\mathrm{p}}(1)$

Proof. (i) By Lemma D.11 (i) and the fact that $\inf w^{\pi}(\boldsymbol{x}, a) \geq c_0^{-1}$, we have

$$\lambda_{\min}(\boldsymbol{H}_k) \geq c_0^{-1}\lambda_{\min}\left((N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\mathbb{E}\Big[\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^{\top}\Big]\right) \geq 3^{-1}c_0^{-1}c.$$

(ii) By Lemma D.8 (i) and D.11 (i), we have

$$\lambda_{\max}(\boldsymbol{H}_k) \leq (c_1+2c_2)^2\lambda_{\max}\left((N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\mathbb{E}\Big[\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t}^{\top}\Big]\right) = \mathcal{O}(1).$$

(iii) Applying Lemma D.9 and D.11, we have

$$\begin{aligned}\big\|\widetilde{\boldsymbol{\Omega}}_k - \widetilde{\boldsymbol{H}}_k\big\|_2 &= (N_kT)^{-1}\left\|\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{z}_{i,t}^{\top}\boldsymbol{z}_{i,t}\Big((\varepsilon_{1,i,t}+\widetilde{\varepsilon}_{2,i,t})^2 - \varepsilon_{1,i,t}^2\Big)\right\|_2 \\ &\leq \sup_{i,t}\Big|(\varepsilon_{1,i,t}+\widetilde{\varepsilon}_{2,i,t})^2 - \varepsilon_{1,i,t}^2\Big| \cdot \left\|(N_kT)^{-1}\sum_{i\in\mathcal{G}_k}\sum_{t\in[T]}\boldsymbol{z}_{i,t}^{\top}\boldsymbol{z}_{i,t}\right\|_2 \\ &= \mathrm{o}_{\mathrm{p}}(1).\end{aligned}$$

(iv) The proof uses similar arguments in bounding $\Big\|\widetilde{\Sigma}_k - \mathbb{E}\Big[\widetilde{\Sigma}_k\Big]\Big\|_2$ and is omitted here. □

Lemma D.15. *Under the conditions in Theorem D.2, we have*

$$\Big\|\widetilde{\Sigma}_k^{-1}\widetilde{\boldsymbol{\Omega}}_k(\widetilde{\Sigma}_k^{\top})^{-1} - \Sigma_k^{-1}\boldsymbol{H}_k(\Sigma_k^{\top})^{-1}\Big\|_2 = \mathrm{o}_{\mathrm{p}}(1).$$

Proof. We have

$$
\begin{aligned}
&\left\|\widetilde{\Sigma}_k^{-1}\widetilde{\boldsymbol{\Omega}}_k(\widetilde{\Sigma}_k^{\top})^{-1} - \mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1}\boldsymbol{H}_k(\mathbb{E}\left[\widetilde{\Sigma}_k\right]^{\top})^{-1}\right\|_2 \\
&\leq \left\|\widetilde{\Sigma}_k^{-1}\widetilde{\boldsymbol{\Omega}}_k(\widetilde{\Sigma}_k^{\top})^{-1} - \widetilde{\Sigma}_k^{-1}\boldsymbol{H}_k(\widetilde{\Sigma}_k^{\top})^{-1}\right\|_2 + \left\|\widetilde{\Sigma}_k^{-1}\boldsymbol{H}_k(\widetilde{\Sigma}_k^{\top})^{-1} - \mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1}\boldsymbol{H}_k(\mathbb{E}\left[\widetilde{\Sigma}_k\right]^{\top})^{-1}\right\|_2 \\
&\leq \left\|\widetilde{\Sigma}_k^{-1}\right\|_2^2\left\|\widetilde{\boldsymbol{\Omega}}_k - \boldsymbol{H}_k\right\|_2 \\
&\quad + \left\|\left(\widetilde{\Sigma}_k^{-1} - \mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1}\right)\boldsymbol{H}_k(\widetilde{\Sigma}_k^{\top})^{-1}\right\| + \left\|\mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1}\boldsymbol{H}_k\left(\widetilde{\Sigma}_k^{-1} - \mathbb{E}\left[\widetilde{\Sigma}_k\right]^{-1}\right)^{\top}\right\|_2 \\
&= \mathrm{o}_{\mathrm{p}}(1),
\end{aligned}
$$

where we use the results in Lemma D.11 and D.14. □