DOES THE MARKOV DECISION PROCESS FIT THE DATA: TESTING FOR THE MARKOV PROPERTY IN SEQUENTIAL DECISION MAKING

A PREPRINT

Chengchun Shi
London School of Economics and Political Science

Runzhe Wan
North Carolina State University

Rui Song
North Carolina State University

Wenbin Lu
North Carolina State University

Ling Leng
Amazon

ABSTRACT

The Markov assumption (MA) is fundamental to the empirical validity of reinforcement learning. In this paper, we propose a novel Forward-Backward Learning procedure to test MA in sequential decision making. The proposed test does not assume any parametric form on the joint distribution of the observed data and plays an important role for identifying the optimal policy in high-order Markov decision processes and partially observable MDPs. We apply our test to both synthetic datasets and a real data example from mobile health studies to illustrate its usefulness.

1 Introduction

Reinforcement learning (RL) is a general technique that allows an agent to learn and interact with an environment. In RL, the state-action-reward triplet is typically modelled by the Markov decision process (MDP, see e.g. Puterman, 1994). Central to the empirical validity of various RL algorithms is the Markov assumption (MA). Under MA, there exists an optimal stationary policy that is no worse than any non-stationary or history dependent policies (Puterman, 1994; Sutton & Barto, 2018). When this assumption is violated, the optimal policy might depend on lagged variables and any stationary policy can be sub-optimal. Thus, MA forms the basis for us to select the set of state variables to implement RL algorithms. The focus of this paper is to test MA in sequential decision making problems.

1.1 Contributions and advances of our test

First, our test is useful in identifying the optimal policy in high-order MDPs (HMDPs). Under HMDPs, the optimal policy at time t depends not only on the current covariates $S_{0,t}$, but also the past state-action pairs $(S_{0,t-1}, A_{0,t-1})$, $\cdots$, $(S_{0,t-\kappa_0+1}, A_{0,t-\kappa_0+1})$ for some $\kappa_0 > 1$ (see Lemma 2 for a formal statement). In real-world applications, it remains challenging to properly select the look-back period κ_0. On one hand, κ_0 shall be sufficiently large to guarantee MA holds. On the other hand, including too many lagged variables will result in a very noisy policy. To determine κ_0, we propose to construct the state by concatenating measurements taken at time points $t, \cdots, t-k+1$ and sequentially apply our test for $k = 1, 2, \cdots$, until the null hypothesis MA is not rejected. Then we use existing RL algorithms based on the constructed state to estimate the optimal policy. We apply such a procedure to both synthetic and real

datasets in Section 5.2. Results show that the estimated policy based on our constructed states achieves the largest value in almost all cases.

Second, our test is useful in detecting partially observable MDPs. Suppose we concatenate measurements over sufficiently many decision points and our test still rejects MA. Then we shall consider modelling the system dynamics by partially observable MDPs (POMDPs) or other non-Markovian problems. Applying RL algorithms designed for these settings have been shown to yield larger value functions than those for standard MDPs (see e.g. Hausknecht & Stone, 2015). In Section 5.3, we illustrate the usefulness of our test in detecting POMDPs.

Third, we propose a novel testing procedure to test MA. To the best of our knowledge, this is the first work on developing valid statistical tests for MA in sequential decision making. Major challenges arise when the state vector is high-dimensional. This is certainly the case as we convert the process into an MDP by concatenating data over multiple decision points. To deal with high-dimensionality, we proposed a novel forward-backward learning procedure to construct the test statistic. The key ingredient of our test lies in constructing a doubly robust estimating equation to alleviate biases of modern machine learning algorithms. This ensures our test statistic has a tractable limiting distribution. In addition, since the test is constructed based on forward and backward learners (see Section 3.2 for details) estimated using the state-of-the-art machine learning estimation methods, it is well-suited to high-dimensional settings.

Lastly, our test is valid as either the number of trajectories n or the number of decision points T in each trajectory diverges to infinity. It can thus be applied to a variety of sequential decision making problems ranging from the Framingham heart study (Tsao & Vasan, 2015) with over two thousand trajectories to the OhioT1DM dataset (Marling & Bunescu, 2018a) that contains eight weeks' worth of data for six trajectories. Our test can also be applied to applications from video games where both n and T approach infinity.

1.2 Related work

There exists a huge literature on developing RL algorithms. Some recent popular methods include fitted Q-iteration (Riedmiller, 2005), deep Q-network (Mnih et al., 2015), double Q-learning (Van Hasselt et al., 2016), asynchronous advantage actor-critic (Mnih et al., 2016), etc. All the above mentioned methods model the sequential decision making problems by MDPs. When the Markov assumption is violated, the foundation of these algorithms is shaking hence may lead to deterioration of their performance to different degrees.

Currently, only a few methods have been proposed to test the Markov assumption. Among those available, Chen & Hong (2012) developed such a test in time series analysis. Constructing their test statistic requires to estimate the conditional characteristic function (CCF) of the current measurements given those taken in the past. Chen & Hong (2012) proposed to estimate the CCF based on local polynomial regression (Stone, 1977). We note their method cannot be directly used to test MA in MDP. Even though we can extend their method to our setup, the resulting test will perform poorly in settings where the dimension of the state vector is large, since local polynomial fitting suffers from the curse of dimensionality.

Our work is also related to the literature on conditional independence testing (see e.g. Zhang et al., 2012; Su & White, 2014; Wang et al., 2015; Huang et al., 2016; Wang & Hong, 2018; Berrett et al., 2020). However, all the above methods require observations to be independent and are not suitable to our settings where measurements are time dependent.

1.3 Organization of the paper

The rest of the paper is organized as follows. In Section 2, we introduce the MDP, HMDP and POMDP models, and establish the existence of the optimal stationary policy under MA. In Section 3, we introduce our testing procedure for MA and prove the validity of our test. In Section 4, we introduce a forward procedure based on our test for model selection. Empirical studies are presented in Section 5.

2 Model setup

2.1 MDP and existence of the optimal stationary policy

Let $(S_{0,t}, A_{0,t}, R_{0,t})$ denote the state-action-reward triplet collected at time t. For any integer $t \geq 0$, let $\bar{\boldsymbol{S}}_{0,t} = (S_{0,0}, A_{0,0}, S_{0,1}, A_{0,1}, \cdots, S_{0,t})^\top$ denote the state and action history. For simplicity, we assume the action set $\mathcal{A}$ is finite and the rewards are uniformly bounded. In MDPs, it is typically assumed that the following Markov assumption holds,

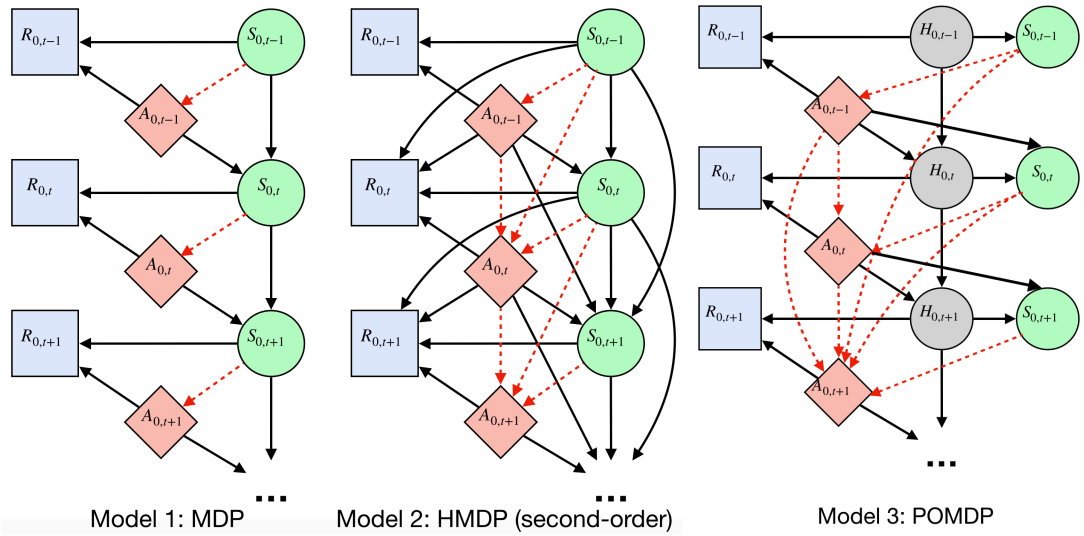

Figure 1: Causal diagrams for MDPs, HMDPs and POMDPs. The solid lines represent the causal relationships and the dashed lines indicate the information needed to implement the optimal policy.

$$\mathbb{P}(S_{0,t+1} \in \mathcal{S}, R_{0,t} \in \mathcal{R} | A_{0,t}, \bar{\boldsymbol{S}}_{0,t}, \{R_{0,j}\}_{j<t}) = \mathcal{P}(\mathcal{S}, \mathcal{R}; A_{0,t}, S_{0,t}),$$

for some Markov transition kernel $\mathcal{P}$ and any $\mathcal{S} \subseteq \mathbb{S}$, $\mathcal{R} \subseteq \mathbb{R}$, $t \geq 0$ where $\mathbb{S} \in \mathbb{R}^p$ denotes the state space.

A *history-dependent* policy π is a sequence of decision rules $\{\pi_t\}_{t\geq 0}$ where each π_t maps $\bar{\boldsymbol{S}}_{0,t}$ to a probability mass function $\pi_t(\cdot|\bar{\boldsymbol{S}}_{0,t})$ on $\mathcal{A}$. When there exists some function π^* such that $\pi_t(\cdot|\bar{\boldsymbol{S}}_{0,t}) = \pi^*(\cdot|S_{0,t})$ for any $t \geq 0$ almost surely, we refer to π as a *stationary* policy.

For a given discounted factor $0 < \gamma < 1$, the objective of RL is to learn an optimal policy $\pi = \{\pi_t\}_{t\geq 0}$ that maximizes the value function

$$V(\pi; s) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^{\pi_t}(R_{0,t} | S_{0,0} = s),$$

for any $s \in \mathbb{S}$, where the expectation $\mathbb{E}^{\pi_t}$ is taken by assuming that the system follows π_t. Let HR and SR denote the class of history-dependent and stationary policies, respectively. The following lemma forms the basis of existing RL algorithms.

Lemma 1 *Under MA, there exists some $\pi^{opt} \in$ SR such that $V(\pi^{opt}; s) = \sup_{\pi \in HR} V(\pi; s)$ for any $s \in \mathbb{S}$.*

Lemma 1 implies that under MA, it suffices to restrict attention to stationary policies. This greatly simplifies the estimating procedure of the optimal policy. When MA is violated however, we need to focus on history-dependent policies as they may yield larger value functions.

When the state space is discrete, Lemma 1 is implied by Theorem 6.2.10 of Puterman (1994). For completeness, we provide a proof in Appendix C.1 assuming $\mathbb{S}$ belongs to a general vector space. In the following, we introduce two variants of MDPs, including HMDPs and POMDPs. These models are illustrated in Figure 1.

2.2 HMDP

It can be seen from Figure 1 that HMDPs are very similar to MDPs. The difference lies in that in HMDPs, $S_{0,t+1}$ and $R_{0,t}$ depend not only on $(S_{0,t}, A_{0,t})$, but $(S_{0,t-1}, A_{0,t-1}), \cdots, (S_{0,t-\kappa_0+1}, A_{0,t-\kappa_0+1})$ for some integer $\kappa_0 > 1$ as well. Formally, we have

$$\mathbb{P}(S_{0,t+1} \in \mathcal{S}, R_{0,t} \in \mathcal{R} | A_{0,t}, \bar{\boldsymbol{S}}_{0,t}, \{R_{0,j}\}_{j<t}) = \mathcal{P}(\mathcal{S}, \mathcal{R}; \{A_{0,j}\}_{t-\kappa_0<j\leq t}, \{S_{0,j}\}_{t-\kappa_0<j\leq t}), \quad (1)$$

for some $\mathcal{P}$, κ_0 and any $\mathcal{S} \subseteq \mathbb{S}$, $\mathcal{R} \subseteq \mathbb{R}$, $t > \kappa_0$. For any integer $k > 0$, define a new state variable

$$S_{0,t}(k) = (S_{0,t}^\top, A_{0,t}, S_{0,t+1}^\top, A_{0,t+1}, \cdots, S_{0,t+k-1}^\top)^\top.$$

Let $A_{0,t}(k) = A_{0,t+k-1}$ and $R_{0,t}(k) = R_{0,t+k-1}$ for any t, k. It follows from (1) that the new process formed by the triplets $(S_{0,t}(\kappa_0), A_{0,t}(\kappa_0), R_{0,t}(\kappa_0))_{t\geq 0}$ satisfies MA.

For any $k > 0$, let SR(k) denote the set of stationary policies $\pi = \{\pi_t\}_{t\geq 0}$ such that π_t depend on $\bar{\boldsymbol{S}}_{0,t}$ only through $S_{0,t-k}(k)$. Suppose we are interested in identifying a policy that maximizes the following k-step value function

$$V^{(k)}(\pi; s) = \sum_{t\geq 0} \gamma^t \mathbb{E}^{\pi_t}\{R_{0,t}(k)|S_{0,0}(k) = s\},$$

for any $s \in \mathbb{S}(k)$, the state space for $S_{0,t}(k)$. By Lemma 1, we obtain the following results.

Lemma 2 *Assume* (1) *holds. Then there exists some* $\pi^{opt} \in$ *SR*(κ_0) *such that* $V^{(k)}(\pi^{opt}; s) = \sup_{\pi\in HR} V^{(k)}(\pi; s)$ *for any* $s \in \mathbb{S}(k)$ *and* $k \geq \kappa_0$.

Lemma 2 suggests that in HMDPs, identification of the optimal policy relies on correct specification of the look-back period κ_0. To determine κ_0, we can sequentially test whether the triplets $\{(S_{0,t}(k), A_{0,t}(k), R_{0,t}(k))\}_{t\geq 0}$ satisfy MA for $k = 1, 2, \cdots$, until the null MA is not rejected.

2.3 POMDP

The POMDP model can be described as follows. At time $t-1$, suppose the environment is in some hidden state $H_{0,t-1}$. The hidden variables $\{H_{0,t}\}_{t\geq 0}$ are unobserved. Suppose the agent chooses an action $A_{0,t-1}$. Similar to MDPs, this will cause the environment to transition to a new state $H_{0,t}$ at time t. At the same time, the agent receives an observation $S_{0,t} \in \mathbb{S}$ and a reward $R_{0,t}$ that depend on $H_{0,t}$ and $A_{0,t-1}$. The goal is to estimate an optimal policy based on the observed state-action pairs.

The observations in POMDPs do not satisfy the Markov property. To better illustrate this, consider the causal diagram for POMDP depicted in Figure 1. The path $S_{0,t-1} \leftarrow H_{0,t-1} \rightarrow H_{0,t} \rightarrow H_{0,t+1} \rightarrow S_{0,t+1}$ connects $S_{0,t-1}$ and $S_{0,t+1}$ without traversing $S_{0,t}$ and $A_{0,t}$. As a result, $S_{0,t+1}$ and $S_{0,t-1}$ are not d-separated (see the definition of d-separation on Page 16, Pearl, 2000) given $S_{0,t}$ and $A_{0,t}$. Under the faithfulness assumption (see e.g. Kalisch & Bühlmann, 2007), $S_{0,t-1}$ and $S_{0,t+1}$ are mutually dependent conditional on $S_{0,t}$ and $A_{0,t}$. Similarly, we can show $S_{0,t+k}$ and $S_{0,t-1}$ are mutually dependent conditional on $\{(S_{0,j}, A_{0,j})\}_{t\leq j<t+k}$ for any $k > 1$. As a result, the Markov assumption will not hold no matter how many past measurements the state variable includes. This suggests in POMDPs, the optimal policy could be history dependent.

3 Testing the Markov assumption

3.1 A CCF-based characterization of MA

For simplicity, suppose $R_{0,t}$ is a deterministic function of $S_{0,t+1}$, $A_{0,t}$ and $S_{0,t}$. This condition automatically holds if we include $R_{0,t}$ in the set of state variables $S_{0,t+1}$. It is also satisfied in our real dataset (see Section 5.2.1 for details). Under this condition, MA is equivalent to the following,

$$\mathbb{P}(S_{0,t+1} \in \mathcal{S}|A_{0,t}, \bar{\boldsymbol{S}}_{0,t}) = \mathcal{P}(\mathcal{S}; A_{0,t}, S_{0,t}), \tag{2}$$

for any $\mathcal{S} \subseteq \mathbb{S}$ and $t \geq 0$. Let $\{(S_{1,t}, A_{1,t}, R_{1,t})\}_{0\leq t\leq T}$, $\{(S_{2,t}, A_{2,t}, R_{2,t})\}_{0\leq t\leq T}$, $\cdots$, $\{(S_{n,t}, A_{n,t}, R_{n,t})\}_{0\leq t\leq T}$ be i.i.d. copies of $\{(S_{0,t}, A_{0,t}, R_{0,t})\}_{0\leq t\leq T}$. Given the observed data, we focus on testing the following hypothesis:

H_0: The system is a MDP, i.e, (2) holds v.s
H_1: The system is a HMDP or POMDP.

In the rest of this section, we present a CCF characterization of H_0. For any random vectors Z_1, Z_2, Z_3, we use the notation $Z_1 \perp\!\!\!\perp Z_2|Z_3$ to indicate that Z_1 and Z_2 are independent conditional on Z_3. To test H_0, it suffices to test the following conditional independence assumptions:

$$S_{0,t} \perp\!\!\!\perp \{(S_{0,j}, A_{0,j})\}_{0\leq j\leq t-2}|S_{0,t-1}, A_{0,t-1}, \forall t > 1. \tag{3}$$

For any t, let $X_{0,t} = (S_{0,t}^\top, A_{0,t})^\top$ denote the state-action pair. For any $\mu \in \mathbb{R}^p$, define the following CCF,

$$\varphi_t(\mu|x) = \mathbb{E}\{\exp(i\mu^\top S_{0,t+1})|X_{0,t} = x\}. \tag{4}$$

In the following, we present an equivalent representation for (3) based on (4).

Theorem 1 (3) *is equivalent to the following: for any $t > 0$, $q \geq 0$, $\mu \in \mathbb{R}^p$, $\nu \in \mathbb{R}^{p+1}$, we have almost surely,*

$$\begin{aligned} &\varphi_{t+q}(\mu|X_{0,t+q})\mathbb{E}[\exp(i\nu^\top X_{0,t-1})|\{X_{0,j}\}_{t\leq j\leq t+q}] \\ &= \mathbb{E}[\exp(i\mu^\top S_{0,t+q+1} + i\nu^\top X_{0,t-1})|\{X_{0,j}\}_{t\leq j\leq t+q}]. \end{aligned} \tag{5}$$

Under H_0, there exists some φ^* such that $\varphi_t = \varphi^*$ for any t. By Theorem 1, we can show that

$$\begin{aligned} &\mathbb{E}\{\exp(i\mu^\top S_{0,t+q+1}) - \varphi^*(\mu|X_{0,t+q})\}\exp(i\nu^\top X_{0,t-1}) \\ = \quad &\mathbb{E}\exp(i\mu^\top S_{0,t+q+1} + i\nu^\top X_{0,t-1}) - \mathbb{E}\varphi^*(\mu|X_{0,t+q})\exp(i\nu^\top X_{0,t-1}) = 0, \end{aligned}$$

for any t, q, μ, ν. This motivates us to consider the test statistic based on

$$\frac{1}{n(T-q-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-q-1}\{\exp(i\mu^\top S_{j,t+q+1}) - \widehat{\varphi}(\mu|X_{j,t+q})\}\{\exp(i\nu^\top X_{j,t-1}) - \bar{\varphi}(\nu)\}, \tag{6}$$

where $\widehat{\varphi}$ denotes some nonparametric estimator for φ^* and $\bar{\varphi}(\nu) = n^{-1}(T+1)^{-1}\sum_{1\leq j\leq n, 0\leq t\leq T}\exp(i\nu^\top X_{j,t-1})$.

Modern machine learning (ML) algorithms are well-suited to estimating φ^* in high-dimensional cases. However, naively plugging ML estimators for $\widehat{\varphi}$ will cause a heavy bias in (6). Because of that, the resulting estimating equation does not have a tractable limiting distribution. Kernel smoothers (Härdle, 1990) or local polynomial regression can be used to reduce the estimation bias by properly choosing the bandwidth parameter. However, as commented in Section 1.2, these methods suffer from the curse of dimensionality and will perform poorly in cases as we concatenate data over multiple decision points.

In the next section, we address these concerns by presenting a doubly-robust estimating equation to alleviate the estimation bias. When observations are time independent, our method shares similar spirits with the double machine learning method proposed by Chernozhukov et al. (2018) for statistical inference of the average treatment effects in causal inference.

3.2 Forward-Backward Learning

To introduce our method, we define another CCF

$$\psi_t(\nu|x) = \mathbb{E}\{\exp(i\nu^\top X_{0,t-1})|X_{0,t} = x\}. \tag{7}$$

We need the following two conditions.

(C1) Actions are generated by a fixed behavior policy.
(C2) Suppose the process $\{S_{0,t}\}_{t\geq 0}$ is strictly stationary.

Condition (C1) requires the agent to select actions based on information contained in the current state variable only. Under H_0, the process $\{S_{0,t}\}_{t\geq 0}$ forms a time-invariant Markov chain. When its initial distribution equals its stationary distribution, (C2) is automatically satisfied. This together with (C1) implies $\{X_{0,t}\}_{t\geq 0}$ is strictly stationary as well. As a result, we have $\psi_t = \psi^*$ for some ψ^* and any $t > 0$.

Theorem 2 *Suppose H_0, (C1) and (C2) hold. Then for any $t > 0$, $q \geq 0$, $\mu \in \mathbb{R}^p$, $\nu \in \mathbb{R}^{p+1}$, we have*

$$\mathbb{E}\Gamma_0(q, \mu, \nu) \equiv \mathbb{E}\{\exp(i\mu^\top S_{0,t+q+1}) - \varphi^*(\mu|X_{0,t+q})\}\{\exp(i\nu^\top X_{0,t-1}) - \psi^*(\nu|X_{0,t})\} = 0.$$

Moreover, the above equation is doubly-robust. That is, for any CCFs φ and ψ, the following holds as long as either $\varphi = \varphi^*$ or $\psi = \psi^*$,

$$\mathbb{E}\{\exp(i\mu^\top S_{0,t+q+1}) - \varphi(\mu|X_{0,t+q})\}\{\exp(i\nu^\top X_{0,t-1}) - \psi(\nu|X_{0,t})\} = 0. \tag{8}$$

Proof: *When $\varphi = \varphi^*$, we have*

$$\mathbb{E}[\exp(i\mu^\top S_{0,t+q+1}) - \varphi^*(\mu|X_{0,t+q})|\{X_{0,j}\}_{j\leq t+q}] = 0,$$

under MA. Assertion (8) *thus follows. Under (C1), we have $X_{0,t-1} \perp\!\!\!\perp \{X_{0,j}\}_{j>t}|X_{0,t}$ for any $t > 1$. When $\psi = \psi^*$, we can similarly show that*

$$\mathbb{E}[\exp(i\nu^\top X_{0,t-1}) - \psi^*(\nu|X_{0,t})|\{X_{0,j}\}_{j>t}] = 0.$$

The doubly-robustness property thus follows.

The propose algorithm estimates both φ^* and ψ^* using ML methods without specifying their parametric forms. Let $\widehat{\varphi}$ and $\widehat{\psi}$ denote the corresponding estimators. Note that computing φ^* is essentially estimating the characteristic function of $S_{0,t}$ given $S_{0,t-1}$. This corresponds to a forward prediction task. Similarly, estimating ψ^* is a backward prediction task. Thus, we refer to $\widehat{\varphi}$ and $\widehat{\psi}$ as **forward** and **backward learners**, respectively. Our proposed method is referred to as the **forward-backward learning** algorithm. It is worth mentioning that although we focus on the problem of testing MA in this paper, the proposed method can be applied to more general estimation and inference problems with time-dependent observations.

Consider the following estimating equation,

$$\frac{1}{n(T-q-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-q-1}\{\exp(i\mu^\top S_{j,t+q+1})-\widehat{\varphi}(\mu|X_{j,t+q})\}\{\exp(i\nu^\top X_{j,t-1})-\widehat{\psi}(\nu|X_{j,t})\}. \tag{9}$$

Unlike (6), the above estimating equation is doubly robust. This helps alleviate the impact of the biases in $\widehat{\varphi}$ and $\widehat{\psi}$.

Our test statistic is constructed based on a slightly modified version of (9) with cross-fitting. The use of cross-fitting allows us to establish the limiting distribution of the estimating equation under minimal conditions.

Suppose we have at least two trajectories, i.e, $n \geq 2$. We begin by randomly dividing $\{1,\cdots,n\}$ into $\mathbb{L}$ subsets $\mathcal{I}^{(1)},\cdots,\mathcal{I}^{(\mathbb{L})}$ of equal size. Denote by $\mathcal{I}^{(-\ell)}=\{1,\cdots,n\}-\mathcal{I}^{(\ell)}$ for $\ell=1,\cdots,\mathbb{L}$. Let $\widehat{\varphi}^{(-\ell)}$ and $\widehat{\psi}^{(-\ell)}$ denote the forward and backward learners based on the data in $\mathcal{I}^{(-\ell)}$. For any μ,ν,q, define

$$\widehat{\Gamma}(q,\mu,\nu)=\frac{n^{-1}}{T-q-1}\sum_{\ell=1}^{\mathbb{L}}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T-q-1}\{\exp(i\mu^\top S_{j,t+q+1})-\widehat{\varphi}^{(-\ell)}(\mu|X_{j,t+q})\}\{\exp(i\nu^\top X_{j,t-1})-\widehat{\psi}^{(-\ell)}(\nu|X_{j,t})\}.$$

Notice that $\widehat{\Gamma}$ is a complex-valued function. We use $\widehat{\Gamma}_R$ and $\widehat{\Gamma}_I$ to denote its real and imaginary part.

Algorithm 1 Forward-Backward Learning

Input: B, Q, $\mathbb{L}$, α and the observed data.
Step 1: Randomly generate i.i.d. pairs $\{(\mu_b,\nu_b)\}_{1\leq b\leq B}$ from $N(0,I)$; Randomly divide $\{1,\cdots,n\}$ into $\bigcup_\ell \mathcal{I}^{(\ell)}$ for $\ell=1,\cdots,\mathbb{L}$, set $\mathcal{I}^{(-\ell)}=\{1,\cdots,n\}-\mathcal{I}^{(\ell)}$.
Step 2: Compute the forward and backward learners $\widehat{\varphi}^{(-\ell)}(q,\mu_b,\cdot)$ and $\widehat{\psi}^{(-\ell)}(q,\nu_b,\cdot)$ for $q=0,\cdots,Q$, $b=1,\cdots,B$ based on modern ML methods.
Step 3: Compute $\widehat{\Gamma}(q,\mu_b,\nu_b)$ for $q=0,\cdots,Q$, $b=1,\cdots,B$; Compute $\widehat{S}$ according to (10).
Step 4: For $q=0,\cdots,Q$, compute an estimated covariance matrix $\widehat{\Sigma}^{(q)}$ according to (11) (see Appendix A.1 for details).
Step 5: Use Monte Carlo to simulate the upper $\alpha/2$-th critical value of $\max_{q\in\{0,\ldots,Q\}}\|\{\widehat{\mathbf{\Sigma}}^{(q)}\}^{1/2}\mathbb{Z}_q\|_\infty$ where $\mathbb{Z}_2,\cdots,\mathbb{Z}_Q$ are i.i.d. $2B$-dimensional random vectors with identity covariance matrix. Denote this critical value by $\widehat{c}_\alpha$.
Reject H_0 if $\widehat{S}$ is greater than $\widehat{c}_\alpha$.

To implement our test, we randomly sample i.i.d. pairs $\{(\mu_b,\nu_b)\}_{1\leq b\leq B}$ according to a multivariate normal distribution with zero mean and identity covariance matrix, where B is allowed to diverge with the number of observations. Let Q be some large integer that is allowed to be proportion to T (see the condition in Theorem 3 below for details). We calculate $\widehat{\Gamma}_R(q,\mu_b,\nu_b)$ and $\widehat{\Gamma}_I(q,\mu_b,\nu_b)$ for $b=1,\cdots,B$, $q=0,\cdots,Q$. Under H_0, $\widehat{\Gamma}_R(q,\mu_b,\nu_b)$ and $\widehat{\Gamma}_I(q,\mu_b,\nu_b)$ are close to zero. Thus, we reject H_0 when one of these quantities has large absolute value. Our test statistic is given by

$$\widehat{S}=\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\sqrt{n(T-q-1)}\max(|\widehat{\Gamma}_R(q,\mu_b,\nu_b)|,|\widehat{\Gamma}_I(q,\mu_b,\nu_b)|). \tag{10}$$

Under H_0, each $\widehat{\Gamma}_R(q,\mu_b,\nu_b)$ (or $\widehat{\Gamma}_I(q,\mu_b,\nu_b)$) is asymptotically normal. As a result, $\widehat{S}$ converges in distribution to a maximum of some Gaussian random variables. For a given significance level $\alpha>0$, we reject H_0 when $\widehat{S}>\widehat{c}_\alpha$ for some threshold $\widehat{c}_\alpha$ computed by wild bootstrap (Wu, 1986). We detail our procedure in Algorithm 1.

Step 2 of our algorithm requires to estimate $\widehat{\varphi}^{(-\ell)}(\mu_b|\cdot)$ and $\widehat{\psi}^{(-\ell)}(\nu_b|\cdot)$ for $b=1,\cdots,B$. The integer B shall be large enough to guarantee that our test has good power properties. Our method allows B to grow at an arbitrary polynomial order of $n\times T$ (see the condition in Theorem 3 below for details). Separately applying ML algorithms

B times to compute these leaners is computationally intensive. In Section 5.1, we use the random forests (Breiman, 2001) algorithm as an example to illustrate how these leaners can be simultaneously calculated. Other ML algorithms could also be used.

3.3 Bidirectional asymptotics

In this section, we prove the validity of our test under a bidirectional-asymptotic framework where either n or T grows to infinity. We begin by introducing some conditions.

(C3) Under H_0, suppose the Markov chain $\{X_{0,t}\}_{t\geq 0}$ is geometrically ergodic when $T \to \infty$.
(C4) Suppose there exists some $c_0 > 1/2$ such that

$$\max_{1\leq b\leq B} \int_x |\widehat{\varphi}^{(-\ell)}(\mu_b|x) - \varphi^*(\mu_b|x)|^2 \mathbb{F}(dx) = O_p((nT)^{-c_0}),$$
$$\max_{1\leq b\leq B} \int_x |\widehat{\psi}^{(-\ell)}(\nu_b|x) - \psi^*(\nu_b|x)|^2 \mathbb{F}(dx) = O_p((nT)^{-c_0}),$$

where $\mathbb{F}$ denotes the distribution function of $X_{0,0}$. In addition, suppose $\widehat{\varphi}^{(-\ell)}$ and $\widehat{\psi}^{(-\ell)}$ are bounded functions.

Condition (C3) enables us to establish the limiting distribution of our test under the setting where $T \to \infty$. Notice that this condition is not needed when T is bounded. The geometric ergodicity assumption (see e.g. Tierney, 1994, for definition) is weaker than the uniform ergodicity condition imposed in the existing reinforcement learning literature (see e.g. Bhandari et al., 2018; Zou et al., 2019). There exist Markov chains that are not uniformly ergodic but may still be geometrically ergodic (Mengersen & Tweedie, 1996).

The first part of Condition (C4) requires the prediction errors of estimated CCFs to satisfy certain uniform convergence rates. This is the key condition to ensure valid control of the type-I error rate of our test. In practice, the capacity of modern ML algorithms and their success in prediction tasks even in high-dimensional samples make this a reasonable assumption. In theory, the uniform convergence rates in (C4) can be derived for popular ML methods such as random forests (Biau, 2012) and deep neural networks (Schmidt-Hieber, 2020). The boundedness assumption in (C4) is reasonable since φ^* and ψ^* are bounded by 1.

Theorem 3 *Assume (C1)-(C4) hold. Suppose* $\log B = O((nT)^{c^*})$ *for any finite* $c^* > 0$ *and* $Q \leq \max(\rho_0 T, T-2)$ *for some constant* $\rho_0 < 1$. *In addition, suppose there exists some* $\epsilon_0 > 0$ *such that the real and imaginary part of* $\Gamma_0(q,\mu,\nu)$ *have variances greater than* ϵ_0 *for any* μ, ν *and* $q \in \{0, \cdots, Q\}$. *Then we have as either* $n \to \infty$ *or* $T \to \infty$, $\mathbb{P}(\widehat{S} > \widehat{c}_\alpha) = \alpha + o(1)$.

Theorem 3 implies the type-I error rate of our test is well-controlled. Our proof relies on the high-dimensional martingale central limit theorem that is recently developed by Belloni & Oliveira (2018). This enables us to show the asymptotic equivalence between the distribution of $\widehat{S}$ and that of the bootstrap samples given the data, under settings where B diverges with n and T. It is worthwhile to mention that the stationarity condition in (C2) is imposed to simplify the presentation. Our test remains valid when (C2) is violated. To save space, we move the related discussions to Appendix A.2.

4 Model selection

Algorithm 2 RL Model Selection

Input: B, Q, $\mathbb{L}$, α and the observed data.
for $k = 1, 2, \cdots, K$ **do**
  Apply algorithm 1 with B, Q, $\mathbb{L}$, α specified above to the data $\{(S_{j,t}(k), A_{j,t}(k))\}_{1\leq j\leq n, 0\leq t\leq T-k+1}$.
  if H_0 is not rejected **then**
    Conclude the system is a k-th order MDP; **Break**.
  end if
end for
Conclude the system is a POMDP.

Based on our test, we can choose which RL model to use to model the system dynamics. For any j, k, t, let

$$S_{j,t}(k) = (S_{j,t}^\top, A_{j,t}, S_{j,t+1}^\top, A_{j,t+1}, \cdots, S_{j,t+k}^\top)^\top,$$

and $A_{j,t}(k) = A_{j,t+k}$. Given a large integer K, our procedure sequentially test the null hypothesis MA based on the concatenated data $\{(S_{j,t}(k), A_{j,t}(k))\}_{1\le j\le n, 0\le t\le T-k}$ for $k = 0, 1, \cdots, K$. Once the null is not rejected, we can conclude the system is a k-th order MDP and terminate our procedure. Otherwise, we conclude the system is a POMDP. We summarize our method in Algorithm 2.

5 Numerical examples

This section is organized as follows. We discuss some implementation details in Section 5.1. In Section 5.2, we apply our test to mobile health applications. We use both synthetic and real datasets to demonstrate the usefulness of our test in detecting HMDPs. In Section 5.3, we apply our test to a POMDP problem to illustrate its consistency.

5.1 Implementation details

We first describe the algorithm we use to simultaneously compute $\{\widehat{\varphi}^{(-\ell)}(\mu_b|\cdot)\}_{1\le b\le B}$. The algorithm for computing backward learners can be similarly derived. Our method is motivated by the quantile regression forest algorithm (Meinshausen, 2006). We detail our procedure below. %vspace-0.4cm

1. Apply the random forests algorithm with the response-predictor pairs $\{(S_{j,t}, X_{j,t-1})\}_{j\in\mathcal{I}^{(-\ell)}, 1\le t\le T}$ to grow M trees $T(\theta_m)$ for $m = 1, \ldots, M$. Here θ_m denotes the parameters associated with the m-th tree. Denote by $l(x, \theta_m)$ the leaf space of the m-th tree that predictor x fails into. %vspace-0.2cm
2. For any $m \in \{1, \cdots, T\}$, $(j, t) \in \mathcal{I}^{(-\ell)}$ and x, compute the weight $w_{j,t}^{(-\ell)}(x, \theta_m)$ as
$$\frac{\mathbb{I}\{X_{j,t} \in l(x, \theta_m)\}}{\#\{(l_1, l_2) : l_1 \in \mathcal{I}^{(-\ell)}, X_{l_1,l_2} \in l(x, \theta_m)\}}.$$
Average over all trees to calculate the weight of each training data as $w_{j,t}^{(-\ell)}(x) = \sum_{m=1}^{M} w_{j,t}^{(-\ell)}(x, \theta_m)/M$.
3. For any x and $b \in \{1, \ldots, B\}$, compute the forward learner $\widehat{\varphi}^{(-\ell)}(\mu_b|x)$ as the weighted average $\sum_{j\in\mathcal{I}^{(-\ell)}, 1\le t\le T} w_{j,t}^{(-\ell)}(x) \exp(i\mu_b^\top S_{j,t})$.

To implement this algorithm, the number of trees M is set to 100 and other tuning parameters are selected via 5-fold cross-validation. To construct our test, the hyperparameters B, Q and $\mathbb{L}$ are fixed as 100, 8 and 3 respectively. All state variables are normalized to have unit sampling variance before running the test. Normalization will not affect the Type I error rate of our test but helps improve its power. Our experiments are run on an c5d.24xlarge instance on the AWS EC2 platform, with 96 cores and 192GB RAM.

5.2 Applications in HMDP problems

5.2.1 THE OHIOT1DM Dataset

There has been increasing interest in applying RL algorithms to mobile health (mHealth) applications. In this section, we use the OhioT1DM dataset Marling & Bunescu (2018b) as an example to illustrate the usefulness of test in mHealth applications. The data contains continuous measurements for six patents with type 1 diabetes over eight weeks. In order to apply RL algorithms, it is crucial to determine how many lagged variables we should include to construct the state vector.

In our experiment, we divide each day of follow-up into one hour intervals and a treatment decision is made every hour. We consider three important time-varying variables to construct $S_{0,t}$, including the average blood glucose levels $\mathrm{G}_{0,t}$ during the one hour interval $(t-1, t]$, the carbohydrate estimate for the meal $\mathrm{C}_{0,t}$ during $(t-1, t]$ and $\mathrm{Ex}_{0,t}$ which measures exercise intensity during $(t-1, t]$. At time t, we define $A_{0,t}$ by discretizing the amount of insulin $\mathrm{In}_{0,t}$ injected and define $R_{0,t}$ according to the Index of Glycemic Control (Rodbard, 2009) that is a deterministic function $\mathrm{G}_{0,t+1}$. To save space, we present detailed definitions of $A_{0,t}$ and $R_{0,t}$ in Appendix B.1.

5.2.2 synthetic data

We first simulate patients with type I diabetes to mimic the OhioT1DM dataset. According to our findings in Section 5.2.3, we model this sequential decision problem by a fourth order MDP. Specifically, we consider the following model for $\mathrm{G}_{0,t}$:
$$\mathrm{G}_{0,t} = \alpha + \sum_{i=1}^{4} (\boldsymbol{\beta}_i^T S_{0,t-i} + c_i A_{0,t-i}) + E_{0,t},$$

where α, $\{\boldsymbol{\beta}_i\}_{i=1}^4$ and $\{c_i\}_{i=1}^4$ are computed by least-square estimation based on the OhioT1DM dataset. The error term $E_{0,t}$ is set to follow $N(0, 9)$.

At each time point, a patient randomly choose to consume food with probability p_1 and take physical activity with probability p_2, where the amounts and intensities are independently generated from normal distributions. The initial values of $\mathrm{G}_{0,t}$ are also randomly sampled from a normal distribution. Actions are independently generated from a multinoulli distribution. Parameters p_1, p_2 as well as other parameters in the above distributions are all estimated from the data.

For each simulation, we generate $N = 10, 15$ or 20 trajectories according to the above model. For each trajectory, we generate measurements with $T = 1344$ time points (8 weeks) after an initial burn-in period of 10 time points. For $k \in \{1, \ldots, 10\}$, we use our test to determine whether the system is a k-th order MDP. Under our generative model, we have H_0 holds when $k \geq 4$ and H_1 holds otherwise.

%vspace-0.2cm Empirical rejection rates of our test with different combinations of k, N and the significance level

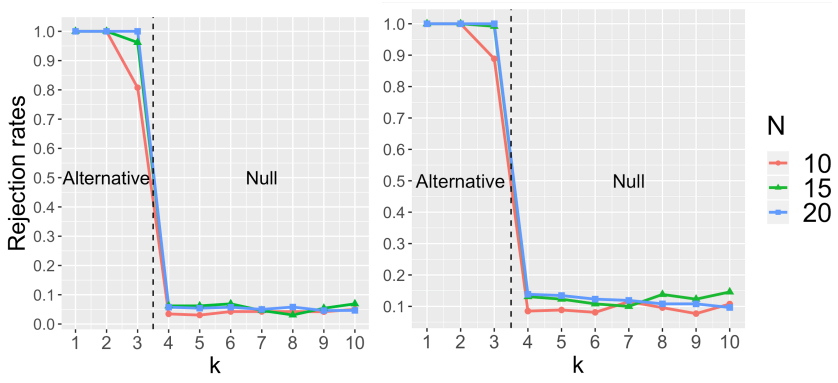

Figure 2: Empirical rejection rates aggregated over 500 simulations with different combinations of α, N and k. $\alpha = (0.05, 0.1)$ from left plot to right plot.

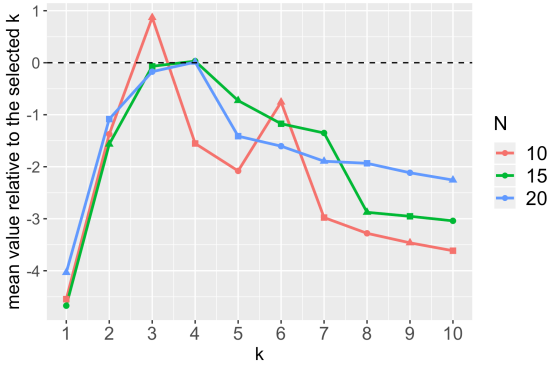

Figure 3: Value differences with different combinations of k and N.

α are reported in Figure 2. Results are aggregated over 500 simulations. It can be seen that the Type I error rate of our test is close to the nominal level in almost all cases. In addition, its power increases with N, demonstrating the consistency of our test.

Table 1: Policy evaluation results for the OhioT1DM dataset.

k	1	2	3	4	5	6	7	8	9	10
Estimated value $\bar{V}_k$	-90.82	-57.53	-63.77	**-52.57**	-56.23	-60.05	-63.70	-54.85	-65.08	-59.59

To further illustrate the usefulness of our test, we apply Algorithm 2 with $\alpha = 0.01$, $K = 10$ for model selection and evaluate the policy learned based on the selected model. Specifically, let $\widehat{\kappa}_0^{(l)}$ denote the order of MDP estimated by Algorithm 2 in the l-th simulation. For each $k \in \{1, \cdots, 10\}$, we apply the fitted-Q iteration algorithm (Ernst

et al., 2005, see Section B.2 for details) to the data $\{S_{j,t}(k), A_{j,t}(k), R_{j,t}(k)\}_{1\le j\le N, 0\le t\le T-k+1}$ generated in the l-th simulation to learn an optimal policy $\widehat{\pi}^{(l)}(k)$ and then simulate 100 trajectories following $\widehat{\pi}^{(l)}(k)$ to compute the average discounted reward $V^{(l)}(k)$ (see Appendix B.2 for details). Finally, for each $k = 1, \cdots, 10$, we compute the value difference

$$\text{VD}(k) = \frac{1}{500}\sum_{l=1}^{500}\{V^{(l)}(k) - V^{(l)}(\widehat{\kappa}_0^{(l)})\},$$

to compare the policy learned based on our selected model with those by assuming the system is a k-th order MDP. We report these value differences with different choices of N in Figure 3. It can be seen that $\text{VD}(k)$ is smaller than or close to zero in almost all cases. When $k = 4$, the value differences are very close to zero for large N. This suggests that our method is useful in identifying the optimal policy in HMDPs.

5.2.3 real data analysis

The lengths of trajectories in the OhioT1DM dataset range from 1119 to 1288. To implement our test, we set $T = 1100$ and apply Algorithm 1 to test whether the system is a k-th order MDP. The corresponding p-values are reported in Table 1. To apply Algorithm 2 for model selection, we set $\alpha = 0.01$. Our algorithm stops after the fourth iteration. The first four p-values are 0, 0, 0.001 and 0.068, respectively. Thus, we conclude the system is a 4-th order MDP.

Next, we use cross-validation to evaluate our selected model. Specifically, we split the six trajectories into training and testing sets, with each containing three trajectories. This yields a total of $L = \binom{6}{3} = 20$ combinations. Then for each combination and $k \in \{1, \cdots, 10\}$, we apply FQI to learn an optimal policy based on the training dataset by assuming the system is a k-th order MDP and apply the Fitted Q evaluation algorithm Le et al. (2019) on the testing dataset to evaluate its value (see Appendix B.3 for details). Finally, we aggregated these values over different combinations and report them in Table 1. It can be seen that the policy learned based on our selected model achieves the largest value.

5.3 Applications in POMDP problems

We apply our test to the Tiger problem (Cassandra et al., 1994). The model is defined as follows: at the initial time point, a tiger is randomly placed behind either the left or the right door with equal probability. At each time point, the agent can select from one of the following three actions: (i) open the left door; (ii) open the right; (iii) listen for tiger noises. But listening is not entirely accurate. If the agent chooses to listen, it will receive an observation $S_{0,t}$ that corresponds to the estimated location of the tiger. Let $H_{0,t}$ denote the observed correct location of the tiger, we have $\mathbb{P}(H_{0,t} = S_{0,t}) = 0.7$ and $\mathbb{P}(H_{0,t} \neq S_{0,t}) = 0.3$. If the agent chooses to open one of two doors, it receives a penalty of -100 if the tiger is behind that door or a reward $R_{0,t}$ of +10 otherwise. The game is then terminated.

We set T to 20. To generate the data, the behaviour policy is set to listening at time points $t = 0, 1, 2, \cdots, T-1$ and randomly choosing a door to open with equal probability at time T. For each simulation, we generate a total of N trajectories and then apply Algorithm 1 to the data $\{(S_{j,t}(k), A_{j,t}(k))\}_{1\le j\le N, 0\le t\le T-k+1}$ for $k = 1, \ldots, 10$. The empirical rejection rates with $N = 50, 100$ and 200 and the significance level $\alpha = 0.05$ and 0.1 are reported in the top plots of Figure 4. It can be seen that our test has nonnegligible powers for detecting POMDPs. Take $\alpha = 0.1$ as an example. The rejection rate is well above 50% in almost all cases. Moreover, the power of our test increases as either N increases or k decreases, as expected.

To evaluate the validity our test in this setting, we define a new state vector $S^*_{0,t} = (S_{0,t}, H_{0,t})^\top$ and repeat the above experiment with this new state. Since the hidden variable is included in the state vector, the Markov property is satisfied. The empirical rejection rates with different combinations of N, α and k are reported in the bottom plots of Figure 4. It can be seen that the Type I error rates are well-controlled in almost all cases.

References

Belloni, A. and Oliveira, R. I. A high dimensional central limit theorem for martingales, with applications to context tree models. *arXiv preprint arXiv:1809.02741*, 2018.

Bercu, B. and Touati, A. Exponential inequalities for self-normalized martingales with applications. *Ann. Appl. Probab.*, 18(5):1848–1869, 2008. ISSN 1050-5164. doi: 10.1214/07-AAP506.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2020.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*, 2018.

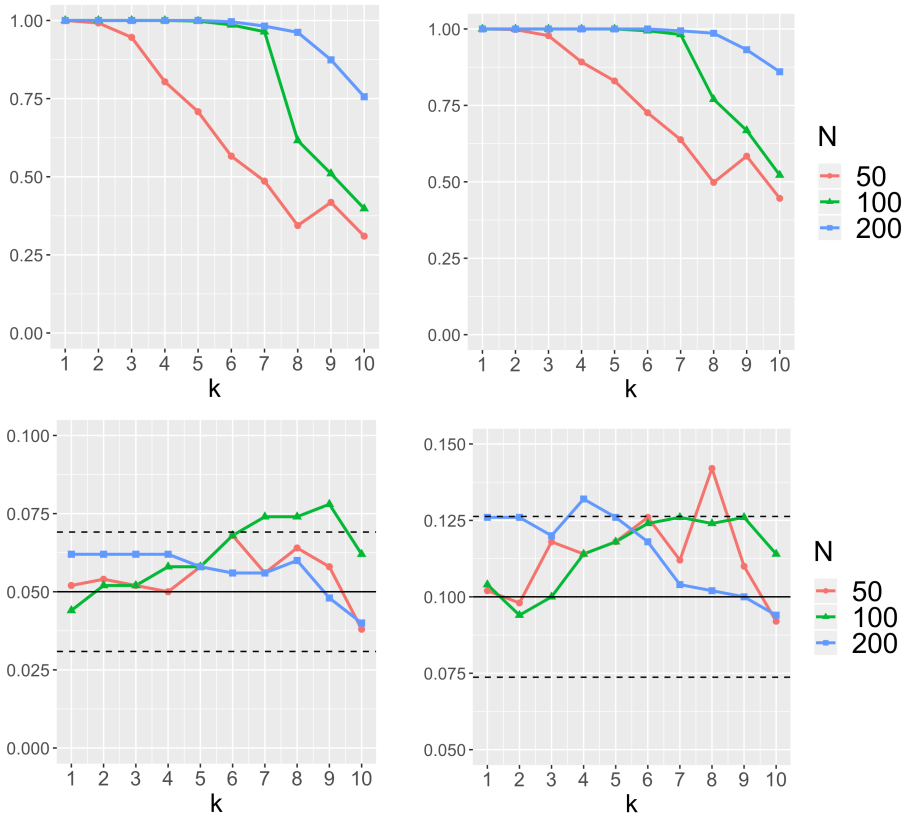

Figure 4: Empirical rejection rates aggregated over 500 simulations with different combinations of α, K and N. $\alpha = (0.05, 0.1)$ from left plots to right plots. H_1 holds in top plots. H_0 holds in bottom plots. Dashed lines correspond to $y = \alpha \pm 1.96\text{MCE}$ where MCE denotes the Monte Carlo error $\sqrt{\alpha(1-\alpha)/500}$.

Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13:1063–1095, 2012. ISSN 1532-4435.

Bradley, R. C. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2: 107–144, 2005. ISSN 1549-5787. doi: 10.1214/154957805100000104. Update of, and a supplement to, the 1986 original.

Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001.

Cassandra, A. R., Kaelbling, L. P., and Littman, M. L. Acting optimally in partially observable stochastic domains. In *AAAI*, volume 94, pp. 1023–1028, 1994.

Chen, B. and Hong, Y. Testing for the Markov property in time series. *Econometric Theory*, 28(1):130–178, 2012. ISSN 0266-4666. doi: 10.1017/S0266466611000065.

Chen, X. and Christensen, T. M. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *J. Econometrics*, 188(2):447–465, 2015. ISSN 0304-4076. doi: 10.1016/j.jeconom.2015.03.010.

Chernozhukov, V., Chetverikov, D., and Kato, K. Detailed proof of nazarov's inequality. *arXiv preprint arXiv:1711.10696*, 2017.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econom. J.*, 21(1):C1–C68, 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

Härdle, W. *Applied nonparametric regression*, volume 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, 1990. ISBN 0-521-38248-3. doi: 10.1017/CCOL0521382483.

Hausknecht, M. and Stone, P. Deep recurrent q-learning for partially observable mdps. In *2015 AAAI Fall Symposium Series*, 2015.

Huang, M., Sun, Y., and White, H. A flexible nonparametric test for conditional independence. *Econometric Theory*, 32(6):1434–1482, 2016. ISSN 0266-4666. doi: 10.1017/S0266466615000286.

Kalisch, M. and Bühlmann, P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

Le, H. M., Voloshin, C., and Yue, Y. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.

Marling, C. and Bunescu, R. C. The ohiot1dm dataset for blood glucose level prediction. In *KHD@ IJCAI*, pp. 60–63, 2018a.

Marling, C. and Bunescu, R. C. The ohiot1dm dataset for blood glucose level prediction. In *KHD@ IJCAI*, pp. 60–63, 2018b.

Meinshausen, N. Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999, 2006. ISSN 1532-4435.

Mengersen, K. L. and Tweedie, R. L. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121, 1996. ISSN 0090-5364. doi: 10.1214/aos/1033066201.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937, 2016.

Pearl, J. *Causality*. Cambridge University Press, Cambridge, 2000. ISBN 0-521-77362-8. Models, reasoning, and inference.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1994. ISBN 0-471-61977-9. A Wiley-Interscience Publication.

Riedmiller, M. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, pp. 317–328. Springer, 2005.

Rodbard, D. Interpretation of continuous glucose monitoring data: glycemic variability and quality of glycemic control. *Diabetes technology & therapeutics*, 11(S1):S–55, 2009.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, To appear, 2020.

Stone, C. J. Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–645, 1977. ISSN 0090-5364. With discussion and a reply by the author.

Su, L. and White, H. Testing conditional independence via empirical likelihood. *J. Econometrics*, 182(1):27–44, 2014. ISSN 0304-4076. doi: 10.1016/j.jeconom.2014.04.006.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: an introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2018. ISBN 978-0-262-03924-6.

Tierney, L. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–1762, 1994. ISSN 0090-5364. doi: 10.1214/aos/1176325750. With discussion and a rejoinder by the author.

Tsao, C. W. and Vasan, R. S. Cohort profile: The framingham heart study (fhs): overview of milestones in cardiovascular epidemiology. *International journal of epidemiology*, 44(6):1800–1813, 2015.

Van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

Wang, X. and Hong, Y. Characteristic function based testing for conditional independence: a nonparametric regression approach. *Econometric Theory*, 34(4):815–849, 2018. ISSN 0266-4666. doi: 10.1017/S026646661700010X.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. Conditional distance correlation. *J. Amer. Statist. Assoc.*, 110 (512):1726–1734, 2015. ISSN 0162-1459. doi: 10.1080/01621459.2014.993081.

Wu, C.-F. J. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1350, 1986. ISSN 0090-5364. doi: 10.1214/aos/1176350142. With discussion and a rejoinder by the author.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, pp. 8665–8675, 2019.

A Additional details regarding our test

A.1 The covariance estimator $\widehat{\Sigma}^{(q)}$

For any $\ell = 1, \cdots, \mathbb{L}$, $j \in \mathcal{I}^{(\ell)}$ and $0 < t < T - q$, define vectors $\lambda_{R,q,j,t}, \lambda_{I,q,j,t} \in \mathbb{R}^{\mathbb{B}}$ such that the b-th element of $\lambda_{R,q,j,t}, \lambda_{I,q,j,t}$ correspond to the real and imaginary part of

$$\{\exp(i\mu^\top S_{j,t+q+1}) - \widehat{\varphi}^{(-\ell)}(\mu|X_{j,t+q})\}\{\exp(i\nu^\top X_{j,t-1}) - \widehat{\psi}^{(-\ell)}(\nu|X_{j,t})\},$$

respectively. The matrix $\widehat{\Sigma}^{(q)}$ is defined by

$$\sum_{\ell=1}^{\mathbb{L}} \sum_{j \in \mathcal{I}^{(\ell)}} \sum_{t=1}^{T-q-1} \frac{(\lambda_{R,q,j,t}^\top, \lambda_{I,q,j,t}^\top)^\top (\lambda_{R,q,j,t}^\top, \lambda_{I,q,j,t}^\top)}{n(T-q-1)}. \tag{11}$$

A.2 Validity of our test without the stationary assumption

When (C2) is violated, the relation $\psi_1 = \psi_2 = \cdots = \psi_{T-1}$ might no longer hold. However, under (C1), (C3) and H_0, the marginal distribution function of $X_{0,t}$ can be well-approximated by some $\mathbb{F}$ on average. As a result, ψ_t's can be well-approximated by some ψ^* on average. Let $\mathbb{F}_t$ denote the distribution function of $X_{0,t}$. As long as the prediction error satisfies

$$\max_{1 \le b \le B} \frac{1}{T} \sum_{t=1}^{T} \int_x |\widehat{\psi}^{(-\ell)}(\nu_b|x) - \psi_t(\nu_b|x)|^2 \mathbb{F}_t(dx) = O_p((nT)^{-c_0}),$$

for some $c_0 > 1/2$, our test remains valid.

B More on the OhioT1DM dataset

B.1 Detailed definitions of actions and rewards

We define $A_{0,t}$ as follows:

$$A_{0,t} = \begin{cases} 0, & \text{In}_{0,t} = 0; \\ m, & 4m - 4 < \text{In}_{0,t} \le 4m \quad (m = 1, 2, 3); \\ 4, & \text{In}_{0,t} > 12. \end{cases}$$

The Index of Glycemic Control is chosen as the immediate reward $R_{0,t}$, defined by

$$R_{0,t} = \begin{cases} -\frac{1}{30}(80 - \text{G}_{0,t+1})^2, & G_{0,t+1} < 80; \\ 0, & 80 \le \text{G}_{0,t+1} \le 140; \\ -\frac{1}{30}(\text{G}_{0,t+1} - 140)^{1.35}, & 140 \le G_{0,t+1}. \end{cases}$$

B.2 Detailed procedure for value evaluation in simulations

In Section 5.2.2, we compare the policies learned with the selected order $\widehat{\kappa}_0$ and fixed orders $k \in \{1, \cdots, 10\}$. Below, we provide more details on computing the value $V^{(l)}(k)$.

1. In the l-th simulation, generate N trajectories $\{(S_{j,t}, A_{j,t})\}_{1 \le j \le N, 0 \le t \le 1344}$, and apply Algorithm 2 with $\alpha = 0.01$ and $K = 10$ to estimate an order $\widehat{\kappa}_0^{(l)}$. Also generate 100 trajectories of length 10 with the model described in Section 5.2.2, denoted by $\{(S_{j,t}^e, A_{j,t}^e)\}_{1 \le j \le 100, 0 \le t < 10}$.
2. For $k = 1, \ldots, 10$, apply FQI (see below) to the concatenated data $\{(S_{j,t}(k), A_{j,t}(k), R_{j,t}(k))\}_{1 \le j \le N, 0 \le t \le 1344-k}$ to learn an optimal policy $\widehat{\pi}^{(l)}(k)$.
3. For each initial trajectory $\{(S_{j,t}^e, A_{j,t}^e)\}_{0 \le t < 10}$, generate the data $\{(S_{j,t}^e, A_{j,t}^e, R_{j,t}^e)\}_{10 \le t < 60}$ following $\widehat{\pi}^{(l)}(k)$. Compute the value $V^{(l)}(k)$ by

$$V^{(l)}(k) = \frac{1}{100} \sum_{j=1}^{100} \sum_{t=10}^{50} \gamma^{t-10} R_{j,t}^e,$$

Algorithm 3 Fitted-Q iteration

Input: Data $\{S_{j,t}, A_{j,t}, R_{j,t}, S_{j,t+1}\}_{j,t}$, function class $\mathcal{F}$, decay rate γ, action space $\mathcal{A}$
Randomly pick $Q_0 \in \mathcal{F}$
for $k = 1, \ldots, K$ **do**
  Update target values $Z_{j,t} = R_{j,t} + \gamma \max_{a\in\mathcal{A}} Q_{k-1}(S_{j,t+1}, a)$ for all (j,t);
  Solve a regression problem to update the Q-function:
    $Q_k = \arg\min_{Q\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{Q(S_{j,t}, A_{j,t}) - Z_{j,t}\}^2$
end for
Output: The estimated optimal policy $\widehat{\pi}(\cdot) = \arg\max_{a\in\mathcal{A}} Q_K(\cdot, a)$

with $\gamma = 0.9$.

In our experiment,we use random forests to estimate the Q function during each iteration. The number of trees are set as 100 and the other hyperparameters are selected by 5-fold cross-validation. The decay rate γ is set to 0.9.

B.3 Detailed procedure for value evaluation in real data analysis

In Section 5.2.3, we compare policies learned by assuming the data follows a k-th order MDP for $k \in \{1, \cdots, 10\}$. The policies are estimated by FQI. To evaluate the values of these policies based on the real dataset, we apply the Fitted-Q evaluation (FQE) algorithm. Similar to FQI, it is an iterative algorithm based on the Bellman equation. We recap the steps below.

Algorithm 4 Fitted-Q evaluation

Input: Data $\{S_{j,t}, A_{j,t}, R_{j,t}, S_{j,t+1}\}_{j,t}$, policy π, function class $\mathcal{F}$, decay rate γ
Randomly pick $Q_0 \in \mathcal{F}$
for $k = 1, \ldots, K$ **do**
  Update target values $Z_{j,t} = R_{j,t} + \gamma Q_{k-1}(S_{j,t+1}, \pi(S_{j,t+1}))$ for all (j,t);
  Solve a regression problem to update the Q-function:
    $Q_k = \arg\min_{Q\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{Q(S_{j,t}, A_{j,t}) - Z_{j,t}\}^2$
end for
Output: The estimated value $\widehat{V}(\cdot) = Q_K(\cdot, \pi(\cdot))$

Denote the trajectories for the six patients in the OhioT1DM dataset by $\{(S_{i,t}, A_{i,t})\}_{1\le i\le 6, 1\le t\le 1100}$, and let the index set $\mathcal{I} = \{1, 2, 3, 4, 5, 6\}$. We now describe the evaluation procedure in more details:

1. In $l = 1, \ldots, 20$, divide $\mathcal{I}$ into a training set $\mathcal{D}_1^{(l)}$ and an validation set $\mathcal{D}_2^{(l)} = (\mathcal{D}_1^{(l)})^c$ with $|\mathcal{D}_1^{(l)}| = |\mathcal{D}_2^{(l)}| = 3$.
2. For each $l \in \{1, \ldots, 20\}$, $k \in \{1, \ldots, 10\}$, apply FQI to the data $\{(S_{j,t}(k), A_{j,t}(k), R_{j,t}(k))\}_{j\in\mathcal{D}_1^{(l)}, 0\le t\le 1100-k+1}$ to learn an optimal policy $\widehat{\pi}^{(l)}(k)$.
3. For each $l \in \{1, \ldots, 20\}$, $k \in \{1, \ldots, 10\}$, apply FQE to the data $\{(S_{j,t}(k), A_{j,t}(k), R_{j,t}(k))\}_{j\in\mathcal{D}_2^{(l)}, 0\le t\le 1100-k+1}$ to estimate the state-value function of $\widehat{\pi}^{(l)}(k)$, denoted by $\widehat{V}_k^{(l)}(\cdot)$. Generate 100 trajectories of length 10 according to the simulation model in Section 5.2.2. Denote them by $\{(S_{j,t}^e, A_{j,t}^e)\}_{1\le j\le 100, 0\le t<10}$. Calculate the value under $\widehat{\pi}^{(l)}(k)$ by
$$V^{(l)}(k) = \frac{1}{100}\sum_{j=1}^{100} \widehat{V}_k^{(l)}(S_{j,(10-k)}^e(k)).$$
4. Average over the 20 splits to compute the average value for each k by $V(k) = \sum_{l=1}^{20} V^{(l)}(k)/20$.

For both FQI and FQE, we use random forests to estimate the regression function. The number of trees are set to 75 and the other hyperparameters are selected by 5-fold cross-validation. We set $\gamma = 0.9$ in our experiments.

C Technical proofs

C.1 Proof of Lemma 1

Consider a policy $\pi = \{\pi_t\}_{t\geq 0} \in$ HR. Suppose there exists some $\{\pi_t^*\}_{t\geq 0}$ such that $\pi_t(\cdot|\bar{\boldsymbol{S}}_{0,t}) = \pi_t^*(\cdot|S_{0,t})$ almost surely for any $t \geq 0$. We refer to such a policy π as a *Markov* policy. In addition, π is a *deterministic* policy if and only if $\pi_t(a|\bar{\boldsymbol{S}}_{0,t}) \in \{0,1\}$ almost surely for any $t \geq 0$ and $a \in \mathcal{A}$. Let MR denotes the set of Markov policies and SD denote the set of deterministic stationary policies, we have SD $\subseteq$ SR $\subseteq$ MR $\subseteq$ HR. In the following, we focus on proving

$$\sup_{\pi\in\text{HR}} V(\pi;s) = \sup_{\pi\in\text{SD}} V(\pi;s), \quad \forall s \in \mathbb{S}.$$

Since SD $\subseteq$ SR, the assertion in Lemma 1 is thus satisfied.

We begin by providing a sketch of the proof. Our proof is divided into three steps. In the first step, we show

$$\sup_{\pi\in\text{HR}} V(\pi;s) = \sup_{\pi\in\text{MR}} V(\pi;s), \quad \forall s \in \mathbb{S}.$$

To prove this, we show in Section C.1.1 that for any such $\pi \in$ HR and any s, there exists a Markov policy $\pi^* = \{\pi_t^*\}_{t\geq 0}$ where each π_t^* depends on $S_{0,t}$ only such that

$$\mathbb{P}^{\pi}(A_{0,t} = a, S_{0,t} \in \mathcal{S}|S_{0,0} = s) = \mathbb{P}^{\pi^*}(A_{0,t} = a, S_{0,t} \in \mathcal{S}|S_{0,0} = s), \tag{12}$$

for any $t \geq 0$, $a \in \mathcal{A}$, $\mathcal{S} \subseteq \mathbb{S}$ and $s \in \mathbb{S}$ where the probabilities $\mathbb{P}^{\pi}$ and $\mathbb{P}^{\pi^*}$ are taken by assuming the system dynamics follow π and π^*, respectively. Under MA, we have

$$\mathbb{E}^{\pi}(Y_{0,t}|S_{0,0} = s) = \mathbb{E}^{\pi}\{\mathbb{E}^{\pi}(Y_{0,t}|A_{0,t}, S_{0,t}, S_{0,0} = s)|S_{0,0} = x\} = \mathbb{E}^{\pi}\{r(A_{0,t}, S_{0,t})|S_{0,0} = x\},$$

for some function r. This together with (12) yields that

$$\mathbb{E}^{\pi}(Y_{0,t}|S_{0,0} = s) = \mathbb{E}^{\pi^*}(Y_{0,t}|S_{0,0} = s), \quad \forall t \geq 0,$$

and hence $V(\pi;s) = V(\pi^*;s)$. This completes the proof for the first step.

With a slight abuse of notation, for any $\pi \in$ SD, we denote by $\pi(s)$ the action that the agent chooses according to π, given that the current state equals s. In the second step, we show for any bounded function $\nu(\cdot)$ on $\mathbb{S}$ that satisfies the optimal Bellman equation

$$\nu(s) = \sup_{\pi\in\text{SD}} \left\{ r(\pi(s), s) + \gamma \int_{s'} \nu(s')\mathcal{P}(ds';\pi(s),s) \right\}, \quad \forall s \in \mathbb{S},$$

it satisfies

$$\nu(s) = \sup_{\pi^*\in\text{MR}} V(\pi^*;s), \quad \forall s \in \mathbb{S}. \tag{13}$$

The proof of (13) is given in Section C.1.2.

For any function ν, define the norm $\|\nu\|_\infty = \sup_{s\in\mathbb{S}} |\nu(s)|$. We have for any ν_1 and ν_2 that

$$\begin{aligned}
\sup_x \left| \sup_{\pi\in\text{SD}} \left\{ r(\pi(s),s) + \gamma \int_{s'} \nu_1(s')\mathcal{P}(ds';\pi(s),s) \right\} - \sup_{\pi\in\text{SD}} \left\{ r(\pi(s),s) + \gamma \int_{s'} \nu_2(s')\mathcal{P}(ds';\pi(s),s) \right\} \right| \\
\leq \gamma \sup_{\pi\in\text{SD}} \sup_{s\in\mathbb{S}} \left| \int_{s'} \nu_1(s')\mathcal{P}(ds';\pi(s),s) - \int_{s'} \nu_2(s')\mathcal{P}(ds';\pi(s),s) \right| \\
\leq \gamma \sup_{\pi\in\text{SD}} \sup_{s\in\mathbb{S}} \left| \int_{s'} \|\nu_1 - \nu_2\|_\infty \mathcal{P}(ds';\pi(s),s) \right| \leq \gamma \|\nu_1 - \nu_2\|_\infty.
\end{aligned}$$

By Banach's fix point theorem, there exists a unique value function ν_0 that satisfies the optimal Bellman equation. Combining this together with the results obtained in the first two steps, we obtain that ν_0 satisfies $\nu_0(s) = \sup_{\pi\in\text{HR}} V(\pi;s)$ for any $s \in \mathbb{S}$. The proof is thus completed if we can show there exists a deterministic stationary policy π^{**} that satisfies

$$\nu_0(s) = V(\pi^{**};s), \quad \forall s \in \mathbb{S}. \tag{14}$$

We put the proof of (14) in Section C.1.3.

C.1.1 Proof of (12)

Apparently, (12) holds with $t = 0$. Suppose (12) holds for $t = k$. We now show (12) holds for $t = k + 1$. Under MA, we have

$$\begin{aligned}&\mathbb{P}^{\pi}(S_{0,k+1} \in \mathcal{S}|S_{0,0} = s) = \mathbb{E}^{\pi}\{\mathbb{P}^{\pi}(S_{0,t+1} \in \mathcal{S}|A_{0,t}, S_{0,t}, S_{0,0} = s)|S_{0,0} = x\}\\ = \quad &\mathbb{E}^{\pi}\{\mathcal{P}(\mathcal{S}; A_{0,t}, S_{0,t})|S_{0,0} = x\} = \mathbb{E}^{\pi^*}\{\mathcal{P}(\mathcal{S}; A_{0,t}, S_{0,t})|S_{0,0} = x\} = \mathbb{P}^{\pi^*}(S_{0,k+1} \in \mathcal{S}|S_{0,0} = s) \triangleq \mathbb{G}_{k+1}(\mathcal{S}; s).\end{aligned}$$

Set π^*_{k+1} to be the decision rule that satisfies

$$\mathbb{P}^{\pi}(A_{0,k+1} = a|S_{0,k+1}, S_{0,0} = s) = \mathbb{P}^{\pi^*_{k+1}}(A_{0,k+1} = a|S_{0,k+1}), \quad \forall a \in \mathcal{A},$$

it follows that

$$\begin{aligned}&\mathbb{P}^{\pi}(A_{0,k+1} = a, S_{0,k+1} \in \mathcal{S}|S_{0,0} = s) = \int_{s'} \mathbb{P}^{\pi}(A_{0,k+1} = a|S_{0,k+1} = s', S_{0,0} = s)\mathbb{G}_{k+1}(ds'; s)\\ = &\int_{s'} \mathbb{P}^{\pi^*}(A_{0,k+1} = a|S_{0,k+1} = s', S_{0,0} = s)\mathbb{G}_{k+1}(ds'; s) = \mathbb{P}^{\pi^*}(A_{0,k+1} = a, S_{0,k+1} \in \mathcal{S}|S_{0,0} = s).\end{aligned}$$

Thus, (12) holds for $t = k + 1$ as well. The proof is hence completed.

C.1.2 Proof of (13)

We first show for any bounded function ν that satisfies

$$\nu(s) \geq \sup_{\pi \in \text{SD}} \left\{ r(\pi(s), s) + \gamma \int_{s'} \nu(s')\mathcal{P}(ds'; \pi(s), s) \right\}, \quad \forall s \in \mathbb{S}, \tag{15}$$

we have

$$\nu(s) \geq \sup_{\pi^* \in \text{MR}} V(\pi^*; s), \quad \forall s \in \mathbb{S}. \tag{16}$$

Then, we show for any bounded function ν that satisfies

$$\sup_{s \in \mathbb{S}} \left[\nu(s) - \sup_{\pi \in \text{SD}} \left\{ r(\pi(s), s) + \gamma \int_{s'} \nu(s')\mathcal{P}(ds'; \pi(s), s) \right\} \right] \leq 0,$$

we have

$$\nu(s) \leq \sup_{\pi^* \in \text{MR}} V(\pi^*; s), \quad \forall s \in \mathbb{S}. \tag{17}$$

The proof is hence completed.

Proof of (16)*:* Consider an arbitrary deterministic Markov policy $\pi^* = \{\pi^*_t\}_{t \geq 0}$. With a slight abuse of notation, we denote by $\pi^*_t(s)$ the action that the agent chooses following π^*_t, given that the current state equals s. It follows from (15) that

$$\nu(s) \geq r(\pi^*_0(s), s) + \gamma \int_{s'} \nu(s')\mathcal{P}(ds'; \pi^*_0(s), s), \quad \forall s \in \mathbb{S}.$$

By iteratively applying (15), we have

$$\nu(s) \geq r(\pi^*_0(s), s) + \sum_{k=1}^{K} \gamma^k \mathbb{E}^{\pi^*}\{r(A_{0,k}, X_{0,k})|S_{0,0} = x\} + \gamma^{K+1}\mathbb{E}^{\pi^*}\{\nu(X_{0,K+1})|S_{0,0} = x\}, \quad \forall s \in \mathbb{S}.$$

Since ν is bounded, the last term on the right-hand-side (RHS) converges to zero uniformly in x, as $K \to \infty$. Let $K \to \infty$, we obtain $\nu(s) \geq V(\pi^*; s)$, for any $s \in \mathbb{S}$ and any deterministic Markov policy π^*. Using Lemma 4.3.1 of Puterman (1994), we can similarly show $\nu(s) \geq V(\pi^*; s)$ for any $s \in \mathbb{S}$ and $\pi^* \in$ MR. This completes the proof of (16).

Proof of (17)*:* By definition, we have

$$\inf_{\pi \in \text{SD}} \sup_{s \in \mathbb{S}} \left[\nu(s) - \left\{ r(\pi(s), s) + \gamma \int_{s'} \nu(s')\mathcal{P}(ds'; \pi(s), s) \right\} \right] \leq 0.$$

Thus, for any $\epsilon > 0$, there exists some $\pi_0 \in \text{SD}$ that satisfies

$$\sup_{s\in\mathbb{S}}\left[\nu(s) - \left\{r(\pi_0(s), s) + \gamma\int_{s'}\nu(s')\mathcal{P}(ds';\pi_0(s), s)\right\}\right] \le \epsilon. \tag{18}$$

Consider the following bounded linear operator $\mathcal{L}_0$,

$$\mathcal{L}_0\nu(s) = \int_{s'}\nu(s')\mathcal{P}(ds';\pi_0(s), s),$$

defined on the space of bounded functions. Let $\mathcal{I}_0$ denote the identity operator. Since $\gamma < 1$, the operator $\mathcal{I}_0 - \gamma\mathcal{L}_0$ is invertible and its inverse equals $\sum_{k=0}^{+\infty}\gamma^k\mathcal{L}_0^k$. It follows from (18) that

$$\nu(s) \le \sum_{k=0}^{+\infty}\gamma^k\mathcal{L}_0^k\{r(\pi_0(s), s) + \epsilon\}, \qquad \forall s \in \mathbb{S}.$$

Since $V(\pi_0; s) = \sum_{k=0}^{+\infty}\gamma^k\mathcal{L}_0^k r(\pi_0(s), s)$ and $\sum_{k=0}^{+\infty}\gamma^k\mathcal{L}_0^k\epsilon \le \epsilon/(1-\gamma)$, we obtain

$$\nu(s) \le V(\pi_0; s) + \frac{\epsilon}{1-\gamma}.$$

Let $\epsilon \to 0$, we obtain $\nu(s) \le \sup_{\pi^*\in\text{MR}} V(\pi^*; s)$ for any x. The proof is hence completed.

C.1.3 Proof of (14)

Since $\nu_0(\cdot)$ satisfies the optimal Bellman equation, we have

$$\nu_0(s) = \arg\max_{\pi\in\text{SD}}\left\{r(\pi(s), s) + \gamma\int_{s'}\nu_0(s')\mathcal{P}(ds';\pi(s), s)\right\}.$$

Let $\mathcal{A}_s$ be the available set of actions at a given state s. As a result, we have

$$\nu_0(s) = \arg\max_{a\in\mathcal{A}_s}\left\{r(a, s) + \gamma\int_{s'}\nu_0(s')\mathcal{P}(ds';a, s)\right\}.$$

Since $\mathcal{A}$ is finite, so is $\mathcal{A}_s$. As a result, the above argmax is achievable. Let $\pi^{**}(s)$ be the action such that the above argmax is achieved, we have

$$\nu_0(s) = r(\pi^{**}(s), s) + \gamma\int_{s'}\nu_0(s')\mathcal{P}(ds';\pi^{**}(s), s).$$

Similar to the proof of (13), we can show $\nu_0(s) = V(\pi^{**}; s)$, for all $s \in \mathbb{S}$. The proof is hence completed.

C.2 Proof of Theorem 1

The proof is divided into two parts. In the first part, we show (3) $\Rightarrow$ (5). In the second part, we show (5) $\Rightarrow$ (3).

C.2.1 Part 1

Under (3), $S_{0,t+q+1} \perp\!\!\!\perp \{X_{0,j}\}_{j<t+q}|X_{0,t+q}$. It follows that

$$\mathbb{E}[\exp(i\mu^\top S_{0,t+q+1} + i\nu^\top X_{0,t-1})|\{X_{0,j}\}_{t\le j\le t+q}] = \varphi_{t+q}(\mu|X_{0,t+q})\mathbb{E}[(i\nu^\top X_{0,t-1})|\{X_{0,j}\}_{t\le j\le t+q}].$$

The proof is hence completed.

C.2.2 Part 2

We introduce the following lemma before presenting the proof.

Lemma 3 *For any random vectors $Z_1 \in \mathbb{R}^{q_1}, Z_2 \in \mathbb{R}^{q_2}, Z_3 \in \mathbb{R}^{q_3}$, suppose $\mathbb{E}\{\exp(i\mu_1^\top Z_1)|Z_3\}\mathbb{E}\{\exp(i\mu_2^\top Z_2)|Z_3\} = \mathbb{E}\{\exp(i\mu_1^\top Z_1 + i\mu_2^\top Z_2)|Z_3\}$ for any $\mu_1 \in \mathbb{R}^{q_1}$, $\mu_2 \in \mathbb{R}^{q_2}$ almost surely. Then we have $Z_1 \perp\!\!\!\perp Z_2|Z_3$.*

Let $q = 0$. By (5), we obtain

$$\varphi_t(\mu|X_{0,t})\mathbb{E}\{\exp(i\nu^\top X_{0,t-1})|X_{0,t}\} = \mathbb{E}[\exp(i\mu^\top S_{0,t+1} + i\nu^\top X_{0,t-1})|X_{0,t}],$$

for any $t > 0$, $\mu \in \mathbb{R}^p$, $\nu \in \mathbb{R}^{p+1}$. By Lemma 3, we obtain

$$S_{0,t+1} \perp\!\!\!\perp X_{0,t-1}|X_{0,t}, \quad \forall t > 0. \tag{19}$$

Set $q = 1$, we have by (5) that

$$\varphi_{t+1}(\mu|X_{0,t+1})\mathbb{E}\{\exp(i\nu^\top X_{0,t-1})|X_{0,t}, X_{0,t+1}\} = \mathbb{E}[\exp(i\mu^\top S_{0,t+2} + i\nu^\top X_{0,t-1})|X_{0,t}, X_{0,t+1}], \tag{20}$$

for any $t > 0$, $\mu \in \mathbb{R}^p$, $\nu \in \mathbb{R}^{p+1}$. For any $v \in \mathbb{R}^{p+1}$, multiply both sides of (20) by $\exp(iv^\top X_{0,t})$ and take expectation with respect to $X_{0,t}$ conditional on $X_{0,t+1}$, we obtain

$$\mathbb{E}\{\exp(i\mu^\top S_{0,t+2})|X_{0,t+1}\}\mathbb{E}\{\exp(iv^\top X_{0,t-1} + i\nu^\top X_{0,t})|X_{0,t+1}\} = \mathbb{E}[\exp(i\mu^\top S_{0,t+2} + iv^\top X_{0,t-1} + i\nu^\top X_{0,t})|X_{0,t+1}].$$

By Lemma 3, we obtain

$$S_{0,t+2} \perp\!\!\!\perp X_{0,t-1}, X_{0,t}|X_{0,t+1}, \quad \forall t > 0. \tag{21}$$

Similarly, we can show

$$S_{0,t+q+1} \perp\!\!\!\perp \{S_{0,j}\}_{t-1\le j<t+q}|X_{0,t+q}, \quad \forall t. \tag{22}$$

Combining (19) with (21) and (22) yields (3). The proof is hence completed.

C.2.3 Proof of Lemma 3

Let $\widetilde{Z}_1, \widetilde{Z}_2$ be independent copies of Z_1, Z_2 such that $\widetilde{Z}_1|Z_3 \stackrel{d}{=} Z_1|Z_3$, $\widetilde{Z}_2|Z_3 \stackrel{d}{=} Z_2|Z_3$ and that $\widetilde{Z}_1 \perp\!\!\!\perp \widetilde{Z}_2|\widetilde{Z}_3$. Consider any $\mu_1 \in \mathbb{R}^{q_1}, \mu_2 \in \mathbb{R}^{q_2}, \mu_3 \in \mathbb{R}^{q_3}$, we have

$$\begin{aligned} &\mathbb{E}\exp(i\mu_1^\top \widetilde{Z}_1 + i\mu_2^\top \widetilde{Z}_2 + i\mu_3^\top Z_3) = \mathbb{E}[\exp(i\mu_3^\top Z_3)\mathbb{E}\{\exp(i\mu_1^\top \widetilde{Z}_1 + i\mu_2^\top \widetilde{Z}_2)|Z_3\}] \qquad (23)\\ = \;& \mathbb{E}[\exp(i\mu_3^\top Z_3)\mathbb{E}\{\exp(i\mu_1^\top \widetilde{Z}_1)|Z_3\}\mathbb{E}\{\exp(i\mu_2^\top \widetilde{Z}_2)|Z_3\}] = \mathbb{E}[\exp(i\mu_3^\top Z_3)\mathbb{E}\{\exp(i\mu_1^\top Z_1)|Z_3\}\mathbb{E}\{\exp(i\mu_2^\top Z_2)|Z_3\}]. \end{aligned}$$

Under the condition in Lemma 3, we have

$$\begin{aligned} \mathbb{E}[\exp(i\mu_3^\top Z_3)\mathbb{E}\{\exp(i\mu_1^\top Z_1)|Z_3\}\mathbb{E}\{\exp(i\mu_2^\top Z_2)|Z_3\}] &= \mathbb{E}[\exp(i\mu_3^\top Z_3)\mathbb{E}\{\exp(i\mu_1^\top Z_1 + i\mu_2^\top Z_2)|Z_3\}] \\ &= \mathbb{E}\exp(i\mu_1^\top Z_1 + i\mu_2^\top Z_2 + i\mu_3^\top Z_3). \end{aligned}$$

This together with (23) yields

$$\mathbb{E}\exp(i\mu_1^\top \widetilde{Z}_1 + i\mu_2^\top \widetilde{Z}_2 + i\mu_3^\top Z_3) = \mathbb{E}\exp(i\mu_1^\top Z_1 + i\mu_2^\top Z_2 + i\mu_3^\top Z_3).$$

As a result, $(\widetilde{Z}_1, \widetilde{Z}_2, Z_3)$ and (Z_1, Z_2, Z_3) have same characteristic functions. Therefore, we have $(\widetilde{Z}_1, \widetilde{Z}_2, Z_3) \stackrel{d}{=} (Z_1, Z_2, Z_3)$. By construction, we have $\widetilde{Z}_1 \perp\!\!\!\perp \widetilde{Z}_2|Z_3$. It follows that $Z_1 \perp\!\!\!\perp Z_2|Z_3$.

C.3 Proof of Theorem 3

We focus on proving Theorem 3 in the more challenging setting where $T \to \infty$. The number of trajectories n can be either bounded or growing to ∞. The case where T is bounded can be proven using similar arguments. We begin by providing an outline of the proof. For any q, μ, ν, define

$$\Gamma^*(q, \mu, \nu) = \frac{1}{n(T-q-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-q-1}\{\exp(i\mu^\top S_{j,t+q+1}) - \varphi^*(\mu|X_{j,t+q})\}\{\exp(i\nu^\top X_{j,t-1}) - \psi^*(\nu|X_{j,t})\}.$$

Denote by Γ_R^* and Γ_I^* the real and imaginary part of Γ^*, respectively.

We break the proof into three steps. In the first step, we show

$$\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\sqrt{n(T-q-1)}|\widehat{\Gamma}(q, \mu_b, \nu_b) - \Gamma^*(q, \mu_b, \nu_b)| = o_p(\log^{-1/2}(nT)). \tag{24}$$

Proof of (24) relies largely on Condition (C4) which requires $\widehat{\varphi}$ and $\widehat{\psi}$ to satisfy certain uniform convergence rates. This further implies that

$$\widehat{S} = S^* + o_p(\log^{-1/2}(nT)), \tag{25}$$

where

$$S^* = \max_{b\in\{1,\cdots,B\}} \max_{q\in\{0,\cdots,Q\}} \sqrt{n(T-q-1)}\max(|\Gamma_R^*(q,\mu_b,\nu_b)|, |\Gamma_I^*(q,\mu_b,\nu_b)|).$$

In the second step, we show for any $z \in \mathbb{R}$ and any sufficiently small $\varepsilon > 0$,

$$\mathbb{P}(S^* \le z) \ge \mathbb{P}(\|N(0,V_0)\|_\infty \le z - \varepsilon\log^{-1/2}(nT)) - o(1),$$
$$\mathbb{P}(S^* \le z) \le \mathbb{P}(\|N(0,V_0)\|_\infty \le z + \varepsilon\log^{-1/2}(nT)) + o(1),$$

where the matrix V_0 is defined in Step 2 of the proof. This together with (25) yields that

$$\mathbb{P}(\widehat{S} \le z) \ge \mathbb{P}(\|N(0,V_0)\|_\infty \le z - 2\varepsilon\log^{-1/2}(nT)) - o(1), \tag{26}$$
$$\mathbb{P}(\widehat{S} \le z) \le \mathbb{P}(\|N(0,V_0)\|_\infty \le z + 2\varepsilon\log^{-1/2}(nT)) + o(1). \tag{27}$$

The proposed Bootstrap algorithm repeatedly generate random variables from $\|N(0,\widehat{V})\|_\infty$ where the detailed form of $\widehat{V}$ is given in the third step of the proof. The critical values $\widehat{c}_\alpha$ is chosen to be the upper α-th quantile of $\|N(0,\widehat{V})\|_\infty$. In the third step, we show $\|V_0 - \widehat{V}\|_{\infty,\infty} = O((nT)^{-c^{**}})$ for some $c^{**} > 0$ with probability tending to 1, where $\|\cdot\|_{\infty,\infty}$ denotes the elementwise max-norm. Combining this upper bound with some arguments used in proving (26) and (27), we can show with probability tending to 1 that

$$\mathbb{P}(\widehat{S} \le z) \ge \mathbb{P}(\|N(0,\widehat{V})\|_\infty \le z - 2\varepsilon\log^{-1/2}(nT)|\widehat{V}) - o(1),$$
$$\mathbb{P}(\widehat{S} \le z) \le \mathbb{P}(\|N(0,\widehat{V})\|_\infty \le z + 2\varepsilon\log^{-1/2}(nT)|\widehat{V}) + o(1),$$

for any sufficiently small $\varepsilon > 0$ where $\mathbb{P}(\cdot|\widehat{V})$ denotes the conditional probability given $\widehat{V}$. Set $z = \widehat{c}_\alpha$. It follows from that

$$\mathbb{P}(\widehat{S} \le \widehat{c}_\alpha) \ge \mathbb{P}(\|N(0,\widehat{V})\|_\infty \le \widehat{c}_\alpha - 2\varepsilon\log^{-1/2}(nT)|\widehat{V}) - o(1), \tag{28}$$
$$\mathbb{P}(\widehat{S} \le \widehat{c}_\alpha) \le \mathbb{P}(\|N(0,\widehat{V})\|_\infty \le \widehat{c}_\alpha + 2\varepsilon\log^{-1/2}(nT)|\widehat{V}) + o(1), \tag{29}$$

with probability tending to 1. Under the given conditions in Theorem 3, the diagonal elements in V_0 are bounded away from zero. With probability tending to 1, the diagonal elements in $\widehat{V}$ is bounded away from zero as well. It follows from Theorem 1 of Chernozhukov et al. (2017) that conditional on $\widehat{V}$,

$$\begin{aligned}\mathbb{P}(\|N(0,\widehat{V})\|_\infty \le \widehat{c}_\alpha + 2\varepsilon\log^{-1/2}(nT)|\widehat{V}) - \mathbb{P}(\|N(0,\widehat{V})\|_\infty \le \widehat{c}_\alpha - 2\varepsilon\log^{-1/2}(nT)|\widehat{V})& \\ \le O(1)\varepsilon\log^{1/2}(BQ)\log^{-1/2}(nT),&\end{aligned}$$

with probability tending to 1, where $O(1)$ denotes some positive constant that is independent of ε. Under the given conditions on B and Q, we obtain with probability tending to 1 that,

$$\mathbb{P}(\|N(0,\widehat{V})\|_\infty \le \widehat{c}_\alpha + 2\varepsilon\log^{-1/2}(nT)|\widehat{V}) - \mathbb{P}(\|N(0,\widehat{V})\|_\infty \le \widehat{c}_\alpha - 2\varepsilon\log^{-1/2}(nT)|\widehat{V}) \le C^*\varepsilon,$$

for some constant $C^* > 0$. This together with (28) and (29) yields

$$|\mathbb{P}(\widehat{S} \le \widehat{c}_\alpha) - \mathbb{P}(\|N(0,\widehat{V})\|_\infty \le \widehat{c}_\alpha|\widehat{V})| \le C^*\varepsilon + o(1),$$

with probability tending to 1. Notice that ε can be made arbitrarily small. The validity of our test thus follows.

In the following, we present our proof for each of the step. Suppose $\{\mu_b,\nu_b\}_{1\le b\le B}$ are fixed throughout the proof. Denote by $\widehat{\varphi}_R^{(\ell)}$, $\widehat{\varphi}_I^{(\ell)}$ the real and imaginary part of $\widehat{\varphi}^{(\ell)}$ respectively. Without loss of generality, we assume the absolute values of $\widehat{\varphi}_R^{(\ell)}$, $\widehat{\varphi}_I^{(\ell)}$ are uniformly bounded by 1.

C.3.1 Step 1

With some calculations, we can show that for any q,μ,ν,

$$\widehat{\Gamma}(q,\mu,\nu) = \Gamma^*(q,\mu,\nu) + R_1(q,\mu,\nu) + R_2(q,\mu,\nu) + R_3(q,\mu,\nu),$$

where the remainder terms R_1, R_2 and R_3 are given by

$$R_1(q,\mu,\nu) = \frac{1}{n(T-q-1)}\sum_{\ell=1}^{\mathbb{L}}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T-q-1}\{\varphi^*(\mu|X_{j,t+q}) - \widehat{\varphi}^{(-\ell)}(\mu|X_{j,t+q})\}\{\psi^*(\nu|X_{j,t}) - \widehat{\psi}^{(-\ell)}(\nu|X_{j,t})\},$$

$$R_2(q,\mu,\nu) = \frac{1}{n(T-q-1)}\sum_{\ell=1}^{\mathbb{L}}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T-q-1}\{\exp(i\mu^\top S_{j,t+q+1}) - \varphi^*(\mu|X_{j,t+q})\}\{\psi^*(\nu|X_{j,t}) - \widehat{\psi}^{(-\ell)}(\nu|X_{j,t})\},$$

$$R_3(q,\mu,\nu) = \frac{1}{n(T-q-1)}\sum_{\ell=1}^{\mathbb{L}}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T-q-1}\{\varphi^*(\mu|X_{j,t+q}) - \widehat{\varphi}^{(-\ell)}(\mu|X_{j,t+q})\}\{\exp(i\nu^\top X_{j,t-1}) - \psi^*(\nu|X_{j,t})\}.$$

It suffices to show

$$\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\sqrt{n(T-q-1)}|R_m(q,\mu_b,\nu_b)| = o_p(\log^{-1/2}(nT)), \tag{30}$$

for $m = 1, 2, 3$. In the following, we show (30) holds with $m = 1, 2$. Using similar arguments, one can show (30) holds with $m = 3$.

Proof of (30) **with** $m = 1$**:** Since $\mathbb{L}$ is fixed, it suffices to show

$$\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\sqrt{n(T-q-1)}|R_{1,\ell}(q,\mu_b,\nu_b)| = o_p(\log^{-1/2}(nT)), \tag{31}$$

where $R_{1,\ell}(q,\mu_b,\nu_b)$ is defined by

$$\frac{1}{n(T-q-1)}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T-q-1}\{\varphi^*(\mu_b|X_{j,t+q}) - \widehat{\varphi}^{(-\ell)}(\mu_b|X_{j,t+q})\}\{\psi^*(\nu_b|X_{j,t}) - \widehat{\psi}^{(-\ell)}(\nu_b|X_{j,t})\}.$$

Similarly, let φ_R^* and φ_I^* denote the real and imaginary part of φ^*. We can rewrite $R_{1,\ell}(q,\mu_b,\nu_b)$ as $R_{1,\ell}^{(1)}(q,\mu_b,\nu_b) - R_{1,\ell}^{(2)}(q,\mu_b,\nu_b) + iR_{1,\ell}^{(3)}(q,\mu_b,\nu_b) + iR_{1,\ell}^{(4)}(q,\mu_b,\nu_b)$ where

$$R_{1,\ell}^{(1)}(q,\mu_b,\nu_b) = \frac{1}{n(T-q-1)}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T-q-1}\{\varphi_R^*(\mu_b|X_{j,t+q}) - \widehat{\varphi}_R^{(-\ell)}(\mu_b|X_{j,t+q})\}\{\psi_R^*(\nu_b|X_{j,t}) - \widehat{\psi}_R^{(-\ell)}(\nu_b|X_{j,t})\},$$

$$R_{1,\ell}^{(2)}(q,\mu_b,\nu_b) = \frac{1}{n(T-q-1)}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T-q-1}\{\varphi_I^*(\mu_b|X_{j,t+q}) - \widehat{\varphi}_I^{(-\ell)}(\mu_b|X_{j,t+q})\}\{\psi_I^*(\nu_b|X_{j,t}) - \widehat{\psi}_I^{(-\ell)}(\nu_b|X_{j,t})\},$$

$$R_{1,\ell}^{(3)}(q,\mu_b,\nu_b) = \frac{1}{n(T-q-1)}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T-q-1}\{\varphi_R^*(\mu_b|X_{j,t+q}) - \widehat{\varphi}_R^{(-\ell)}(\mu_b|X_{j,t+q})\}\{\psi_I^*(\nu_b|X_{j,t}) - \widehat{\psi}_I^{(-\ell)}(\nu_b|X_{j,t})\},$$

$$R_{1,\ell}^{(4)}(q,\mu_b,\nu_b) = \frac{1}{n(T-q-1)}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T-q-1}\{\varphi_R^*(\mu_b|X_{j,t+q}) - \widehat{\varphi}_R^{(-\ell)}(\mu_b|X_{j,t+q})\}\{\psi_I^*(\nu_b|X_{j,t}) - \widehat{\psi}_I^{(-\ell)}(\nu_b|X_{j,t})\}.$$

To prove (31), it suffices to show

$$\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\sqrt{n(T-q-1)}|R_{1,\ell}^{(s)}(q,\mu_b,\nu_b)| = o_p(\log^{-1/2}(nT)), \tag{32}$$

for $s = 1, 2, 3, 4$. For brevity, we only show (32) holds with $s = 1$.

By the Cauchy-Schwarz inequality, it suffices to show

$$\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\frac{1}{\sqrt{n(T-q-1)}}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T}\{\varphi_R^*(\mu_b|X_{j,t}) - \widehat{\varphi}_R^{(-\ell)}(\mu_b|X_{j,t})\}^2 = o_p(\log^{-1/2}(nT)), \tag{33}$$

$$\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\frac{1}{\sqrt{n(T-q-1)}}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T}\{\psi_R^*(\nu_b|X_{j,t}) - \widehat{\psi}_R^{(-\ell)}(\nu_b|X_{j,t})\}^2 = o_p(\log^{-1/2}(nT)). \tag{34}$$

In the following, we focus on proving (33). Proof of (34) is similar and is thus omitted.

Under (C2) and (C3), it follows from Theorem 3.7 of Bradley (2005) that $\{X_{0,t}\}_{t\geq 0}$ is exponentially β-mixing, that is, the β-mixing coefficient of $\{X_{0,t}\}_{t\geq 0}$ $\beta_0(\cdot)$ satisfies $\beta_0(t) = O(\rho^t)$ for some $\rho < 1$ and any $t \geq 0$. Let $n_0 = |\mathcal{I}^{(\ell)}| = n/\mathbb{L}$ and suppose $\mathcal{I}^{(\ell)} = \{\ell_1, \ell_2, \cdots, \ell_{n_0}\}$. Since $\{X_{\ell_1,t}\}_{t\geq 0}, \{X_{\ell_2,t}\}_{t\geq 0}, \cdots, \{X_{\ell_{n_0},t}\}_{t\geq 0}$ are i.i.d copies of $\{X_{0,t}\}_{t\geq 0}$, the β-mixing coefficient of

$$\{X_{\ell_1,1}, X_{\ell_1,2}, \cdots, X_{\ell_1,T}, X_{\ell_2,1}, X_{\ell_2,2}, \cdots, X_{\ell_2,T}, \cdots, X_{\ell_{n_0},1}, X_{\ell_{n_0},2}, \cdots, X_{\ell_{n_0},T}\}$$

satisfies $\beta(t) = O(\rho^t)$ for any $t \geq 0$ as well.

Let $\phi_{j,t,b}$ denote $\varphi_R^*(\mu_b|X_{j,t}) - \widehat{\varphi}_R^{(-\ell)}(\mu_b|X_{j,t})$. By (C2), we have

$$\max_{j,t,b}\mathbb{E}^{X_{j,t}}\phi_{j,t,b}^4 \leq 4\max_{b\in\{1,\cdots,B\}}\int_x\{\varphi_R^*(\mu_b|x) - \widehat{\varphi}_R^{(-\ell)}(\mu_b|x)\}^2\mathbb{F}(dx) \equiv \Delta, \tag{35}$$

where the expectation $\mathbb{E}^{X_{j,t}}$ is taken with respect to $X_{j,t}$. Notice that Δ is a random variable that depends on $\{\mu_b, \nu_b\}_{1\le b\le B}$ and $\{X_{j,t}\}_{j\in\mathcal{I}^{(-\ell)},0\le t\le T}$. By (35), we have

$$\max_{j,t,b} \mathbb{E}^{X_{j,t}}(\phi^2_{j,t,b} - \mathbb{E}^{X_{j,t}}\phi^2_{j,t,b})^2 \le \Delta.$$

Under the boundedness assumption, we have $|\phi_{j,t,b}| \le 2$ and hence $|\phi^2_{j,t,b} - \mathbb{E}^{X_{j,t}}\phi^2_{j,t,b}| \le 4$.

By Theorem 4.2 of Chen & Christensen (2015), we have for any integers $\tau \ge 0$ and $1 < d < n_0T/2$ that

$$\mathbb{P}\left(\left|\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T}(\phi^2_{j,t,b} - \mathbb{E}^{X_{0,0}}\phi^2_{0,0,b})\right| \ge 6\tau \,\middle|\, \Delta\right) \le \frac{n_0T}{d}\beta(d) + \mathbb{P}\left(\left|\sum_{(j,t)\in\mathcal{I}_r}(\phi^2_{j,t,b} - \mathbb{E}^{X_{0,0}}\phi^2_{0,0,b})\right|_2 \ge \tau \,\middle|\, \Delta\right) + 4\exp\left(-\frac{\tau^2/2}{n_0Td\Delta + 4d\tau/3}\right),$$

where $\mathcal{I}_r$ denotes the last $n_0T - d\lfloor n_0T/d\rfloor$ elements in the list

$$\{(\ell_1,1),(\ell_1,2),\cdots,(\ell_1,T),(\ell_2,1),(\ell_2,2),\cdots,(\ell_2,T),\cdots,(\ell_{n_0},1),(\ell_{n_0},2),\cdots,(\ell_{n_0},T)\}, \tag{36}$$

and $\lfloor z\rfloor$ denote the largest integer that is smaller than or equal to z for any z. Suppose $\tau \ge 4d$. Notice that $|\mathcal{I}_r| \le d$. It follows that

$$\mathbb{P}\left(\left|\sum_{(j,t)\in\mathcal{I}_r}(\phi^2_{j,t,b} - \mathbb{E}^{X_{0,0}}\phi^2_{0,0,b})\right|_2 \ge \tau \,\middle|\, \Delta\right) = 0.$$

Notice that $\beta(t) = O(\rho^t)$. Set $d = -(c^*+3)\log(n_0T)/\log\rho$, we obtain $n_0T\beta(d)/d = O(n_0^{-2}T^{-2}B^{-1}) = O(B^{-1}Q^{-1}n^{-2}T^{-2})$, since $Q \le T$, $B = O((nT)^{c_*})$ and $n_0 = n/\mathbb{L}$. Here, the big-O notation is uniform in $b \in \{1,\cdots,B\}$ and $q \in \{0,\cdots,Q\}$. Set $\tau = \max\{3\sqrt{\Delta n_0Td\log(Bn_0T)}, 11d\log(Bn_0T)\}$, we obtain that

$$\frac{\tau^2}{4} \ge 2n_0Td\Delta\log(BTn_0) \quad\text{and}\quad \frac{\tau^2}{4} \ge 8d\tau\log(BTn_0)/3 \quad\text{and}\quad \tau \ge 4d,$$

as either $n\to\infty$ or $T\to\infty$. It follows that $\tau^2/(2n_0Td\Delta + 8d\tau/3) \ge 2\log(Bn_0T)$ and hence

$$\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\mathbb{P}\left(\left|\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T}(\phi^2_{j,t,b} - \mathbb{E}^{X_{0,0}}\phi^2_{0,0,b})\right| \ge 6\tau \,\middle|\, \Delta\right) = O(B^{-1}Q^{-1}n^{-1}T^{-1}).$$

By Bonferroni's inequality, we obtain

$$\mathbb{P}\left(\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\left|\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T}(\phi^2_{j,t,b} - \mathbb{E}^{X_{0,0}}\phi^2_{0,0,b})\right| \ge 6\tau \,\middle|\, \Delta\right) = O(n^{-1}T^{-1}).$$

Thus, with probability $1 - O(n^{-1}T^{-1})$, we have

$$\max_{b\in\{1,\cdots,B\}}\max_{q\in\{0,\cdots,Q\}}\left|\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T}(\phi^2_{j,t,b} - \mathbb{E}^{X_{0,0}}\phi^2_{0,0,b})\right| = O(\sqrt{\Delta n_0T}\log(Bn_0T), \log^2(Bn_0T)). \tag{37}$$

Under the given conditions on Q, we have $T - q - 1$ is proportional to T for any $q \le Q$. Combining (C4) and the condition on B with (37) yields (33).

Proof of (30) **with** $m = 2$**:** Similar to the proof of (31), it suffices to show $\max_{q,b}\sqrt{n(T-q-1)}|R_{2,\ell}(q,\mu_b,\nu_b)| = o_p(\log^{-1/2}(nT))$, or $\max_{q,b}\sqrt{n(T-q-1)}|R^{(r)}_{2,\ell}(q,\mu_b,\nu_b)| = o_p(\log^{-1/2}(nT))$ for any $\ell = 1,\cdots,\mathbb{L}$ and $r =$

$1, 2, 3, 4$ where

$$R_{2,\ell}(q,\mu,\nu) = \frac{1}{n(T-q-1)} \sum_{j\in\mathcal{I}^{(\ell)}} \sum_{t=1}^{T-q-1} \{\exp(i\mu^\top S_{j,t+q+1}) - \varphi^*(\mu|X_{j,t+q})\}\{\psi^*(\nu|X_{j,t}) - \widehat{\psi}^{(-\ell)}(\nu|X_{j,t})\},$$

$$R^{(1)}_{2,\ell}(q,\mu,\nu) = \frac{1}{n(T-q-1)} \sum_{j\in\mathcal{I}^{(\ell)}} \sum_{t=1}^{T-q-1} \{\cos(\mu^\top S_{j,t+q+1}) - \varphi^*_R(\mu|X_{j,t+q})\}\{\psi^*_R(\nu|X_{j,t}) - \widehat{\psi}^{(-\ell)}_R(\nu|X_{j,t})\},$$

$$R^{(2)}_{2,\ell}(q,\mu,\nu) = \frac{1}{n(T-q-1)} \sum_{j\in\mathcal{I}^{(\ell)}} \sum_{t=1}^{T-q-1} \{\sin(\mu^\top S_{j,t+q+1}) - \varphi^*_I(\mu|X_{j,t+q})\}\{\psi^*_I(\nu|X_{j,t}) - \widehat{\psi}^{(-\ell)}_I(\nu|X_{j,t})\},$$

$$R^{(3)}_{2,\ell}(q,\mu,\nu) = \frac{1}{n(T-q-1)} \sum_{j\in\mathcal{I}^{(\ell)}} \sum_{t=1}^{T-q-1} \{\cos(\mu^\top S_{j,t+q+1}) - \varphi^*_R(\mu|X_{j,t+q})\}\{\psi^*_I(\nu|X_{j,t}) - \widehat{\psi}^{(-\ell)}_I(\nu|X_{j,t})\},$$

$$R^{(4)}_{2,\ell}(q,\mu,\nu) = \frac{1}{n(T-q-1)} \sum_{j\in\mathcal{I}^{(\ell)}} \sum_{t=1}^{T-q-1} \{\sin(\mu^\top S_{j,t+q+1}) - \varphi^*_I(\mu|X_{j,t+q})\}\{\psi^*_R(\nu|X_{j,t}) - \widehat{\psi}^{(-\ell)}_R(\nu|X_{j,t})\}.$$

In the following, we only show $\max_{q,b}\sqrt{n(T-q-1)}|R^{(1)}_{2,\ell}(q,\mu_b,\nu_b)| = o_p(\log^{-1/2}(nT))$ to save space.

Define the list

$$\{(\ell_1,1),(\ell_1,2),\cdots,(\ell_1,T-q),(\ell_2,1),(\ell_2,2),\cdots,(\ell_2,T-q)\cdots,(\ell_{n_0},1),(\ell_{n_0},2),\cdots,(\ell_{n_0},T-q)\}.$$

For any $1\le g\le n_0(T-q)$, denote by (n_g,T_g) the g-th element in the list. Let $\mathcal{F}^{(0)}_q = \{X_{\ell_1,1}, X_{\ell_1,2},\cdots,X_{\ell_1,1+q}\}\cup\{X_{j,t}: 0\le t\le T, j\in\mathcal{I}^{(-\ell)}\}\cup\{\mu_1,\cdots,\mu_B,\nu_1,\cdots,\nu_B\}$. Then we recursively define $\mathcal{F}^{(g)}_q$ as

$$\mathcal{F}^{(g)}_q = \begin{cases} \mathcal{F}^{(g-1)}_q \cup \{X_{n_g,t_g+q+1}\}, & \text{if } g=1 \text{ or } n_g = n_{g-1};\\ \mathcal{F}^{(g-1)}_q \cup \{X_{n_{g-1},T}, X_{n_g,1}, X_{n_g,2},\cdots,X_{n_g,1+q}\}, & \text{otherwise.}\end{cases}$$

Let $\phi^*_{g,q,b} = \{\cos(\mu_b^\top S_{n_g,t_g+q+1}) - \varphi^*_R(\mu_b|X_{n_g,t_g+q})\}\{\psi^*_R(\nu_b|X_{n_g,t_g}) - \widehat{\psi}^{(-\ell)}_R(\nu_b|X_{n_g,t_g})\}$. Under MA, $R^{(1)}_{2,\ell}(q,\mu_b,\nu_b)$ can be rewritten as $\{n(T-q-1)\}^{-1}\sum_{g=1}^{n_0(T-q)}\phi^*_{g,q,b}$ and forms a sum of martingale difference sequence with respect to the filtration $\{\sigma(\mathcal{F}^{(g)}_q): g\ge 0\}$ where $\sigma(\mathcal{F}^{(g)}_q)$ denotes the σ-algebra generated by variables in $\mathcal{F}^{(g)}_q$. In the following, we apply concentration inequalities for martingales to bound $\max_{q,b}|R^{(1)}_{2,\ell}(q,\mu_b,\nu_b)|$.

Under the boundedness condition, we have $|\phi^*_{g,q,b}|^2 \le 4\{\psi^*_R(\nu_b|X_{n_g,t_g}) - \widehat{\psi}^{(-\ell)}_R(\nu_b|X_{n_g,t_g})\}^2$. In addition, we have by MA that

$$\mathbb{E}\{(\phi^*_{g+1,q,b})^2|\sigma(\mathcal{F}^{(g)}_q)\} = \mathbb{E}[\{\cos(\mu_b^\top S_{n_g,t_g+q+1}) - \varphi^*_R(\mu_b|X_{n_g,t_g+q})\}^2|X_{n_g,t_g+q}]$$
$$\times\{\psi^*_R(\nu_b|X_{n_g,t_g}) - \widehat{\psi}^{(-\ell)}_R(\nu_b|X_{n_g,t_g})\}^2 \le 4\{\psi^*_R(\nu_b|X_{n_g,t_g}) - \widehat{\psi}^{(-\ell)}_R(\nu_b|X_{n_g,t_g})\}^2.$$

It follows from Theorem 2.1 of Bercu & Touati (2008) that

$$\mathbb{P}\left(\left|\sum_{g=1}^{n_0(T-q)}\phi^*_{g,q,b}\right|\ge\tau,\ \sum_{g=1}^{n_0(T-q)}4\{\psi^*_R(\nu_b|X_{n_g,t_g}) - \widehat{\psi}^{(-\ell)}_R(\nu_b|X_{n_g,t_g})\}^2\le y\right)\le 2\exp\left(-\frac{\tau^2}{2y}\right),\quad\forall y,\tau,$$

and hence

$$\mathbb{P}\left(\left|\sum_{g=1}^{n_0(T-q)}\phi^*_{g,q,b}\right|\ge\tau,\ \max_{b\in\{1,\cdots,B\}}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T}\{\psi^*_R(\nu_b|X_{j,t}) - \widehat{\psi}^{(-\ell)}_R(\nu_b|X_{j,t})\}^2\le\frac{y}{4}\right)\le 2\exp\left(-\frac{\tau^2}{2y}\right),\quad\forall y,\tau,$$

By Bonferroni's inequality, we obtain

$$\mathbb{P}\left(\max_{\substack{q\in\{0,\cdots,Q\}\\ b\in\{1,\cdots,B\}}}\left|\sum_{g=1}^{n_0(T-q)}\phi^*_{g,q,b}\right|\ge\tau,\ \max_{b\in\{1,\cdots,B\}}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T}\{\psi^*_R(\nu_b|X_{j,t}) - \widehat{\psi}^{(-\ell)}_R(\nu_b|X_{j,t})\}^2\le\frac{y}{4}\right)\le 2BQ\exp\left(-\frac{\tau^2}{2y}\right),$$

for any y, τ. Set $y = 4\varepsilon\sqrt{nT}$, we obtain

$$\mathbb{P}\left(\max_{\substack{q\in\{0,\cdots,Q\}\\ b\in\{1,\cdots,B\}}}\left|\sum_{g=1}^{n_0(T-q)}\phi^*_{g,q,b}\right| \geq \tau, \max_{b\in\{1,\cdots,B\}}\sum_{j\in\mathcal{I}^{(\ell)}}\sum_{t=1}^{T}\{\psi^*_R(\nu_b|X_{j,t}) - \widehat{\psi}^{(-\ell)}_R(\nu_b|X_{j,t})\}^2 \leq \sqrt{nT}\right) \leq 2BQ\exp\left(-\frac{\tau^2}{2\sqrt{nT}}\right),$$

It follows from (34) that

$$\mathbb{P}\left(\max_{\substack{q\in\{0,\cdots,Q\}\\ b\in\{1,\cdots,B\}}}\left|\sum_{g=1}^{n_0(T-q)}\phi^*_{g,q,b}\right| \geq \tau\right) \leq 2BQ\exp\left(-\frac{\tau^2}{2\sqrt{nT}}\right) + o(1). \tag{38}$$

Set $\tau = (nT)^{1/4}\sqrt{2\log(BQnT)}$, the right-hand-side (RHS) of (38) is $o(1)$. Under the given conditions on B and Q, we obtain $\max_{q,b}\sqrt{n(T-q-1)}|R^{(1)}_{2,\ell}(q,\mu_b,\nu_b)| = o_p(\log^{-1/2}(nT))$.

C.3.2 Step 2

For any $j \in \mathcal{I}^{(\ell)}$ and $0 < t < T-q$, define vectors $\lambda^*_{R,q,j,t}, \lambda^*_{I,q,j,t} \in \mathbb{R}^{\mathbb{B}}$ such that the b-th element of $\lambda^*_{R,q,j,t}, \lambda^*_{I,q,j,t}$ correspond to the real and imaginary part of

$$\frac{1}{\sqrt{n(T-q-1)}}\{\exp(i\mu_b^\top S_{j,t+q+1}) - \varphi^*(\mu_b|X_{j,t+q})\}\{\exp(i\nu_b^\top X_{j,t-1}) - \psi^*(\nu_b|X_{j,t})\},$$

respectively. Let $\lambda^*_{q,j,t}$ denote the (2B)-dimensional vector $(\lambda^{*\top}_{R,q,j,t}, \lambda^{*\top}_{I,q,j,t})^\top$. In addition, we define a (2B(Q+1))-dimensional vector $\lambda^*_{j,t}$ as $(\lambda^{*\top}_{0,j,t}, \lambda^{*\top}_{1,j,t-1}\mathbb{I}(t>1), \cdots, \lambda^{*\top}_{1,j,t-Q}\mathbb{I}(t>Q))^\top$. Define the list

$$(1,1),(1,2),\cdots,(1,T-1),(2,1),(2,2),\cdots,(2,T-1),\cdots,(n,1),(n,2),\cdots,(n,T-1). \tag{39}$$

For any $1 \leq g \leq n(T-1)$, let (n_g, t_g) be the g-th element in the list. Let $\mathcal{F}^{(0)} = \{X_{1,0}\} \cup \{\mu_1,\cdots,\mu_B,\nu_1,\cdots,\nu_B\}$ and recursively define $\mathcal{F}^{(g)}$ as

$$\mathcal{F}^{(g)} = \begin{cases} \mathcal{F}^{(g-1)} \cup \{X_{n_g,t_g}\}, & \text{if } g = 1 \text{ or } n_g = n_{g-1};\\ \mathcal{F}^{(g-1)} \cup \{X_{n_{g-1},T}, X_{n_g,0}\}, & \text{otherwise.}\end{cases}$$

The high-dimensional vector $M_{n,T} = \sum_{g=1}^{n(T-1)}\lambda^*_{n_g,t_g}$ forms a sum of martingale difference sequence with respect to the filtration $\{\sigma(\mathcal{F}^{(g)}) : g \geq 0\}$. Notice that $S^* = \|\sum_{g=1}^{n(T-1)}\lambda^*_{n_g,t_g}\|_\infty$. In this step, we apply the high-dimensional martingale central limit theorem developed by Belloni & Oliveira (2018) to establish the limiting distribution of S^*.

For $1 \leq g \leq n(T-1)$, let

$$\Sigma_g = \sum_{g=1}^{n(T-1)}\mathbb{E}\left(\lambda^*_{n_g,t_g}\lambda^{*\top}_{n_g,t_g}\middle|\mathcal{F}^{(g-1)}\right).$$

Let $V^* = \sum_{g=1}^{n(T-1)}\Sigma_g$. Using similar arguments in proving (37), we can show $\|V^* - V_0\|_{\infty,\infty} = O((nT)^{-1/2}\log(BnT)) + O((nT)^{-1}\log^2(BnT))$, with probability $1 - O(n^{-1}T^{-1})$, where $V_0 = \mathbb{E}V^*$. Under the given conditions on B, we have $\|V^* - V_0\|_{\infty,\infty} \leq \kappa_{B,n,T}$ for some $\kappa_{B,n,T} = O((nT)^{-1/2}\log(nT))$, with probability $1 - O(n^{-1}T^{-1})$.

In addition, under the boundedness assumption in (C4), all the elements in V^* and V_0 are uniformly bounded by some constants. It follows that

$$\mathbb{E}\|V^* - V_0\|_{\infty,\infty} \leq \kappa_{B,n,T} + \mathbb{P}(\|V^* - V_0\|_{\infty,\infty} > \kappa_{B,n,T}) = O((nT)^{-1/2}\log(nT)).$$

By Theorem 3.1 of Belloni & Oliveira (2018), we have for any Borel set $\mathcal{R}$ and any $\delta > 0$ that

$$\begin{aligned}&\mathbb{P}(S^* \in \mathcal{R}) \leq \mathbb{P}(\|N(0,V_0)\|_\infty \in \mathcal{R}^{C\delta})| \\ \leq\ & C\left(\frac{1}{nT} + \frac{\log(BnT)\log(BQ)}{\delta^2\sqrt{nT}} + \frac{\log^3(BQ)}{\delta^3\sqrt{nT}} + \frac{\log^3(BQ)}{\delta^3}\sum_{g=1}^{n(T-1)}\mathbb{E}\|\eta_g\|^3_\infty\right),\end{aligned} \tag{40}$$

for some constant $C > 0$.

Under the boundedness assumption in (C4), the absolute value of each element in Σ_g is uniformly bounded by $16(n(T-q-1))^{-1} = O(n^{-1}T^{-1})$. With some calculations, we can show that $\sum_{g=1}^{n(T-1)} \mathbb{E}\|\eta_g\|_\infty^3 = O((nT)^{-1/2}\log^{3/2}(BQ))$. In addition, we have $Q = O(T)$ and $B = O((nT)^{c_*})$. Combining these together with (40) yields

$$\mathbb{P}(S^* \in \mathcal{R}) \leq \mathbb{P}(\|N(0, V_0)\|_\infty \in \mathcal{R}^{C\delta})| + O(1)\left(\frac{1}{nT} + \frac{\log^2(nT)}{\delta^2\sqrt{nT}} + \frac{\log^{9/2}(nT)}{\delta^3\sqrt{nT}}\right), \tag{41}$$

where $O(1)$ denotes some positive constant.

Set $\mathcal{R} = (z, +\infty)$ and $\delta = \varepsilon \log^{-1/2}(nT)/C$, we obtain

$$\mathbb{P}(S^* \leq z) \geq \mathbb{P}(\|N(0, V_0)\|_\infty \leq z - \varepsilon \log^{-1/2}(nT)) - o(1).$$

Set $\mathcal{R} = (-\infty, z]$, we can similarly show

$$\mathbb{P}(S^* \leq z) \leq \mathbb{P}(\|N(0, V_0)\|_\infty \leq z + \varepsilon \log^{-1/2}(nT)) + o(1).$$

This completes the proof of Step 2.

C.3.3 Step 3

We break the proof into two parts. In Part 1, we show V_0 is a block diagonal matrix. Specifically, let V_{0,q_1,q_2} denote the $(2B) \times (2B)$ submatrix of V_0 formed by rows in $\{2q_1B+1, 2q_1B+2, \cdots, 2(q_1+1)B\}$ and columns in $\{2q_2B+1, 2q_2B+2, \cdots, 2(q_2+1)B\}$. For any $q_1 \neq q_2$, we show $V_{0,q_1,q_2} = O_{(2B)\times(2B)}$.

Let $\Sigma^{(q)}$ denote $V_{0,q,q}$. In Part 2, we provide an upper bound for $\max_{q\in\{0,\cdots,Q\}} \|\Sigma^{(q)} - \widehat{\Sigma}^{(q)}\|_{\infty,\infty}$. Let $\widehat{V}$ be a block diagonal matrix where the main diagonal blocks are given by $\widehat{\Sigma}^{(0)}, \widehat{\Sigma}^{(1)}, \cdots, \widehat{\Sigma}^{(Q)}$, we obtain $\|V_0 - \widehat{V}\|_{\infty,\infty}$

Part 1: Let $\lambda^*_{R,q,j,t,b}$ and $\lambda^*_{I,q,j,t,b}$ denote the b-th element of $\lambda^*_{R,q,j,t}$ and $\lambda^*_{I,q,j,t}$, respectively. Each element in V_{0,q_1,q_2} equals $\mathbb{E}(\sum_{j,t}\lambda^*_{Z_1,q_1,j,t,b_1})(\sum_{j,t}\lambda^*_{Z_2,q_2,j,t,b_2})$ for some $b_1, b_2 \in \{1, \cdots, B\}$ and $Z_1, Z_2 \in \{R, I\}$. In the following, we show

$$\mathbb{E}\left(\sum_{j,t}\lambda^*_{R,q_1,j,t,b_1}\right)\left(\sum_{j,t}\lambda^*_{R,q_2,j,t,b_2}\right) = 0, \quad \forall q_1 \neq q_2.$$

Similarly, one can show $\mathbb{E}(\sum_{j,t}\lambda^*_{R,q_1,j,t,b_1})(\sum_{j,t}\lambda^*_{I,q_2,j,t,b_2}) = 0$ and $\mathbb{E}(\sum_{j,t}\lambda^*_{I,q_1,j,t,b_1})(\sum_{j,t}\lambda^*_{I,q_2,j,t,b_2}) = 0$ for any $q_1 \neq q_2$. This completes the proof for Part 1.

Since observations in different trajectories are i.i.d, it suffices to show

$$\sum_j \mathbb{E}\left(\sum_t \lambda^*_{R,q_1,j,t,b_1}\right)\left(\sum_t \lambda^*_{R,q_2,j,t,b_2}\right) = 0, \quad \forall q_1 \neq q_2,$$

or equivalently,

$$\mathbb{E}\left(\sum_t \lambda^*_{R,q_1,0,t,b_1}\right)\left(\sum_t \lambda^*_{R,q_2,0,t,b_2}\right) = 0, \quad \forall q_1 \neq q_2, \tag{42}$$

By definition, we have

$$\lambda^*_{R,q,0,t,b} = \frac{1}{\sqrt{n(T-q-1)}}\{\cos(\mu_b^\top S_{0,t+q+1}) - \varphi^*_R(\mu_b|X_{0,t+q})\}\{\cos(\nu_b^\top X_{0,t-1}) - \psi^*_R(\nu_b|X_{0,t})\}.$$

Since $q_1 \neq q_2$, for any t_1, t_2, we have either $t_1 + q_1 \neq t_2 + q_2$ or $t_1 \neq t_2$. Suppose $t_1 + q_1 > t_2 + q_2$. Under MA, we have

$$\mathbb{E}[\{\cos(\mu_b^\top S_{0,t_1+q_1+1}) - \varphi^*_R(\mu_b|X_{0,t_1+q_1})\}|\{X_{0,j}\}_{j\leq t_1+q_1}] = 0, \quad \forall b,$$

and hence

$$\mathbb{E}\lambda^*_{R,q_1,0,t_1,b_1}\lambda^*_{R,q_2,0,t_2,b_2} = 0, \quad \forall b_1, b_2. \tag{43}$$

Similarly, when $t_1 + q_1 < t_2 + q_2$, we can show (43) holds as well.

Suppose $t_1 < t_2$, under (C1) and H_0, we have

$$\mathbb{E}[\{\cos(\nu_b^\top X_{0,t_1-1}) - \varphi_R^*(\nu_b|X_{0,t_1})\}|\{X_{0,j}\}_{j\geq t_1}] = 0, \qquad \forall b,$$

and hence (43) holds. Similarly, when $t_1 > t_2$, we can show (43) holds as well. This yields (42).

Part 2: For any $q \in \{0, \cdots, Q\}$, we can represent $\widehat{\Sigma}^{(q)} - \Sigma^{(q)}$ by

$$\sum_{\ell=1}^{\mathbb{L}} \sum_{j\in\mathcal{I}^{(\ell)}} \sum_{t=1}^{T-q-1} \frac{(\lambda_{R,q,j,t}^\top, \lambda_{I,q,j,t}^\top)^\top (\lambda_{R,q,j,t}^\top, \lambda_{I,q,j,t}^\top) - (\lambda_{R,q,j,t}^{*\top}, \lambda_{I,q,j,t}^{*\top})^\top (\lambda_{R,q,j,t}^{*\top}, \lambda_{I,q,j,t}^{*\top})}{n(T-q-1)}. \tag{44}$$

Using similar arguments in Step 1 of the proof, we can show with probability tending to 1 that the absolute value of each element in (44) is upper bounded by $c_0^*(nT)^{-c^{**}}$ for any $q \in \{0, \cdots, Q\}$ and some positive constants $c_0, c^* > 0$. Thus we obtain $\max_{q\in\{0,\cdots,Q\}} \|\widehat{\Sigma}^{(q)} - \Sigma^{(q)}\|_{\infty,\infty} = O((nT)^{-c^{**}})$, with probability tending to 1. The proof is hence completed.