Does the Markov Decision Process Fit the Data: Testing for the Markov Property in Sequential Decision Making

Chengchun Shi [1] Runzhe Wan [2] Rui Song [2] Wenbin Lu [2] Ling Leng [3]

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 (MDPs) and partially observable MDPs. Theoretically, we establish the validity of our test. Empirically, 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_t, but also the past state-action pairs (S_{t-1}, A_{t-1}), $\cdots$, $(S_{t-\kappa_0+1}, A_{t-\kappa_0+1})$ for some $\kappa_0 > 1$. 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 (POMDPs). Suppose we concatenate measurements over sufficiently many decision points and our test still rejects MA. Then we shall consider modelling the system dynamics by POMDPs or other non-Markovian problems. Applying RL algorithms designed for these settings has been shown to outperform 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 moderate or high-dimensional. This is certainly the case as we convert the process into an MDP by concatenating data over multiple decision points. Modern machine learning (ML) algorithms are well-suited for prediction tasks in high dimensions. Yet, the large bias of the resulting estimates makes statistical inference (e.g., hypothesis testing) extremely difficult. The key ingredient of our test lies in constructing a doubly robust estimating equation to alleviate the biases. This ensures our test statistic has a tractable limiting distribution even in high dimensions. Consequently,

[1] Department of Statistics, London School of Economics and Political Science, UK [2] Department of Statistics, North Carolina State University, USA [3] Amazon, USA. Correspondence to: Chengchun Shi <c.shi7@lse.ac.uk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

the proposed test well controls the type-I error rate (see Theorem 3).

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.

In the economics literature, Chen & Hong (2012) developed a test for testing the Markov property of a multivariate time series. 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 moderate or high-dimensions, since local polynomial fitting suffers from the curse of dimensionality. Naively replacing local polynomial fitting with ML estimates will invalidate their test due to the large bias of the estimator.

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; Shah & Peters, 2018; Berrett et al., 2020; Shi 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. Finally, we conclude our paper in Section 6.

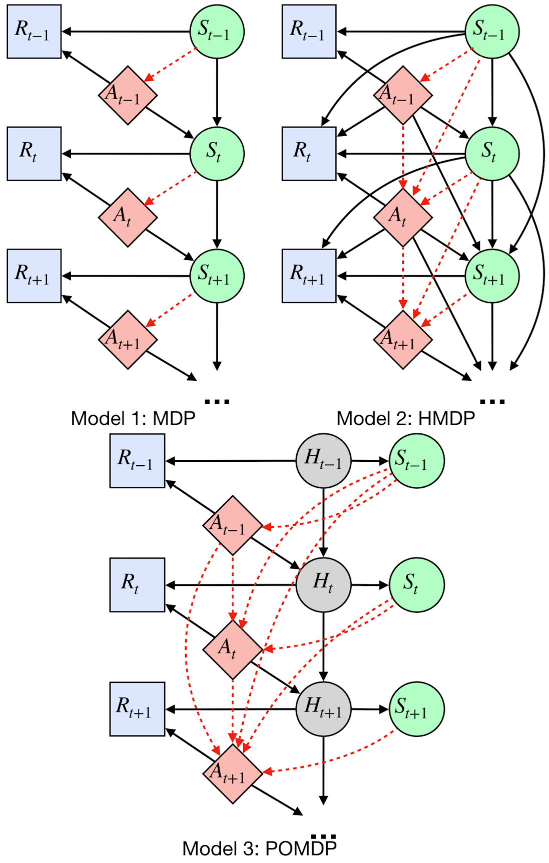

Figure 1. Causal diagrams for MDPs, HMDPs (second order) and POMDPs. The solid lines represent the causal relationships and the dashed lines indicate the information needed to implement the optimal policy.

2. Model Setup

2.1. MDP and Existence of the Optimal Stationary Policy

Consider a single trajectory $\{(S_t, A_t, R_t)\}_{t\geq 0}$ where (S_t, A_t, R_t) denotes the state-action-reward triplet collected at time t. For any integer $t \geq 0$, let $\bar{\boldsymbol{S}}_t = (S_0, A_0, S_1, A_1, \cdots, S_t)^\top$ denote the state and action history. Similarly define $\bar{\boldsymbol{R}}_t = (R_0, R_1, \cdots, R_t)^\top$. 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,

$$\mathbb{P}(S_{t+1} \in \mathcal{S}, R_t \in \mathcal{R} | A_t, \bar{\boldsymbol{S}}_t, \bar{\boldsymbol{R}}_t) = \mathcal{P}(\mathcal{S}, \mathcal{R}; A_t, S_t),$$

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 $\bar{\boldsymbol{S}}_t$ to a probability mass function $\pi_t(\cdot|\bar{\boldsymbol{S}}_t)$ on $\mathcal{A}$. When there exists some function π^* such that $\pi_t(\cdot|\bar{\boldsymbol{S}}_t) = \pi^*(\cdot|S_t)$ 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

$$V(\pi; s) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^{\pi_t}(R_t | S_0 = s),$$

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 Section C.1 of the supplementary article 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_{t+1} and R_t depend not only on (S_t, A_t), but $(S_{t-1}, A_{t-1}), \cdots, (S_{t-\kappa_0+1}, A_{t-\kappa_0+1})$ for some integer $\kappa_0 > 1$ as well. Formally, we have

$$\begin{aligned}&\mathbb{P}(S_{t+1} \in \mathcal{S}, R_t \in \mathcal{R} | A_t, \bar{\boldsymbol{S}}_t, \bar{\boldsymbol{R}}_t)\\ =&\mathcal{P}(\mathcal{S}, \mathcal{R}; \{A_j\}_{t-\kappa_0<j\leq t}, \{S_j\}_{t-\kappa_0<j\leq t}),\end{aligned} \tag{1}$$

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

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

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

Similar to Lemma 1, there exists an optimal stationary policy that depends on $\bar{\boldsymbol{S}}_t$ only through $S_t(\kappa_0)$. This 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 $\{(S_t(k), A_t(k), R_t(k))\}_{t\geq 0}$ satisfy MA for $k = 1, 2, \cdots$, until the null of 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_{t-1}. The hidden variables $\{H_t\}_{t\geq 0}$ are unobserved. Suppose the agent chooses an action A_{t-1}. Similar to MDPs, this will cause the environment to transition to a new state H_t at time t. At the same time, the agent receives an observation $S_t \in \mathbb{S}$ and a reward R_t that depend on H_t and A_{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 $S_{t-1} \leftarrow H_{t-1} \rightarrow H_t \rightarrow H_{t+1} \rightarrow S_{t+1}$ connects S_{t-1} and S_{t+1} without traversing S_t and A_t. As a result, S_{t+1} and S_{t-1} are not d-separated (see the definition of d-separation on Page 16, Pearl, 2000) given S_t and A_t. Under the faithfulness assumption (see e.g. Kalisch & Bühlmann, 2007), S_{t-1} and S_{t+1} are mutually dependent conditional on S_t and A_t. Similarly, we can show S_{t+k} and S_{t-1} are mutually dependent conditional on $\{(S_j, A_j)\}_{t\leq j<t+k}$ 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

We introduce our testing procedure in this section. To motivate our test, we begin by presenting an equivalent characterization of MA based on the notion of CCF. For simplicity, suppose R_t is a deterministic function of S_{t+1}, A_t and S_t. This condition automatically holds if we include R_t in the set of state variables S_{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,

$$\mathbb{P}(S_{t+1} \in \mathcal{S} | A_t, \bar{\boldsymbol{S}}_t) = \mathcal{P}(\mathcal{S}; A_t, S_t), \tag{2}$$

for any $\mathcal{S} \subseteq \mathbb{S}$ and $t \geq 0$. The observed data consists of n trajectories. Specifically, let $\{(S_{i,t}, A_{i,t}, R_{i,t})\}_{0\leq t\leq T}$ be the data from the i-th trajectory where T is the termination time. We assume $\{(S_{1,t}, A_{1,t}, R_{1,t})\}_{0\leq t\leq T}$, $\cdots$, $\{(S_{n,t}, A_{n,t}, R_{n,t})\}_{0\leq t\leq T}$ are i.i.d. copies of

$\{(S_t, A_t, R_t)\}_{0\le t\le T}$. Given the observed data, we focus on testing the following pair of hypotheses:

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

The Markov property is closely related to the notion of conditional independence. Informally speaking, it implies that the past and future states are independent conditional on the present. To be more specific, 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 assumptions:

$$S_{t+q+1} \perp\!\!\!\perp (S_{t-1}, A_{t-1}) | \{(S_j, A_j)\}_{t\le j\le t+q}, \tag{3}$$

for any $t > 0$ and $q \ge 0$.

Next, we present an equivalent presentation for conditional independence based on CCF.

Lemma 2 *For any Z_1, Z_2 and Z_3, $Z_1 \perp\!\!\!\perp Z_2|Z_3$ if and only if $\mathbb{E}\{\exp(i\mu_1^\top Z_1)|Z_3\}\mathbb{E}\{\exp(i\mu_2^\top Z_2)|Z_3\} = \mathbb{E}\{\exp(i\mu_1^\top Z_1 + i\mu_2^\top Z_2)|Z_3\}$ for any μ_1, μ_2 almost surely.*

For any t, let $X_t = (S_t^\top, A_t)^\top$ denote the state-action pair. For any $\mu \in \mathbb{R}^p$, define the CCF of S_{t+1} given X_t by $\varphi_t(\mu|x) = \mathbb{E}\{\exp(i\mu^\top S_{t+1})|X_t = x\}$. Based on Lemma 2, we present an equivalent representation for (3) below.

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

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

Under H_0, there exists some φ^* such that $\varphi_t = \varphi^*$ for any t. Take another expectation on both sides of the above equation, we obtain

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

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

$$\begin{aligned}&\frac{1}{n(T-q-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-q-1}\{\exp(i\mu^\top S_{j,t+q+1})\\ &-\widehat{\varphi}(\mu|X_{j,t+q})\}\times\{\exp(i\nu^\top X_{j,t-1}) - \bar{\varphi}(\nu)\},\end{aligned} \tag{4}$$

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

Modern machine learning (ML) algorithms are well-suited to estimating φ^* in moderate or high-dimensional cases. However, naively plugging ML estimators for $\widehat{\varphi}$ will cause a heavy bias in (4). 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

We begin by introducing the following conditions.

(C1) Actions are generated by a fixed behavior policy.
(C2) Suppose the process $\{S_t\}_{t\ge0}$ is strictly stationary.

Condition (C1) requires the agent to select actions based on information contained in the current state variable only. It is commonly assumed in the literature on off-policy policy evaluation (see e.g., Jiang & Li, 2016). Under H_0, the process $\{X_t\}_{t\ge0}$ forms a time-invariant Markov chain. When its initial distribution equals its stationary distribution, (C2) is automatically satisfied. This together with (C1) implies $\{X_t\}_{t\ge0}$ is strictly stationary as well.

Define another CCF of X_{t-1} given X_t by

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

Under stationarity, we have $\psi_t = \psi^*$ for some ψ^* and any $t > 0$. The following theorem forms the basis of our procedure.

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

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

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 $\psi = \psi^*$,

$$\begin{aligned}&\mathbb{E}\{\exp(i\mu^\top S_{t+q+1}) - \varphi(\mu|X_{t+q})\}\\ &\times\{\exp(i\nu^\top X_{t-1}) - \psi(\nu|X_t)\} = 0.\end{aligned} \tag{5}$$

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

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

under MA. Assertion (5) thus follows. Under (C1), we have $X_{t-1} \perp\!\!\!\perp \{X_j\}_{j>t} | X_t$ *for any* $t > 1$. *When* $\psi = \psi^*$, *we can similarly show that*

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

The doubly-robustness property thus follows.

The proposed 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_t given X_{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.

The following estimating equation corresponds to an estimate of $\Gamma_0(q,\mu,\nu)$,

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

Unlike (4), the above estimating equation is doubly robust. As such, (6) is consistent when either the bias of $\widehat{\varphi}$ or $\widehat{\psi}$ is negligible. In addition, its bias decays to zero at a faster rate than that of $\widehat{\varphi}$ and $\widehat{\psi}$ (see Appendix C.3.1 for details). In contrast, the bias of (4) will be of the same order as that of $\widehat{\varphi}$. As such, the test based on (6) requires a much weaker and practically more feasible condition on the ML estimates (see Condition (C4) for details).

Our test statistic is constructed based on a slightly modified version of (6) 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: $B$, $Q$, $\mathbb{L}$, $\alpha$ 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 (7).
Step 4: For $q = 0,\cdots,Q$, compute an estimated covariance matrix $\widehat{\Sigma}^{(q)}$ according to (A.1) (see Appendix A.1 of the supplementary article for details).
Step 5: Use Monte Carlo to simulate the upper $\alpha/2$-th critical value of $\max_{q\in\{0,\ldots,Q\}} \|\{\widehat{\boldsymbol{\Sigma}}^{(q)}\}^{1/2}\mathbb{Z}_q\|_\infty$ 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 $\mathrm{H}_0$ 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

$$\widehat{S} = \max_{b\in\{1,\cdots,B\}} \max_{q\in\{0,\cdots,Q\}} \sqrt{n(T-q-1)}\times \\ \max(|\widehat{\Gamma}_R(q,\mu_b,\nu_b)|, |\widehat{\Gamma}_I(q,\mu_b,\nu_b)|). \tag{7}$$

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 $\{X_t\}_{t\geq 0}$ 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. 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 of the supplementary article.

4. Model Selection

Algorithm 2 RL Model Selection

```
Input: $B$, $Q$, $\mathbb{L}$, $\alpha$, $K$ and the observed data.
for $k=1,2,\cdots,K$ do
    Apply algorithm 1 with $B$, $Q$, $\mathbb{L}$, $\alpha$ 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 most likely 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

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

and $A_{j,t}(k)=A_{j,t+k}$. Given a large integer K, our procedure sequentially tests the null hypothesis MA based on the concatenated data $\{(S_{j,t}(k),A_{j,t}(k))\}_{1\leq j\leq n,0\leq t\leq T-k}$ for $k=1,\cdots,K$. 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 most likely 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 $\{\widehat{\varphi}^{(-\ell)}(\mu_b|\cdot)\}_{1\leq b\leq B}$. 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.

1. Apply the random forests algorithm with the response-predictor pairs $\{(S_{j,t},X_{j,t-1})\}_{j\in\mathcal{I}^{(-\ell)},1\leq t\leq T}$ 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 falls into.

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 $\sum_{j \in \mathcal{I}^{(-\ell)}, 1 \le t \le T} w_{j,t}^{(-\ell)}(x) \exp(i\mu_b^\top S_{j,t})$.

When the above random forests algorithm is applied to compute the forward and backward learners, the computational complexity of Algorithm 1 is dominated by
$$O(N^2 Q(B + M) + p^* N \log^2(N) LMQ^2 + N_{MC} B^2 Q).$$

Here $N = nT$ is the total number of observations, M is the number of trees, p^* is the dimension of the state-action pair and N_{MC} is the number of Monte Carlo samples generated in Step 5 of Algorithm 1. Such a result is derived based on the existing literature on random forests.

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. It takes roughly 15 hours to complete all numerical experiments [1].

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 our test in mHealth applications. The data contains continuous measurements for six patients 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_t, including the average blood glucose levels G_t during the one hour interval $(t-1, t]$, the carbohydrate estimate for the meal C_t during $(t-1, t]$ and Ex_t which measures exercise intensity during $(t-1, t]$. At time t, we define A_t by discretizing the amount of insulin In_t injected and define R_t according to the Index of Glycemic Control (Rodbard, 2009) that is a deterministic function G_{t+1}. To save space, we present detailed definitions of A_t and R_t in Appendix B.1 of the supplementary article.

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 G_t:
$$\mathrm{G}_t = \alpha + \sum_{i=1}^{4} (\boldsymbol{\beta}_i^T S_{t-i} + c_i A_{t-i}) + E_t,$$
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_t is set to follow $N(0, 9)$.

At each time point, a patient randomly chooses 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 value G_0 is 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 $k \ge 4$ and H_1 holds otherwise.

Empirical rejection rates of our test with different combinations of k, n and the significance level 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.

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

[1] Code available at https://github.com/RunzheStat/TestMDP

Table 1. Policy evaluation results for the OhioT1DM dataset.

k	1	2	3	4	5	6	7	8	9	10
Estimated value V_k	-90.82	-57.53	-63.77	**-52.57**	-56.23	-60.05	-63.70	-54.85	-65.08	-59.59

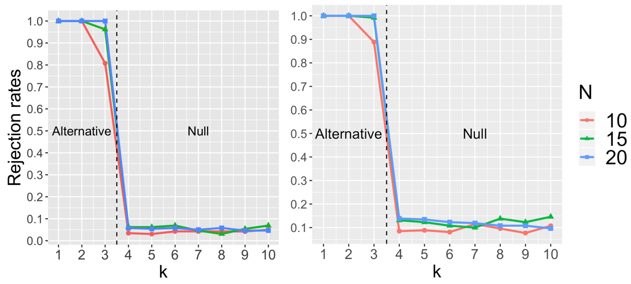

Figure 2. Empirical rejection rates aggregated over 500 simulations with different combinations of α, n and k. $\alpha = (0.05, 0.1)$ from left plot to right plot. When $k < 4$, the alternative hypothesis holds, the empirical rejection rates correspond to the power of the test. When $k \geq 4$, the null holds, the empirical rejection rates corresponds to the type-I error rate.

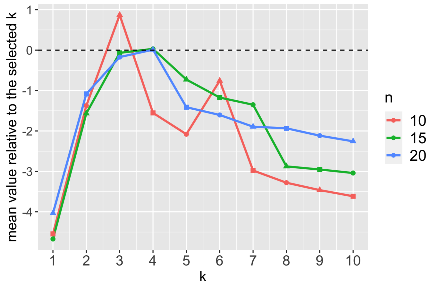

Figure 3. Value differences with different combinations of k and n.

then simulate 100 trajectories following $\widehat{\pi}^{(l)}(k)$ to compute the average discounted reward $V^{(l)}(k)$ (see Section B.2 of the supplementary article 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 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. 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 Section B.3 of the supplementary material 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. In addition, the gain of value under the 4-th order MDP model is significant compared to most other models (see Appendix B.4 for more details).

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_t that corresponds to the estimated location of the tiger. Let H_t denote the observed correct location of the tiger, we have $\mathbb{P}(H_t = S_t) = 0.7$ and $\mathbb{P}(H_t \neq S_t) = 0.3$. 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_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 $\{(S_{j,t}(k), A_{j,t}(k))\}_{1 \leq j \leq N, 0 \leq t \leq T-k+1}$ 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 non-negligible 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 $S_t^* = (S_t, H_t)^\top$ 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.

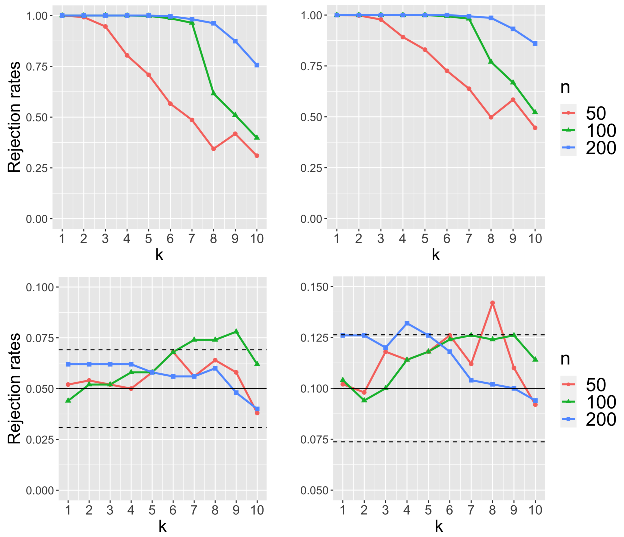

Figure 4. Empirical rejection rates aggregated over 500 simulations with different combinations of α, K and n. $\alpha = (0.05, 0.1)$ from left plots to right plots. H_1 holds in top plots. H_0 holds in bottom plots. Dashed lines correspond to $y = \alpha \pm 1.96$MCE where MCE denotes the Monte Carlo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

6. Discussion

In this paper, we propose a forward and backward learning procedure for testing the goodness of fit of a MDP model. Our test can be naturally coupled with existing state-of-the-art RL algorithms to identify the optimal policy in sequential decision making. RL algorithms have made tremendous achievements in video games, and have found extensive applications in real-world problems including robotics (Kormushev et al., 2013), bidding (Jin et al., 2018), ridesharing (Xu et al., 2018), mobile health (Luckett et al., 2019), etc. We show in our numerical studies that applying our method can help improve the performance of existing RL algorithms in mobile health applications. The proposed method has extensive potential values in other real-world applications as well.

Acknowledgements

The authors wish to thank the meta reviewer and three anonymous referees for their constructive comments, which led to significant improvement of this work.

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.

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.

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

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.

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.

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

Jin, J., Song, C., Li, H., Gai, K., Wang, J., and Zhang, W. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 2193–2201. ACM, 2018.

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.

Kormushev, P., Calinon, S., and Caldwell, D. Reinforcement learning in robotics: Applications and real-world challenges. *Robotics*, 2(3):122–148, 2013.

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

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

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*, 2018.

Shi, C., Xu, T., Bergsma, W., and Li, L. Double generative adversarial networks for conditional independence testing. *arXiv preprint arXiv:2006.02615*, 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.

Xu, Z., Li, Z., Guan, Q., Zhang, D., Li, Q., Nan, J., Liu, C., Bian, W., and Ye, J. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 905–913. ACM, 2018.

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.