Reinforcement Learning in Possibly Nonstationary Environments

Mengbing Li*[1], Chengchun Shi*[2], Zhenke Wu[3], and Piotr Fryzlewicz[4]

[1,3]University of Michigan, Ann Arbor
[2,4]London School of Economics and Political Science

Abstract

We consider reinforcement learning (RL) methods in offline nonstationary environments. Many existing RL algorithms in the literature rely on the stationarity assumption that requires the system transition and the reward function to be constant over time. However, the stationarity assumption is restrictive in practice and is likely to be violated in a number of applications, including traffic signal control, robotics and mobile health. In this paper, we develop a consistent procedure to test the nonstationarity of the optimal policy based on pre-collected historical data, without additional online data collection. Based on the proposed test, we further develop a sequential change point detection method that can be naturally coupled with existing state-of-the-art RL methods for policy optimisation in nonstationary environments. The usefulness of our method is illustrated by theoretical results, simulation studies, and a real data example from the 2018 Intern Health Study[1]. A Python implementation of the proposed procedure is available at `https://github.com/limengbinggz/CUSUM-RL`.

Key Words: Reinforcement learning; Nonstationarity; Hypothesis testing; Change point detection; Policy optimisation.

1 Introduction

Reinforcement learning (RL, see Sutton and Barto, 2018; Levine et al., 2020, for an overview) is a powerful machine learning technique that allows an agent to learn and interact with a

*The first two authors contributed equally to this paper
[1]https://www.srijan-sen-lab.com/intern-health-study

given environment, to maximise the cumulative reward the agent receives. It has been arguably one of the most vibrant research frontiers in machine learning over the last few years. Over 100 papers on RL were accepted for presentation at ICML 2021, a premier conference in the machine learning area, accounting for more than 10% of the accepted papers in total. Significant progress has been made in solving challenging problems across various domains using RL, including games (Silver et al., 2016), robotics (Kormushev et al., 2013), healthcare (Komorowski et al., 2018), bidding (Jin et al., 2018), ridesharing (Xu et al., 2018) and automated driving (de Haan et al., 2019), among many others.

Despite the popularity of developing various RL algorithms in the computer science literature, statistics as a field has only recently begun to engage with RL both in depth and in breadth. Most works in the statistics literature focused on developing data-driven methodologies for precision medicine (see e.g., Murphy, 2003; Robins, 2004; Chakraborty et al., 2010; Qian and Murphy, 2011; Zhang et al., 2013; Zhao et al., 2015; Wallace and Moodie, 2015; Song et al., 2015; Luedtke and van der Laan, 2016; Zhu et al., 2017; Zhang et al., 2018; Shi et al., 2018; Wang et al., 2018; Qi et al., 2020; Nie et al., 2021; Fang et al., 2022). See also Tsiatis et al. (2019); Kosorok and Laber (2019) for overviews. These aforementioned methods were primarily motivated by applications in finite horizon settings with only a few treatment stages. They require a large number of patients in the observed data to achieve consistent estimation and become ineffective in the long or infinite horizon setting where the number of decision stages diverges with the number of observations. The latter setting is widely studied in the computer science literature to formulate many sequential decision making problems in games, robotics, ridesharing, etc. Recently, a few algorithms have been proposed in the statistics literature for policy optimisation or evaluation in infinite horizon settings (Ertefaie and Strawderman, 2018; Luckett et al., 2020; Liao et al., 2020b; Hu et al., 2021a; Liao et al., 2021; Shi et al., 2021; Ramprasad et al., 2021).

Central to the empirical validity of most existing state-of-the-art RL algorithms is the stationarity assumption that requires the state transition and reward functions to be constant over time. Although this assumption is valid in online video games, it is likely violated in a number

of other applications, including traffic signal control (Padakandla et al., 2020), robotic navigation (Niroui et al., 2019), mobile health (Liao et al., 2020a) and infectious disease control Cazelles et al. (2018). It was also mentioned in the seminal book by Sutton and Barto (2018) that "*nonstationarity is the case most commonly encountered in reinforcement learning*". We consider a few examples to elaborate.

One motivating example considered in our paper is from the Intern Health Study (IHS; NeCamp et al., 2020). Medical internship, the first phase of professional medical training in the United States, is a stressful period in the career of physicians. The residents are faced with difficult decisions, long work hours and sleep deprivation. One goal of this ongoing prospective longitudinal study is to investigate when to provide mHealth interventions by sending mobile prompts via a customised study app to provide timely tips for interns to practice anti-sedentary routines that may promote physical well-being. The study data were collected from a 6-month micro-randomised trial on subjects from different specialties. For each subject, daily step counts and sleep minutes were measured via wearable devices (Fitbit) and mood scores measured via ecological momentary assessments (EMAs). The objectively measured physiological variables and the EMA answers have been shown to moderate treatment effects and thus can be used as inputs of treatment policies to improve the interns' physical and psychological well-being (NeCamp et al., 2020). In this paper, we focus on policy optimisation for improving time-discounted cumulative step counts in the presence of potential temporal nonstationarity. Nonstationarity is a serious issue in the mHealth study. For example, in the context of mobile-delivered prompts, the longer a person is under intervention, the more they may habituate to the prompts or become overburdened, resulting in subjects being less responsive to the contents of the suggestions. The treatment effect of activity suggestions may transition from positive to negative, suggesting treatment policies may benefit from adaptation over time. Failure to recognise potential nonstationarity in treatment effects over time may lead to policies that overburden medical interns, resulting in app deletion and study dropouts.

As another example, the coronavirus disease 2019 (COVID-19) has been one of the worst global pandemics in history affecting millions of people. There is a growing interest in ap-

plying RL to develop data-driven intervention policies to contain the spread of the virus (see e.g., Eftekhari et al., 2020; Kompella et al., 2020; Wan et al., 2020). However, the spread of COVID-19 is an extremely complex process and is nonstationary over time. As a result, the optimal policy is likely to vary across time. For instance, at the beginning of the pandemic, stringent lockdown measures have been shown to be highly effective to control the spread of the virus (Anderson et al., 2020). However, these measures would bring enormous costs to the economy (Eichenbaum et al., 2020). When effective vaccines are developed and a large proportion of people are fully vaccinated, it is natural to gradually ease these lockdown restrictions. However, the efficacy of the vaccine is likely to decline over time (Mahase, 2021) and becomes unclear when new variants of the virus arise. To summarise, it is crucial for policy makers to take nonstationarity into consideration to improve global health benefits while balancing the negative impacts of economic and social consequences.

In this paper, we propose a consistent procedure to test the nonstationarity of the optimal policy in infinite horizon settings, based on a pre-collected historical dataset. The proposed test can be naturally coupled with existing state-of-the-art RL algorithms for policy optimisation (e.g., control). Our contributions are summarised as follows.

Scientifically, policy optimisation in nonstationary environments is a vital problem. However, it has been less studied in the existing literature. When the stationary assumption is violated, applying RL algorithms to the entire dataset would yield a sub-optimal policy. A natural idea is to apply RL algorithms to more recent observations to learn the optimal policy. However, in real-world applications, it remains challenging to properly select "the best data segment" without domain knowledge. On the one hand, including too many past observations in the data segment would yield a nonstationary data subset. On the other hand, if the data segment contains only a few most recent observations, it would result in a very noisy policy. To determine the best segment of stationarity, we propose to test if the optimal policy is stationary on a given data segment, and backward sequentially apply our test to a set of candidate data segments for change point detection. Then we apply existing RL algorithms to the data segment after the change. We apply such a procedure to both synthetic and real datasets in Section

6. Results show that the estimated policy based on our constructed data segment is no worse and often better than other baseline methods. In the motivating IHS study, the proposed method reveals the benefit of nonstationarity detection for optimising population physical activities in some medical specialties, leading to on average 170 to 200 additional steps per day for each subject. Promoting the maintenance of healthy behaviors or reducing negative chronic health outcomes often requires longer-term state monitoring and decision-making. As RL continues to drive continuous learning of optimal interventions in mHealth studies, this paper substantiates the need of accommodating nonstationarity with a simple statistical solution.

Methodologically, we propose a novel testing procedure to test the stationarity of the optimal policy and an original change point detection method. To the best of our knowledge, this is the first work on developing statistically sound tests for stationarity in offline RL domains. In the time series literature, a number of works have been developed to test the stationarity of a given time series and detect the change point locations, in models ranging from the simple piecewise-constant signal plus noise setting (Killick et al., 2012; Fryzlewicz, 2014) to high-dimensional panel data and time series (Cho and Fryzlewicz, 2015; Wang and Samworth, 2018); see also Aminikhanghahi and Cook (2017) and Truong et al. (2020) for reviews.

Different from the aforementioned works in time series, the optimal policy is a function of some time-varying state vector. To test its stationarity, we need to check whether the optimal action is stationary over time, for each possible value of the state. In addition, the estimated optimal policy is a highly nonlinear functional of the observed data, making it difficult to derive its limiting distribution. To address these challenges, we notice that the optimal policy is uniquely determined by the optimal state-action value function (also known as the optimal Q-function, see Section 2.3). This motivates us to focus on testing the stationarity of the optimal Q-function. We use the sieve basis to model the optimal Q-function, construct CUSUM-type test statistics and employ multiplier bootstrap to obtain the critical value. It is worth mentioning that our test is able to detect both abrupt and smooth changes, as demonstrated in Section 6. Finally, we apply the proposed test to a set of potential change point locations to compute the p-values, fit an isotonic regression model (see e.g., Brunk et al., 1972; Mukerjee, 1988) to

these p-values and identify the change point location based on the fitted model. The use of isotonic regression allows us to borrow information from all p-values, yielding a more accurate estimator. We remark that, our proposal is an example of harnessing the power of classical statistical inferential tools such as hypothesis testing and isotonic regression to help address an important practical issue in RL.

Theoretically, we systematically establish the size and power properties of the proposed test under a bidirectional asymptotic framework that allows either the number of data trajectories or the number of decision points per trajectory to diverge to infinity. This is useful for different types of applications. There are plenty of mobile health studies that involve a number of subjects and the objective is to develop an optimal policy at the population-level to maximise the overall reward, as in our real data application. Meanwhile, there are other applications where the number of subjects is limited (see e.g., Marling and Bunescu, 2020).

We briefly summarise our theoretical findings here. If the system transitions are stationary over time, the proposed test controls the type-I error even when the sieve approximation error converges slower than the parametric rate. However, the faster the Q-estimator converges to the optimal Q-function, the more powerful the proposed test. Establishing these theoretical results raises a number of challenges. In particular, when the number of sieve basis functions is fixed, the limiting distribution of the test statistic can be established based on classical weak convergence theorems (van der Vaart and Wellner, 1996). However, in our setting, we require the number of sieve basis functions to diverge with the number of observations to allow the approximation error to decay to zero and those theorems are no longer applicable. One of our major technical contributions lies in developing a matrix concentration inequality for nonstationarity Markov decision process (see Lemma B.2 in the supplementary article). The derivation is non-trivial and naively applying the concentration inequality designed for scalar random variables (Alquier et al., 2019) would yield a loose error bound. See Appendix B.4 for details. Another technical contribution is to derive the limiting distribution of the estimated optimal Q-function computed via the fitted Q-iteration (FQI, Ernst et al., 2005) algorithm, one of the most popular Q-learning type algorithms (see Theorem 3 in Section 5.3). We remark

that in the existing literature, most papers focus on establishing non-asymptotic error bound of the estimated Q-function (see e.g., Munos and Szepesvári, 2008; Chen and Jiang, 2019; Fan et al., 2020; Uehara et al., 2021).

The rest of the paper is organised as follows. In Section 2, we introduce the offline RL problem and review some existing algorithms. In Section 3, we illustrate our main idea of learning the optimal policy in nonstationary environment. In Section 4, we detail our test procedure for change point detection. We establish the theoretical properties of our procedure in Section 5, conduct simulation studies in Section 6, and apply the proposed procedure to our data application in Section 7. Finally, we conclude the paper in Section 8.

2 Preliminaries

We first discuss the data structure and formulate our problem of interest. We next discuss the stationarity assumption, which forms the foundation of most existing state-of-the-art RL algorithms. Finally, we review Q-learning (Watkins and Dayan, 1992), one of the most popular RL algorithms, as it is related to our proposal.

2.1 Data and Problem Formulation

We consider an offline setting where the objective is learn an optimal policy based on a pre-collected dataset, without additional online data collection. The offline dataset can be collected from a randomised trial or observational study, and is summarised as

$$\{(S_{i,t}, A_{i,t}, R_{i,t})\}_{1\le i\le N, 0\le t\le T},$$

which consists of N i.i.d. copies of a population trajectory $\{(S_t, A_t, R_t)\}_{t\ge 0}$ censored at some time $T \ge 2$, where i indexes the ith subject, t indexes the tth time point and (S_t, A_t, R_t) denotes the state-action-reward triplet at time t. In the intern health project, the state (e.g., time-varying covariates) corresponds to the time-varying coefficients associated with each medical, such as their mood score, step counts and sleep times. The action (e.g., treatment intervention) is a binary variable, corresponding to whether to send some text message to the doctor or not.

The immediate reward (e.g., clinical outcome) is the step counts. We assume the rewards are uniformly bounded. This assumption is commonly imposed in the literature to simplify the theoretical analysis (see e.g., Fan et al., 2020).

A policy defines the way that a decision maker chooses an action at each decision time. A history dependent policy π is a sequence of decision rules $\{\pi_t\}_{t\geq 0}$ such that each π_t takes the observed data history $\bar{S}_t$ including S_t and the state-action-reward triplets up to time $t-1$ as input, and outputs a probability distribution on the action space, denoted by $\pi_t(\bullet|\bar{S}_t)$. Under π, the decision maker will set $A_t = a$ with probability $\pi_t(a|\bar{S}_t)$ at the tth decision point. When each π_t is binary-valued, π is referred to as a deterministic policy. Suppose the decision rules are stationary over time, i.e., there exists some function π^* such that $\pi_t(\bullet|\bar{S}_t) = \pi^*(\bullet|S_t)$ almost surely for any t, then π is referred to as a *stationary* policy and we write $\pi^*(a|s) = \pi(a|s)$ for any (a, s).

Our objective is to learn an optimal policy based on the observed data to maximise the expected discounted cumulative reward received in the future, starting from time $T+1$,

$$\arg\max_{\pi} \mathbb{E}\Big\{ \underbrace{\mathbb{E}^{\pi}\Big(\sum_{t\geq 0} \gamma^t R_{t+T+1} | S_{T+1} \Big)}_{V^{\pi}_{(T+1):\infty}(S_{T+1})} \Big\}, \tag{1}$$

where $0 < \gamma < 1$ denotes a pre-specified discounted factor that balances the trade-off between the immediate and future outcomes, the first expectation is taken with respect to the marginal distribution of S_{T+1} and the second expectation $\mathbb{E}^{\pi}$ is taken by assuming that all the actions are assigned according to π after time T. $V^{\pi}_{(T+1):\infty}$ is referred to as the (state) value function, as it corresponds to the expected return conditional on the state S_{T+1}.

2.2 The Stationarity Assumption and the Optimal Policy

As commented in the introduction, most existing state-of-the-art RL algorithms focus on a stationary environment. They model the observed data history using the Markov decision process model (Puterman, 1994) and rely on the following assumptions:

MAST (Markov assumption with stationary transitions) There exists some transition function

F and a sequence of i.i.d. random noises $\{\varepsilon_t\}_t$ such that each ε_t is independent of $\{(A_j, S_j)\}_{j\leq t}$ and $\{R_j\}_{j<t}$, and that

$$S_{t+1} = F(S_t, A_t, \varepsilon_t).$$

CMIAST (Conditional mean independence assumption with stationary rewards) There exists some reward function r such that

$$\mathbb{E}(R_t|A_t, \bar{S}_t) = r(A_t, S_t).$$

We make a few remarks. First, by definition, F defines the conditional distribution of the future state given the current state-action pair whereas r corresponds to the conditional mean function of the reward.

Second, both conditions impose certain conditional independence assumptions on the data trajectory. In particular, notice that both F and r are independent of the past state-action-reward triplet given the current state-action pair. These conditional independence assumptions are testable from the observed data; see e.g. the test in Shi et al. (2020b). In practice, one can construct the state by concatenating measurements over sufficiently many decision points to ensure that these conditional independence assumptions hold.

Finally, both conditions require F and r to be stationary over time. Under the stationarity assumption, the value function in (1) is time-homogeneous. More importantly, such an assumption guarantees the existence of an optimal stationary policy π^{opt} whose value $V^{\pi^{opt}}_{(T+1):\infty}(s)$ is no worse than $V^{\pi}_{(T+1):\infty}(s)$ for any history dependence policy π and any s (Puterman, 1994). It allows us to focus on the class of stationary policies and substantially simplifies the calculation. However, as we have commented in the introduction, the stationarity assumption could be violated in some applications, invalidating many RL algorithms developed in the literature.

2.3 Q-Learning

We review Q-learning, one of the most popular algorithms developed under the stationarity assumption. It is model-free in the sense that the optimal policy is derived without directly

estimating the transition and reward functions. We begin by introducing the state-action value function, better known as the Q-function, defined as

$$Q^{\pi}_{t:\infty}(a,s) = \mathbb{E}^{\pi}\left(\sum_{k\geq 0}\gamma^k R_{t+k}|A_t = a, S_t = s\right),$$

for any policy π. Under stationarity, we have $Q^{\pi}_{t:\infty} = Q^{\pi}$ for any t.

The following observation forms the basis of Q-learning: Under MAST and CMIAST, there exists an optimal Q-function Q^{opt} such that the optimal stationary policy is greedy with respect to Q, i.e.,

$$\pi^{opt}(a|s) = \begin{cases} 1, & \text{if } a = \arg\max_{a'} Q^{opt}(a',s); \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

For the ease of notation, we use $\pi^{opt}(s)$ to denote the action a' that maximises $Q^{opt}(a',s)$.

In addition, Q^{opt} satisfies the following Bellman optimality equation,

$$\mathbb{E}\left\{R_t + \gamma \max_a Q^{opt}(a, S_{t+1})|A_t, S_t\right\} = Q^{opt}(A_t, S_t), \tag{3}$$

almost surely for any t.

Existing Q-learning type algorithms propose to learn Q^{opt} based on the Bellman optimality equation and derive the estimated optimal policy based on (2). Examples include the tabular Q-learning algorithm (Watkins and Dayan, 1992), the gradient Q-learning algorithm (Maei et al., 2010; Ertefaie and Strawderman, 2018), FQI, double Q-learning (Hasselt, 2010) and deep Q-network (DQN, Mnih et al., 2015) among many others.

3 Learning under Nonstationarity

We consider policy optimisation in a nonstationary environment. To deal with nonstationarity, we relax MAST and CMIAST by allowing the transition function $\mathcal{P}$ and the reward function r to depend on the decision time t. This yields the following set of conditions.

MA (Markov assumption) There exists some transition function F_t such that $S_{t+1} = F_t(S_t, A_t, \varepsilon_t)$.

CMIA (Conditional mean independent assumption) There exists some reward function r_t such that $\mathbb{E}(R_t|A_t, \bar{S}_t) = r_t(A_t, S_t)$.

As we have commented in Section 2.2, these two conditions are mild. In practice, we can concatenate measurements over sufficiently many decision points to construct the state to convert a nonMarkov process to satisfy these conditions (Mnih et al., 2015; Shi et al., 2020b).

In view of (1), since the future transition and reward functions are unknown to us, the control problem is impossible to solve without additional assumptions. Toward that end, we assume that the reward and transition function are locally stationary at time T. More specifically, we assume

$$\begin{aligned} \sup_{T\le t\le T+M}\sup_{a,s,\mathbb{S}}|\mathbb{P}(F_t(s,a,\varepsilon_1)\in\mathbb{S})-\mathbb{P}(F_{t+1}(s,a,\varepsilon_1)\in\mathbb{S})| &= o(1), \\ \sup_{T\le t\le T+M}\sup_{a,s}|r_t(a,s)-r_{t+1}(a,s)| &= o(1), \end{aligned} \tag{4}$$

for some large integer $M > 0$, where $o(1)$ denotes some quantity that decays to zero as the number of observations $N \times T$ diverges to infinity, and $\mathbb{S}$ denotes an arbitrary measurable subset of the state space. The first condition in (4) essentially requires the total variation norm between the two conditional distributions $S_{t+1}|S_t = s, A_t = a$ and $S_{t+2}|S_{t+1} = s, A_{t+1} = a$ to be negligible for any a, s, t.

Under this assumption, we have the following results.

Lemma 1. *Suppose* (4) *holds and $M \to \infty$ as $NT \to \infty$. Then*

$$\sup_{\pi}|\mathbb{E}V^{\pi}_{(T+1):\infty}(S_{T+1})-\mathbb{E}V^{\pi}_{T}(S_{T+1})| = o(1),$$

where V^{π}_T denotes the value function with the future transition and reward functions $\{\mathcal{P}_t\}_{t>T}$, $\{r_t\}_{t>T}$ replaced by $\mathcal{P}_T$ and r_T, respectively.

Consequently, any policy that maximises $\mathbb{E}V^{\pi}_T(S_{T+1})$ approximately maximises $\mathbb{E}V^{\pi}_{(T+1):\infty}(S_{T+1})$ as well. Since the transition and reward functions are invariant in the definition of V^{π}_T, there exists an optimal stationary policy π^{opt}_T that maximises $\mathbb{E}V^{\pi}_T(S_{T+1})$. To identify π^{opt}_T, we propose to use Q-learning. It suffices to learn the optimal Q-function Q^{opt}_T, which is a version of Q^{opt} with the transition and reward function equal to $\mathcal{P}_T$ and r_T, respectively. Similarly, we define Q^{opt}_t to be the optimal Q-function with the transition and reward function equal to $\mathcal{P}_t$ and r_t, for any t.

We next introduce our method to estimate Q_T^{opt}. Due to the potential nonstationarity in the data, it is not desired to apply existing Q-learning type algorithms to all the data. Our solution is to choose a decision point T^* such that $\{Q_t^{opt}\}_{T^* \le t < T}$ are close to Q_T^{opt} and apply Q-learning to the data subset $\{(S_{i,t}, A_{i,t}, R_{i,t})\}_{1 \le i \le N, T^* \le t \le T}$. It remains critical to determine T^*. A large T^* would yield a biased Q-estimator whereas a small T^* would limit the size of the reduced data subset, yielding a very noise policy.

Toward that end, we propose a data-adaptive method to determine T^*. Specifically, for any candidate change point location t, we first develop a nonparametric test to test whether the optimal Q-function $\{Q_j^{opt}\}_j$ is stationary in the time interval $[t, T]$. We next sequentially apply the proposed test to the time interval $[T - \kappa, T]$ for a monotonic increasing sequence of κ, denoted by $\{\kappa_j\}_j$, for change point detection. Let $T - \kappa_{j_0}$ be the latest change point location. Then we set $T^* = T - \kappa_{j_0 - 1}$. We detail the proposed test in the next section.

To conclude this section, we discuss the potential sources of nonregularity of the observed data. First, the marginal distribution of S_t and the conditional distribution of A_t given the past data history, also known as the behavior policy, are allowed to vary over time. These sources of nonstationarity are not related to decision making as they do not appear in the Q- or value function. Second, the following sources of nonstationarity will affect the optimal Q-function: $\mathcal{P}_t$ and r_t. We focus on testing the nonstationarity of the optimal Q-function as it completely determines the optimal policy. It is also practically interesting to directly test the nonstationarity of the transition and reward functions. See Section 8.2 for details.

4 Hypothesis Testing for Change Point Detection

We first outline the main idea of the proposed hypothesis testing and change point detection methods. We next describe in detail some major steps. A pseudocode summarising the proposed test is presented in Algorithm 1. We remark that our procedure is motivated by existing test techniques in the time series literature and have the practically desirable property of detecting both abrupt and gradual changes that are common in RL applications. It allows Q_t^{opt} to be either piecewise or smooth as a function of t. In Figure 1, we depict two optimal Q-functions

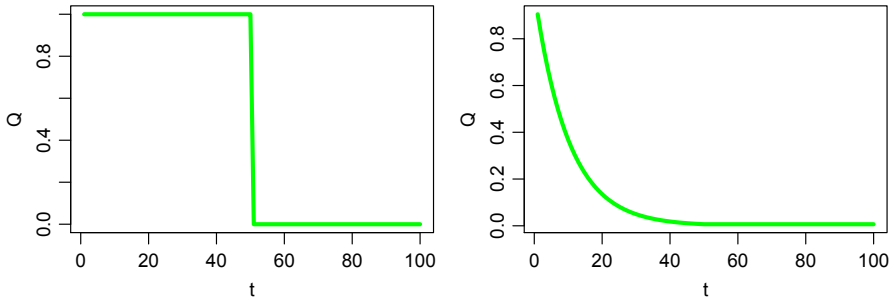


Figure 1: Examples of optimal Q-functions at a given state-action pair with an abrupt change point (left panel) and a gradual change point (right panel).

Algorithm 1 Testing Stationarity for the Optimal Q-Function.

Input: The data $\{(S_{i,t}, A_{i,t}, R_{i,t})\}_{1\le i\le N, T_0\le t\le T}$, and the significance level $0 < \alpha < 1$.

Step 1. For each $T_0 - \epsilon T \le u \le (1-\epsilon)T$, employ the gradient Q-learning or fitted Q-iteration algorithm to compute an estimated Q-function $\widehat{Q}_{[T_0,u]}$ and $\widehat{Q}_{[u,T]}$.

Step 2. Construct the CUSUM-type test statistics TS_1, TS_∞ or $\text{TS}_{n,\infty}$, according to (7), (8) and (9).

Step 3. Employ multiplier bootstrap to compute the bootstrapped test statistic TS_1^b, TS_∞^b or $\text{TS}_{n,\infty}^b$. See (14), (15) and (16). Calculate the p-value according to (17).

Output: Reject the null hypothesis if the p-value is smaller than α.

with an abrupt and a gradual change point.

4.1 The Main Idea

We focus on the null hypothesis that $Q_t^{opt} = Q^{opt}$ for any $t \in \{T_0, T_0+1, \cdots, T\}$ and some integer $T_0 < T$. We propose an integral-type and a maximum-type test statistic. Both test statistics require to estimate the optimal Q-function. First, we propose to use series/sieve method to model Q^{opt}. There are two major advantages of using the sieve method here. First, it ensures the resulting Q-estimator has a tractable limiting distribution (see e.g., Theorem 3), which in turn enables us to derive the asymptotic distribution of the test statistic. Second, the number of sieves (e.g., basis functions) is allowed to diverge with the sample size, alleviating the bias resulting from model misspecification.

Specifically, we propose to model $Q^{opt}(a,s)$ by $\phi_L^\top(a,s)\beta^*$ for some $\beta^* \in \mathbb{R}^L$ where $\phi_L(a,s)$ denotes a vector consisting of L basis functions. In the tabular case where both the state and the action spaces are discrete, one could use a lookup table and set

$$\phi_L(a,s) = [\mathbb{I}\{(a,s) = (a_1,s_1)\}, \cdots, \mathbb{I}\{(a,s) = (a_L,s_L)\}]^\top$$

where $\{(a_j,s_j)\}_j$ correspond to the set consisting all possible action-state pairs. In the case where the action space is discrete and the state space is continuous, we recommend to set

$$\phi_L(a,s) = [\mathbb{I}(a = a_1)\Phi^\top(s), \cdots, \mathbb{I}(a = a_m)\Phi^\top(s)]^\top, \tag{5}$$

where $\{a_j\}_j$ corresponds to the action space and Φ denotes some set of basis functions on the state space, such as power series, Fourier series, splines or wavelets (see e.g., Judd, 1998). Our practical implementation uses the random Fourier features; see Section 6.1 for details. As we have commented in the introduction, if the transition and reward functions are stationary over time, the proposed test controls the type-I error even when the approximation error decays at a rate that is slower than $O\{(NT)^{-1/2}\}$. This implies that the proposed test will not be overly sensitive to the choice of the number of basis functions and "undersmoothing" is not required to guarantee its validity. In practice, we could employ cross validation to determine the number of basis functions; see Section 4.2 for details. However, the smaller the approximation error, the higher the power.

Second, for any time interval $[T_1,T_2] \subseteq [T_0,T]$ with $T_1 = T_0$ or $T_2 = T$, we compute the estimator $\widehat{\beta}_{[T_1,T_2]}$ for β^* using data collected from this interval. A key observation is that, when $Q^{opt}(a,s) = \phi_L^\top(a,s)\beta^*$, it follows from the Bellman optimality equation (3) that β^* satisfies the following equation:

$$\mathbb{E}\phi_L(A_t,S_t)\left\{R_t + \gamma \max_a \beta^{*\top}\phi_L(a,S_{t+1}) - \beta^{*\top}\phi_L(A_t,S_t)\right\} = 0.$$

This motivates us to compute $\widehat{\beta}_{[T_1,T_2]}$ by solving the following estimating equation,

$$\sum_{i=1}^{N}\sum_{t=T_1}^{T_2-1}\phi_L(A_{i,t},S_{i,t})\left\{R_{i,t} + \gamma \max_a \beta^\top\phi_L(a,S_{i,t+1}) - \beta^\top\phi_L(A_{i,t},S_{i,t})\right\} = 0. \tag{6}$$

Under the null hypothesis, it follows from (3) that the left-hand-side (LHS) has zero mean when $\beta = \beta^*$. Consequently, the above estimating equation is consistent. However, it remains challenging to compute $\widehat{\beta}_{[T_1,T_2]}$ due to the existence of the non-smooth max operator in the curly brackets. In practice, we could either employ the gradient Q-learning algorithm or the fitted Q-iteration algorithm to solve the estimating equation. See Section 4.2 for more details. Based on this estimator, we formally define the estimated Q-function $\widehat{Q}_{[T_1,T_2]}(a,s)$ as $\phi_L^\top(a,s)\widehat{\beta}_{[T_1,T_2]}$.

Third, for any candidate change point u, we use $\widehat{Q}_{[u,T]}(a,s) - \widehat{Q}_{[T_0,u-1]}(a,s) = \phi_L^\top(a,s)(\widehat{\beta}_{[u,T]} - \widehat{\beta}_{[T_0,u-1]})$ to measure the difference in the optimal Q-function before and after the change point. Based on this measure, we propose an ℓ_1-type, a maximum-type test statistic, and a normalised maximum-type test statistic, given by

$$\text{TS}_1 = \max_{T_0+\epsilon T<u<(1-\epsilon)T} \sqrt{\frac{(u-T_0)(T-u)}{(T-T_0)^2}} \left\{ \frac{1}{NT} \sum_{i,t} |\widehat{Q}_{[T_0,u]}(A_{i,t},S_{i,t}) - \widehat{Q}_{[u,T]}(A_{i,t},S_{i,t})| \right\}, \quad (7)$$

$$\text{TS}_\infty = \max_{T_0+\epsilon T<u<(1-\epsilon)T} \max_{a,s} \sqrt{\frac{(u-T_0)(T-u)}{(T-T_0)^2}} |\widehat{Q}_{[T_0,u]}(a,s) - \widehat{Q}_{[u,T]}(a,s)|, \quad (8)$$

and

$$\text{TS}_{n,\infty} = \max_{T_0+\epsilon T<u<(1-\epsilon)T} \max_{a,s} \sqrt{\frac{(u-T_0)(T-u)}{(T-T_0)^2}} \widehat{\sigma}_u^{-1}(a,s) |\widehat{Q}_{[T_0,u]}(a,s) - \widehat{Q}_{[u,T]}(a,s)|, \quad (9)$$

respectively, where $\widehat{\sigma}_u^2(a,s)$ denotes some consistent variance estimator of $\widehat{Q}_{[T_0,u]}(a,s) - \widehat{Q}_{[u,T]}(a,s)$ whose detailed form is given in Appendix B.2.2 of the supplementary article.

We make a few remarks. First, the three test statistics are very similar to the classical cumulative sum (CUSUM) statistic in change point analysis (Csörgö et al., 1997). According to the weight scale $\sqrt{u(\kappa-u)/\kappa}$, both test statistics assign less weights on the boundary data points. In addition, ϵ denotes some user-specified boundary cut-off parameter. Removing the boundary points is necessary as it is difficult to estimate the Q-function that is close to the endpoints. Such practice is commonly employed in the time series literature for change point detection in non-Gaussian settings (see e.g., Cho and Fryzlewicz, 2012; Yu and Chen, 2021).

Second, the three test statistics differ in the ways they aggregate the estimated changes $|\widehat{Q}_{[T_0,u]} - \widehat{Q}_{[u,T]}|$ over different state-action pairs. The ℓ_1-type test averages the changes with

weights assigned according to the empirical state-action distribution whereas the two maximum-type tests focuses on the largest change in the (normalised) absolute value. Among the two maximum-type test statistics, the normalised test is likely to be more efficient. This is because when $\widehat{Q}_{[T_0,u]}(a,s) - \widehat{Q}_{[u,T]}(a,s)$ is not consistent for some value of (a,s), the argmax $\arg\max_{a,s} |\widehat{Q}_{[T_0,u]}(a,s) - \widehat{Q}_{[u,T]}(a,s)|$ might differ significant from the oracle maximiser $\arg\max_{a,s} |Q^{opt}_{[T_0,u]}(a,s) - Q^{opt}_{[u,T]}(a,s)|$, thus lowering the power of the unnormalised test. The normalised test alleviates this issue by taking standard errors of these estimators into consideration. For state-action pairs with inconsistent Q-values, their differences $\widehat{Q}_{[T_0,u]} - \widehat{Q}_{[u,T]}$ would have large standard errors. As such, those pairs are unlikely to be the argmax. In addition, the normalised test requires a weaker condition to control the type-I error. See Section 5 for details. We also remark that the studentised supremum type statistics have been used in the economics literature (see e.g., Belloni et al., 2015; Chen and Christensen, 2015, 2018).

Third, we develop a bootstrap-based procedure to compute the critical values for TS_1, TS_∞ and $\text{TS}_{n,\infty}$. As we will show in Section 5.3, each estimator $\widehat{\beta}_{[T_1,T_2]}$ computed by solving (6) is asymptotically normal. So is the estimated Q-function. This motivates us to do employ the multiplier bootstrap to approximate the asymptotic distribution of the Q-estimator and the resulting test statistics. The p-value is obtained based on the empirical distribution of the bootstrap samples. See Section 4.3 for details. Under a given significance level α, we reject the test when the p-value is smaller than α.

Finally, we sequentially apply the proposed test to identify the most recent change point. We begin by specifying a monotonically increasing sequence $\{\kappa_j\}_j$. We next apply our test to examine whether the optimal Q-function is stationarity on the time interval $[T-\kappa_j, T]$, for $j = 1, 2, \cdots$. This yields a set of p-values. When the data interval consists of a single change point, those significant p-values are likely to be monotonic over time. This allows us to apply the isotonic regression to fit these p-values and set the change point to be the location whose fitted value is smaller than the nominal level for the first time. Compared to the standard sequential method that selects the most recent change point location whose p-value is significant, the proposed methods utilise the monotonicity property and can potentially estimate

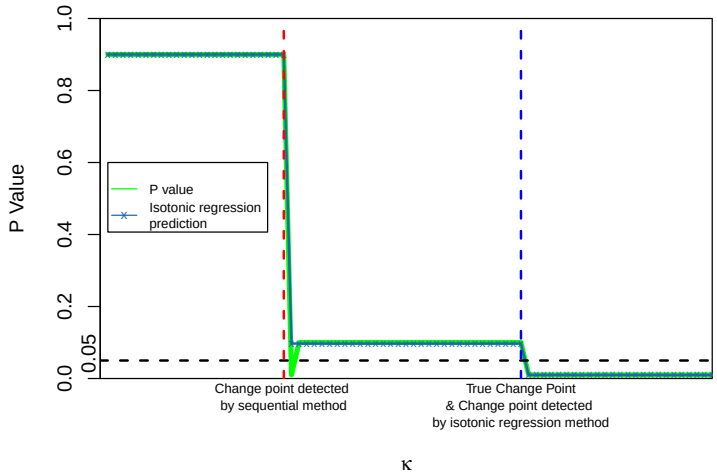


Figure 2: An illustrative example where isotonic regression avoids premature change-point declaration. The p-value next to the red dashed line is below the significance level 0.05 by chance, leading to a false positive. The sequential method is sensitive to the false positive and incorrectly identifies the change point location, whereas the proposed method avoids false positives and identifies the true change point location (next to the blue dashed line) due to that it borrows information from nearby p-values.

the change point location more accurately. See Figure 2 for an graphical illustration. We also remark that our procedure is applicable to settings with multiple change points as well. We elaborate in detail in Section 7. If no changes are detected, we propose to use all the observed data for policy optimisation.

4.2 Estimation of the Q-Function

In this section, we present two algorithms to compute $\widehat{\beta}_{[T_1,T_2]}$ that satisfies the estimating equation (6). Statistical properties of the estimated Q-function are discussed in Section 5.3.

Algorithm 2 (Greedy Gradient Q-Learning). The greedy gradient Q-learning is based on the stochastic gradient descent algorithm. For any β, define the temporal difference error

$$\delta_{i,t}(\beta) = R_{i,t} + \gamma \max_a \beta^\top \phi_L(a, S_{i,t+1}) - \beta^\top \phi_L(A_{i,t}, S_{i,t}). \tag{10}$$

To solve the estimating equation (6), it suffices to minimise the following objective function

$$M(\beta) = \left\{ \sum_{i,t} \phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}(\beta) \right\}^\top \underbrace{\left\{ \sum_{i,t} \phi_L(A_{i,t}, S_{i,t}) \phi_L^\top(A_{i,t}, S_{i,t}) \right\}^{-1}}_{\varpi(\beta)} \left\{ \sum_{i,t} \phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}(\beta) \right\}.$$

Due to the existence of the non-smooth max operator in the temporal difference error, the above loss is not a smooth function of β. Toward that end, we consider calculating the Frechet sub-gradient of $M(\beta)$, given by

$$\nabla M(\beta) = -\left\{ \sum_{i,t} \phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}(\beta) \right\} + \gamma \left\{ \sum_{i,t} \phi_L(A_{i,t}, S_{i,t}) \phi_L(\pi_\beta(S_{i,t+1}), S_{i,t+1}) \right\}^\top \varpi(\beta), \quad (11)$$

where π_β denotes the optimal policy estimated as a function of β, i.e., $\pi_\beta(s) = \arg\max_a \phi_L(a, s)^\top \beta$. See e.g., Ertefaie and Strawderman (2018) for details.

Maei et al. (2010) proposed to use the weight-doubling trick developed by Sutton et al. (2008) for parameter estimation. The main idea is to treat ϖ as an additional parameter, and to update both β and ϖ simultaneously during each iteration using stochastic gradient descent. The purpose of employing the weight-doubling trick is to alleviate the potential bias of the sub-gradient. Specifically, at the kth iteration, we start with the first individual's trajectory, obtain β_{k+1} and ϖ_{k+1} from the following iterative equations:

$$\begin{aligned}
\beta_{k+1} &= \beta_k + \alpha_{k,1} \nu \sum_t [\phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}(\beta_k) - \gamma \{\varpi_k^\top \phi_L(A_{i,t}, S_{i,t})\} \phi_L(\pi^*_{\beta_k}(S_{i,t+1}), S_{i,t+1})], \\
\varpi_{k+1} &= \varpi_k + \alpha_{k,2} \nu \sum_t \phi_L(A_{i,t}, S_{i,t}) [\delta_{i,t}(\beta_k) - \{\varpi_k^\top \phi_L(A_{i,t}, S_{i,t})\}],
\end{aligned}$$

for some tuning parameters $\{\alpha_{k,1}\}_k, \{\alpha_{k,2}\}_k, \nu$, and continue updating the parameters to the last individual. The algorithm is terminated when the difference $\|\beta_{k+1} - \beta_k\|_2$ is smaller than some pre-determined threshold. It is worth mentioning that the step size $\alpha_{k,1}$ is required to converge to zero at a faster rate than $\alpha_{k,2}$, as k diverges to infinity, to ensure the resulting algorithm converges. Ertefaie and Strawderman (2018) established the consistency and asymptotic normality of the resulting estimator by assuming no approximation error exists, e.g., $Q^{opt}(a, s) = \phi_L^\top(a, s)\beta^*$ for any a, s.

Algorithm 3 (Fitted Q-Iteration). The main idea of FQI is to iteratively update the Q-function based on the Bellman optimality equation. During each iteration, it computes $Q^{(k+1)}$ by minimizing

$$Q^{(k+1)} = \arg\min_{Q} \sum_{i,t} \Big\{ R_{i,t} + \gamma \max_{a} Q^{(k)}(a, S_{i,t+1}) - Q(A_{i,t}, S_{i,t}) \Big\}^2 .$$

The above optimisation can be cast into a supervised learning problem with $\{R_{i,t}+\gamma \max_a Q^{(k)}(a, S_{i,t+1})\}_{i,t}$ as the responses and $\{(A_{i,t}, S_{i,t})\}_{i,t}$ as the predictors. When a linear sieve model is imposed for the optimal Q-function, the estimator $\widehat{\beta}_{[T_1,T_2]}$ can be iteratively updated using ordinary least-square regression (OLS). In Section 5.3, we show that such an algorithm converges in the sense that the resulting estimator solves the estimating equation in (6), and derive the asymptotic distribution of $\widehat{\beta}_{[T_1,T_2]}$.

4.3 Bootstrap for P-Value

We develop a multiplier bootstrap method to obtain the p-values. The idea is to generate bootstrap samples to approximate the limiting distribution of TS_1 and TS_∞, defined in (7) and (8), respectively. A key observation is that, under the null hypothesis, when the Q-function is well-approximated and the optimal policy is uniquely defined, the estimated Q-function $\phi(a,s)^\top \widehat{\beta}_{[T_1,T_2]}$ has the following linear representation:

$$\phi(a,s)\widehat{\beta}_{[T_1,T_2]} - Q^{opt}(a,s) = \frac{1}{N(T_2-T_1)} \phi_L^\top(a,s) W_{[T_1,T_2]}^{-1} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}^* + o_p(1), \quad (12)$$

where

$$\begin{aligned} W_{[T_1,T_2]} &= \frac{1}{T_2-T_1} \sum_{t=T_1}^{T_2-1} \mathbb{E}\phi_L(A_{i,t}, S_{i,t}) \{\phi_L(A_{i,t}, S_{i,t}) - \gamma \phi_L(\pi^{opt}(S_{i,t+1}), S_{i,t+1})\}^\top, \\ \delta_{i,t}^* &= R_{i,t} + \gamma \max_a Q^{opt}(a, S_{i,t+1}) - Q^{opt}(A_{i,t}, S_{i,t}). \end{aligned}$$

We refer to the proof of Theorem 3 in the supplementary article for details. By the Bellman optimality equation, the leading term on the RHS of (12) forms a mean-zero martingale. When its quadratic variation process converges, it follows from the martingale central limit theorem (McLeish, 1974) that $\widehat{\beta}_{[T_1,T_2]}$ is asymptotically normal.

It follows from (12) that

$$\widehat{Q}_{[T_0,u]}(a,s) - \widehat{Q}_{[u,T]}(a,s) = \frac{1}{N(u-T_0)}\phi_L^\top(a,s)W_{[T_0,u]}^{-1}\sum_{i=1}^{N}\sum_{t=T_0}^{u-1}\phi_L(A_{i,t},S_{i,t})\delta_{i,t}^* \\ -\frac{1}{N(T-u)}\phi_L^\top(a,s)W_{[u,T]}^{-1}\sum_{i=1}^{N}\sum_{t=u}^{T-1}\phi_L(A_{i,t},S_{i,t})\delta_{i,t}^* + o_p(1). \tag{13}$$

This motivates us to construct bootstrap samples that approximate the asymptotic distribution of the leading term on the RHS of (13). Specifically, we consider the bootstrap sample $\widehat{Q}^b_{[T_0,u]}(a,s) - \widehat{Q}^b_{[u,T]}(a,s)$ where

$$\widehat{Q}^b_{[T_1,T_2]}(a,s) = \frac{1}{N(T_2-T_1)}\phi_L^\top(a,s)\widehat{W}_{[T_1,T_2]}^{-1}\sum_{i=1}^{N}\sum_{t=T_1}^{T_2-1}\phi_L(A_{i,t},S_{i,t})\delta_{i,t}(\widehat{\beta}_{[T_1,T_2]})e_{i,t}, \quad \forall T_1,T_2,$$

where $\delta_{i,t}(\beta)$ is the temporal difference error defined in (10), and $\widehat{W}_{[T_1,T_2]}$ denotes some consistent estimator for $W_{[T_1,T_2]}$, defined by

$$\widehat{W}_{[T_1,T_2]} = \frac{1}{N(T_2-T_1)}\sum_{i=1}^{N}\sum_{t=T_1}^{T_2-1}\phi_L(A_{i,t},S_{i,t})\{\phi_L(A_{i,t},S_{i,t}) - \gamma\phi_L(\pi_{\widehat{\beta}_{[T_1,T_2]}}(S_{i,t+1}),S_{i,t+1})\}^\top,$$

π_β is defined in (11), and $\{e_{i,t}\}_{i,t}$ correspond to a sequence of i.i.d. standard normal random errors that are independent of the observed data. This yields the bootstrapped statistics,

$$\text{TS}_1^b = \max_{T_0+\epsilon T<u<(1-\epsilon)T}\sqrt{\frac{(u-T_0)(T-u)}{(T-T_0)^2}}\left\{\frac{1}{NT}\sum_{i,t}|\widehat{Q}^b_{[T_0,u]}(A_{i,t},S_{i,t}) - \widehat{Q}^b_{[u,T]}(A_{i,t},S_{i,t})|\right\}, \tag{14}$$

$$\text{TS}_\infty^b = \max_{T_0+\epsilon T<u<(1-\epsilon)T}\max_{a,s}\sqrt{\frac{(u-T_0)(T-u)}{(T-T_0)^2}}|\widehat{Q}^b_{[T_0,u]}(a,s) - \widehat{Q}^b_{[u,T]}(a,s)|, \tag{15}$$

$$\text{TS}_{n,\infty}^b = \max_{T_0+\epsilon T<u<(1-\epsilon)T}\max_{a,s}\sqrt{\frac{(u-T_0)(T-u)}{(T-T_0)^2}}\widehat{\sigma}_u^{-1}(a,s)|\widehat{Q}^b_{[T_0,u]}(a,s) - \widehat{Q}^b_{[u,T]}(a,s)|. \tag{16}$$

In Section 5.2, we show that under the null hypothesis, the asymptotic distributions of TS_1, TS_∞ and $\text{TS}_{n,\infty}$ can be well-approximated by the conditional distributions of TS_1^b, TS_∞^b and $\text{TS}_{n,\infty}^b$ given the observed data. The corresponding p-values are given by

$$\mathbb{P}(\text{TS}_1^b > \text{TS}_1|\text{Data}), \quad \mathbb{P}(\text{TS}_\infty^b > \text{TS}_\infty|\text{Data}) \text{ and } \mathbb{P}(\text{TS}_{n,\infty}^b > \text{TS}_{n,\infty}|\text{Data}) \tag{17}$$

respectively. We reject the null when the p-value is smaller than a given significance level α.

5 Theory

We first introduce the technical conditions. We next establish the consistency of the proposed test. Finally, we investigate the asymptotic distribution of the Q-function estimator. To simplify the theoretical analysis, we focus on the setting where the action space $\mathcal{A} = \{0, 1, \cdots, m-1\}$ is discrete with m being the total number of available actions, the state space is continuous, and $\phi_L(a, s) = [\mathbb{I}(a = 0)\Phi^\top(s), \cdots, \mathbb{I}(a = m-1)\Phi^\top(s)]^\top \in \mathbb{R}^L$ for some set of basis functions Φ on the state space $\mathcal{S}$. Without loss of generality, we assume $\mathcal{S} = [0, 1]^d$ where d denotes the dimension of the state. This assumption could be relaxed by assuming $\mathcal{S}$ is a compact subset of $\mathbb{R}^d$. We use $p_t(\bullet|a, s)$ to denote the probability density function of $F_t(a, s, \varepsilon)$. In other words, p_t corresponds to the density function of S_{t+1} given $(A_t, S_t) = (a, s)$. We use π^{opt} to denote the optimal policy that is greedy with respect to Q_t^{opt}.

As commented in the introduction, all the theories in this section are established under a bidirectional asymptotic framework. They are valid as either N or T diverges to infinity.

5.1 Technical Conditions

For any T_1, T_2, to allow for model misspecification, we define the population-level least false parameter $\beta^*_{[T_1,T_2]}$ as follows,

$$\beta^*_{[T_1,T_2]} = \left[\sum_{t=T_1}^{T_2-1} \mathbb{E}\phi_L(A_t, S_t)\{\phi_L(A_t, S_t) - \gamma\phi_L(\pi_t^{opt}(S_{t+1}), S_{t+1})\}^\top\right]^{-1} \left\{\sum_{t=T_1}^{T_2-1} \mathbb{E}\phi_L(A_t, S_t)R_t\right\}.$$

We introduce the following conditions.

(A1) When $Q_t^{opt} = Q^{opt}$ is stationary for any integer $t \in [T_1, T_2 - 1]$, $\widehat{\beta}_{[T_1,T_2]}$ has the following linear representation:

$$\widehat{\beta}_{[T_1,T_2]} - \beta^*_{[T_1,T_2]} = \frac{1}{N(T_2 - T_1)} W^{-1}_{[T_1,T_2]} \sum_{i=1}^{N}\sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t})\delta^*_{i,t} - b_{[T_1,T_2]} + O(N^{-c_1}T^{-c_1}), \quad (18)$$

for some constant $c_1 > 1/2$, with probability at least $1 - O(N^{-1}T^{-1})$, where the big-O term

is uniform in $\{(T_1, T_2) : T_2 - T_1 \geq \epsilon T\}$ and $b_{[T_1,T_2]}$ is given by

$$\frac{W_{[T_1,T_2]}^{-1}}{T_2 - T_1} \sum_{t=T_1}^{T_2-1} \mathbb{E}\phi_L(A_t, S_t)[Q_t^{opt}(A_t, S_t) - \gamma \max_a Q^{opt}(a, S_{t+1}) - \{\phi_L(A_t, S_t) - \gamma\phi_L(\pi_t^{opt}(S_{t+1}), S_{t+1})\}^\top \beta_{[T_1,T_2]}^*].$$

(A2) $\{p_t(s'|a, s)\}_t$, $\{r_t(a, s)\}_t$ are p-smooth (Hölder smooth) functions of s, i.e., $p_t(s'|a, \bullet), r_t(a, \bullet) \in \Lambda(p, c)$ (see Appendix A) for some constant $c > 0$ and some $p \geq 1$.

(A3) There exists some constant $c_2 > 1/2$ such that

$$\sup_{f \in \Lambda(p,C)} \inf_{\beta \in \mathbb{R}^d} \sup_s |\Phi^\top(s)\beta - f(s)| = O(L^{-c_2}), \tag{19}$$

for any sufficiently large constant $C > 0$, where L denotes the number of basis functions.

(A4) $\lambda_{\max}\{\int_s \Phi(s)\Phi^\top(s)ds\} = O(1)$, $\sup_s \|\Phi(s)\|_1 = O(\sqrt{L})$.

(A5) $\sup_{a,t} \mathbb{E}\|F_t(s, a, \varepsilon_1) - F_t(s', a, \varepsilon_1)\|_2 \leq \rho\|s - s'\|_2$ for some $0 \leq \rho < 1$ and $\sup_{t,a,s} \|F_t(s, a, \varepsilon) - F_t(s, a, \varepsilon')\|_2 = O(\|\varepsilon - \varepsilon'\|_2)$. Suppose ε_1 has sub-exponential tails, i.e., for any j, $\mathbb{E}|\varepsilon_{1,j}|^k \leq k! c_3 c_4^{k-2}$ for some constants $c_3, c_4 > 0$ where $\varepsilon_{1,j}$ denotes the jth element of the vector ε_1.

(A6) $\lambda_{\min}\mathbb{E}[\phi_L(A_t, S_t)\phi_L(A_t, S_t)^\top - \gamma^2\phi_L(\pi_t^{opt}(S_t), S_t)\phi_L^\top(\pi_t^{opt}(S_t), S_t)]$ is uniformly bounded away from zero for any t.

(A7) The optimal policy π_t^{opt} is unique for any t.

(A8) The data are generated by a Markov policy $\pi^b = \{\pi_t^b\}_t$, i.e.,

$$\mathbb{P}(A_t = a|\{S_j\}_{j \leq t}, \{A_j\}_{j<t}, \{R_j\}_{j<t}) = \pi_t^b(a|S_t),$$

for any t, where π_t^b is independent of the observed data history.

(A9) The margin $Q_t^{opt}(\pi_t^{opt}(s), s) - \max_{a \in \mathcal{A} \setminus \pi_t^{opt}(s)} Q_t^{opt}(a, s)$ is bounded away from zero, uniformly for any s and t.

(A10) Suppose $\sup_{s_1 \neq s_2}\{\|\Phi(s_1) - \Phi(s_2)\|_2 / \|s_1 - s_2\|_2\} = O(\sqrt{L})$.

We make a few remarks. First, the first two terms on the right-hand-side (RHS) of (18) characterise the variation and bias of $\widehat{\beta}_{[T_1,T_2]} - \beta_{[T_1,T_2]}^*$, respectively. Under settings where L is fixed and the linear model is correctly specified, Ertefaie and Strawderman (2018) obtained a

similar linear representation for $\widehat{\beta}_{[T_1,T_2]}$ computed via the gradient greedy Q-learning algorithm. In Section 5.3, we will show that (A1) holds when FQI is used to compute $\widehat{\beta}_{[T_1,T_2]}$.

Second, the smoothness conditions in (A2) are commonly imposed in the sieve estimation literature (see e.g. Huang, 1998; Chen and Christensen, 2015). These conditions are used to bound the approximation error of the Q-function. When p is an integer, the p-smoothness requires a function to have bounded derivatives up to the pth order. More generally, the definition of the class of p-smooth functions $\Lambda(p, c)$ can be found in Appendix A. Under (A2), the Q-function at each time is p-smooth as a function of the state as well (Fan et al., 2020).

Third, Condition (A3) essentially requires that the class of p-smooth functions could be uniformly approximated by functions that belong to the linear sieve space. It holds for a wide variety of sieve classes. When tensor product polynomial, trigonometric polynomial, B-spline or wavelet bases are employed, (19) is automatically satisfied with $c_2 = p/d$ (DeVore and Lorentz, 1993; Huang, 1998; Chen, 2007). (A3) thus holds as long as $p > d/2$.

Fourth, Condition (A4) is automatically satisfied when a tensor-product B-spline or wavelet basis is used, where the constituents of the wavelet basis are arranged such that the coarser level basis functions come ahead of the finer level ones. Specifically, the first part of (A4) can be proven based on the proof of Theorem 3.3 of Burman and Chen (1989) and that of Theorem 5.1 of Chen and Christensen (2015). The second part of (A4) follows from the fact that the number of nonzero elements in the vector $\Phi(s)$ is bounded by some universal constant and that each of the basis function is uniformly bounded by $O(\sqrt{L})$.

Fifth, Condition (A5) is needed to establish concentration inequalities for nonstationary Markov chains (Alquier et al., 2019). It allows us to develop a matrix concentration inequalities with nonstationary transition functions, which is needed to prove the validity of the bootstrap method (see Lemma B.2 in the supplementary article for details). This assumption is automatically satisfied when the state satisfies a-time-varying AR(1) process:

$$S_{t+1} = \rho_t S_t + \beta_t A_t + \varepsilon_t,$$

for some $\{\rho_t\}_t$ and $\{\beta_t\}$ such that $\sup_t |\rho_t| < 1$, and ε_t has sub-exponential tails. More gener-

ally, it also holds when the auto-regressive model is given by

$$S_{t+1} = f_t(A_t, S_t) + \varepsilon_t,$$

with $\sup_{a,t} |f_t(a,s) - f_t(a,s')| \leq \rho \|s - s'\|_2$ for some $\rho < 1$. When the transition functions are stationary over time, it essentially requires the Markov chain to possess the exponential forgetting property (Dedecker and Fan, 2015).

Sixth, (A6) and (A7) are commonly imposed in the statistics literature on RL (see e.g., Luckett et al., 2020). (A6) is automatically satisfied when the behavior policy that determines A_t is ε-greedy with respect to π_t^{opt} for some $\varepsilon \leq 1 - \gamma^2$ (Shi et al., 2021). (A7) is a necessary condition for establishing the limiting distribution of $\widehat{\beta}_{[T_1,T_2]}$. It is violated in nonregular settings where the optimal policy is not uniquely defined (Chakraborty et al., 2013; Luedtke and van der Laan, 2016; Shi et al., 2020a). The proposed method could be further coupled with data splitting to derive a valid test in nonregular settings. However, the resulting test might suffer from a loss of power, due to the use of data splitting. We discuss this in Section 8.3 in detail.

Seventh, (A8) allows the behavior policy that generates the data to be nonstationary over time. It is automatically satisfied in randomised studies where the behavior policy is usually a constant function of the state.

Eighth, the margin $Q_t^{opt}(\pi_t^{opt}(s), s) - \max_{a \in \mathcal{A} - \pi_t^{opt}(s)} Q_t^{opt}(a, s)$ measures the difference between the state-action value under the best action and the second best action. The condition (A9) is imposed to simplify the theoretical analysis. It could be potentially relaxed to require the margin to converge to zero at certain rate or to require the probability that the margin approaches zero to decay to zero at certain rate (see e.g., Qian and Murphy, 2011; Luedtke and van der Laan, 2016; Hu et al., 2021b).

Finally, (A10) is needed to establish the statistical properties of the maximum-type tests, but is not needed for the ℓ_1-type test. It is automatically satisfied when a tensor-product B-spline is used, since each function in the vector Φ is a Lipschitz continuous function.

5.2 Consistency of the Test

We derive the size and power properties of the proposed test in this section.

Theorem 1 (Size)**.** *Suppose (A1)-(A9) hold and L is proportional to $(NT)^{c_5}$ for some $0 < c_5 < 1/4$. Suppose under the null hypothesis, $\max_{a,s,u} |\phi_L^\top(a,s)(b_{[T_0,u]} - b_{[u,T]})| = O\{(NT)^{-c_6}\}$ and $\max_{a,s,u} \|\phi_L^\top(a,s)(\beta^*_{[T_0,u]} - \beta^*_{[u,T]})\|_2 = O\{(NT)^{-c_6}\}$ for some $c_6 > 1/2$ where $b_{[T_1,T_2]}$ is defined in (A1). Suppose the boundary removal parameter ϵ is proportional to $\log^{-c_7}(NT)$ for some $c_7 > 0$. Then under the null, we have*

$$\sup_z |\mathbb{P}(\sqrt{NT}TS_1^b \le z|Data) - \mathbb{P}(\sqrt{NT}TS_1 \le z)| \xrightarrow{p} 0.$$

In addition, suppose (A10) holds. Then

$$\sup_z |\mathbb{P}(\sqrt{NT}TS_{n,\infty}^b \le z|Data) - \mathbb{P}(\sqrt{NT}TS_{n,\infty} \le z)| \xrightarrow{p} 0.$$

Finally, suppose the constant c_1 in (A1) satisfies that $(NT)^{2c_1-1} \gg L$. Then

$$\sup_z |\mathbb{P}(\sqrt{NT}TS_\infty^b \le z|Data) - \mathbb{P}(\sqrt{NT}TS_\infty \le z)| \xrightarrow{p} 0,$$

as either N or T approaches to infinity.

Theorem 1 implies that the limiting distribution of the proposed test can be well-approximated by the conditional distribution of the bootstrapped statistic given the data. It in turn implies that the rejection probability of the proposed test approaches to the nominal level as the total number of observations diverges to infinity. As commented in the introduction, the derivation of the consistency of the proposed test is complicated due to that we allow L to grow with the number of observations. To address this challenge, we develop a matrix concentration inequality for nonstationary MDP in Lemma B.2.

In the statement of Theorem 1, we require the difference in the approximation error before and after the change point to decay to zero at a rate of $o\{(NT)^{-1/2}\}$. These assumptions are much weaker than requiring the approximation error to be $o\{(NT)^{-1/2}\}$. The latter requires undersmoothing to ensure the bias the Q-estimator to converge to zero at a faster rate than its

standard deviation. To elaborate, notice that when the transition and reward functions are homogeneous over time, we have $b_{[T_0,u]} = b_{[u,T]}$ and $\beta_{[T_0,u]} = \beta_{[u,T]}$. The assumptions in Theorem 1 are automatically satisfied despite that $\phi_L^\top(a,s)b_{[T_0,u]}$ and $\phi_L^\top(a,s)\beta_{[T_0,u]} - Q^{opt}(a,s)$ might converge to zero slower than the parametric rate. As such, undersmoothing is not required and the size property of the proposed test is not overly sensitive to choice of the number of basis functions. The reason why the validity of the proposed test requires a weaker assumption on the approximation error is due to the use of CUSUM-type statistics, which perform scaled differencing at each postulated error location, thereby requiring the difference in approximation errors (rather than absolute approximation errors) to be of a certain order.

Comparatively speaking, the ℓ_1-type test requires weaker conditions than the maximum-type tests. Specifically, the ℓ_1-type test only requires the constant c_1 in (A1) to be larger than $1/2$ whereas the maximum-type test requires $(NT)^{2c_1-1} \gg L$. Moreover, both the normalised and unnormalised maximum-type tests require an additional assumption (A10).

We next establish the power property of the proposed test. In our theoretical analysis, we focus on a particular type of alternative hypothesis where there is a single change-point T^* such that $Q_{T_0}^{opt} = Q_{T_0+1}^{opt} = \cdots = Q_{T^*-1}^{opt}$ and $Q_{T^*}^{opt} = Q_{T^*+1}^{opt} = \cdots = Q_T^{opt}$. We use $\Delta_1 = T^{-1}\sum_a \sum_{t=0}^{T-1} \int_s |Q_{T_0}^{opt}(a,s) - Q_T^{opt}(a,s)|\pi_t(a|s)p_t^b(s)ds$ and $\Delta_\infty = \sup_{a,s} |Q_{T_0}^{opt}(a,s) - Q_T^{opt}(a,s)|$ to characterise the degree of nonstationarity. Specifically, the null holds if Δ_1 or Δ_∞ equals zero and the alternative hypothesis holds if Δ_1 or Δ_∞ is positive. However, we remark that the proposed test is consistent against more general alternative hypothesis as well. See Section 6 for details. Notice that in the definition of Δ_1, we integrate against the observed state-action distribution $T^{-1}\sum_{t=0}^{T-1} \pi_t(a|s)p_t^b(s)$. Alternatively, the reference distribution can be taken for any measure that is absolutely continuous with respect the observed state-action distribution. For any two positive sequences $\{a_{N,T}\}_{N,T}$ and $\{b_{N,T}\}_{N,T}$, the notation $a_{N,T} \gg b_{N,T}$ means that $b_{N,T}/a_{N,T} \to 0$ as $NT \to \infty$.

Theorem 2 (Power)**.** *Suppose (A1)-(A9) hold and L is proportional to $(NT)^{c_5}$ for some $0 < c_5 < 1/2$. Suppose $T_0 + \epsilon T < T^* < (1-\epsilon)T$ and ϵ is proportional to $\log^{-c_7}(NT)$ for some $c_7 > 0$.*

- *If* $\Delta_1 \gg \{\sqrt{L(NT)^{-1}\log(NT)} + L^{-c_2}\}\log^{c_7/2}(NT)$ *and (A10) holds, then the power of the test based on* TS_1 *approaches 1, as either* N *or* T *diverges to infinity;*

- *If* $\Delta_\infty \gg \sqrt{L}\{\sqrt{L(NT)^{-1}\log(NT)} + L^{-c_2}\}\log^{c_7/2}(NT)$*, then the power of the test based on* TS_∞ *approaches 1, as either* N *or* T *diverges to infinity.*

- *If* $\Delta_\infty \gg \{\sqrt{L(NT)^{-1}\log(NT)} + L^{-c_2}\}\log^{c_7/2}(NT)$ *and (A10) holds, then the power of the test based on* $TS_{n,\infty}$ *approaches 1, as either* N *or* T *diverges to infinity.*

The assumption $T_0 + \epsilon T < T^* < (1-\epsilon)T$ is reasonable as we allow ϵ to decay to zero as the number of observations grow to infinity. Under the given assumptions, the bias and standard deviation of the Q-function estimator are proportional to $O(\sqrt{L(NT)^{-1}\log(NT)})$ and $O(L^{-c_2})$, respectively. The conditions on Δ_1 and Δ_∞ essentially require the signal associated with the alternative hypothesis to be much larger than the estimation error. Similar to the findings in Theorem 1, the unnormalised maximum-type test requires a stronger condition on L to detect the alternative hypothesis. To guarantee the proposed test has good power properties, we use cross-validation to select the number of basis functions, as discussed in Section 6.1. This ensures the bias and standard deviation of the Q-function estimator are approximately of the same order of magnitude, thus minimising the requirements for Δ_1 and Δ_∞.

It is worthwhile to mention that establishing the power property of the test requires a less stringent condition on L than deriving the size property. Specifically, we require L to grow at a rate of $o(\sqrt{NT})$ in Theorem 2. In contrast, this condition is strengthened to $L = o(N^{1/4}T^{1/4})$ in Theorem 1.

5.3 Asymptotic Properties of the Q-Function Estimator

In this section, we focus on investigating the asymptotic properties of the estimated Q-function computed by FQI.

Theorem 3. *Suppose (A2)-(A9) hold and* L *is proportional to* $(NT)^{c_5}$ *for some* $0 < c_5 < 1/2$*. Suppose* ϵ *is proportional to* $\log^{-c_7}(NT)$ *for some* $c_7 > 0$*. Suppose that for any interval*

$[T_1, T_2]$ *such that* $Q^{opt}_{T_1} = Q^{opt}_{T_1+1} = \cdots = Q^{opt}_{T_2-1}$*, we have* $p^{opt}_{T_1} = p^{opt}_{T_1+1} = \cdots = p^{opt}_{T_2-1}$ *and* $r^{opt}_{T_1} = r^{opt}_{T_1+1} = \cdots = r^{opt}_{T_2-1}$*. Suppose the maximum number of iterations* K *in FQI satisfies* $\log(NT) \ll K = O(N^{c_8}T^{c_8})$ *for any* $c_8 > 0$*. Then (A1) is satisfied for sufficiently large* NT*.*

We again make a few remarks. First, Theorem 3 implies that the asymptotic linear representation in (A1) holds for the FQI estimator. Under the null hypothesis, it follows from the high-dimensional martingale central limit theorem (Belloni and Oliveira, 2018) that the set of the estimated Q-functions $\{\widehat{Q}_{[T_1,T_2]}\}_{T_1,T_2}$ are jointly asymptotically normal.

Second, as mentioned in the statement of Theorem 3, we require the transition and reward functions to be stationary over time. This is because FQI iteratively updates the Q-function using supervised learning. In contrast, such a condition is not needed when the gradient Q-learning algorithm is employed to learn the optimal Q-function. However, it is worth mentioning that FQI is much easier to implement in practice, as it suffices to implement OLS during each iteration. It could be further extended to employ more general supervised learning algorithms to fit more complicated nonlinear models (e.g., neural networks).

Finally, as commented in the introduction, most works in the literature focused on establishing non-asymptotic error bounds of the FQI estimator. To our knowledge, this is the first work that investigates the limiting distribution of the FQI estimator in infinite horizon settings, based on nonparametric sieve regression.

6 Simulations

In this section, we conduct simulation studies to evaluate the finite sample performance of the proposed methods and compare against common alternatives. In Section 6.1, we detail the implementation the proposed tests (integral, normalised, unnormalised). Section 6.2 presents results based on four generative models with different nonstationarity scenarios (see Table 1). In Section 6.3, we simulate data to mimic data setup in the motivating application of IHS. All simulation results are aggregated over 100 replications.

6.1 Implementation Details

To implement the proposed tests, the boundary removal parameter ϵ is set to 0.1; 2000 bootstrap samples are generated to compute p-values. In our simulations, the state variables are continuous. The set of basis functions ϕ_L is selected according to (5). In particular, we set Φ to the random Fourier features following the Random Kitchen Sinks (RKS) algorithm (Rahimi and Recht, 2007); we use `RBFsampler` function from the Python `scikit-learn` module for implementation. The bandwidth in the radial basis function (RBF) kernel is selected according to the median heuristic (Garreau et al., 2017). The number of basis functions in Φ (denoted by $M = L/m$) is selected via 5-fold cross-validation. Specifically, for each M, let Φ_M denote the resulting set of basis functions. We first divide all data trajectories into 5 non-overlapping data subsets with equal sizes. Let $\mathcal{I}_\ell$ denote the set of these subsamples, and $\mathcal{I}_\ell^c$ denote its complement, $\ell = 1, 2, 3, 4, 5$. For each combination of ℓ and M, we use FQI to compute an estimated optimal Q-function $\widehat{Q}_{\ell,M}$ by setting $\Phi = \Phi_M$, based on the data subsets in $\mathcal{I}_\ell^c$. We next select M to minimise the FQI objective function,

$$\sum_{\ell=1}^{5} \sum_{(i,t)\in\mathcal{I}_\ell} \left\{ R_{i,t} + \gamma \max_a \widehat{Q}_{\ell,M}(a, S_{i,t+1}) - \widehat{Q}_{\ell,M}(A_{i,t}, S_{i,t}) \right\}^2. \tag{20}$$

To mitigate the randomness introduced by the random Fourier features, we repeat each test procedure four times with different random seeds. This yields p-values $\{p_r, r = 1, \ldots, 4\}$, for each candidate change point. We then employ the method developed by Meinshausen et al. (2009) to combine these p-values by defining

$$p_0 = \min\left(1, q_\tau \left\{\tau^{-1} p_r, r = 1, \ldots, 4\right\}\right), \tag{21}$$

to be the final p-value. Here, τ is some constant between 0 and 1, and q_τ is the empirical τ-quantile. Compared to using a single set of Fourier features, such an aggregation method reduces the type-I error and increases the power of the resulting test. In our simulations, results hardly change under $\tau = 0.05, 0.1, 0.15, 0.2$; hereafter, we report results under $\tau = 0.1$.

	State transition function	Reward function
(1)	Time-homogeneous	Piecewise constant
(2)	Time-homogeneous	Smooth
(3)	Piecewise constant	Time-homogeneous
(4)	Smooth	Time-homogeneous

Table 1: Simulation scenarios with different types of nonstationarity in Section 6.2.

6.2 Simulation I

We consider four nonstationary data generating mechanisms with one-dimensional states and binary actions where the nonstationarity occurs in either the state transition function or the reward function, as listed in Table 1. Specifically, in the first two scenarios, the transition function $\mathcal{P}_t$ is stationary whereas the reward function r_t varies over time. The last two scenarios concern stationary reward functions and nonstationary transition functions. For nonstationary setups, abrupt piece-wise constant and smooth changes are considered. See Appendix C.1 for more details about the true parameters of the transition and reward functions in these four scenarios.

In all settings, we set T =100 and simulate data with sample sizes $N = 25, 100$. The discount factor $\gamma = 0.9, 0.95$. The true location of the change point T^* is set to 50. We first apply the each of the proposed tests to the time interval $[T - \kappa, T]$ to detect nonstationarity, where κ takes value from a equally-spaced sequence between 25 and 75 with increments of 5. According to our true data generating mechanisms, when $\kappa \leq 50$, the null of no change point over $[T - \kappa, T]$ is true; the alternative hypothesis is true if $\kappa > 50$. We fix the significance level α to 0.05. The initial state is sampled from a normal random variable with mean zero and variance 0.5. The actions are generated i.i.d. according to a Bernoulli random variable with a success probability of 0.5.

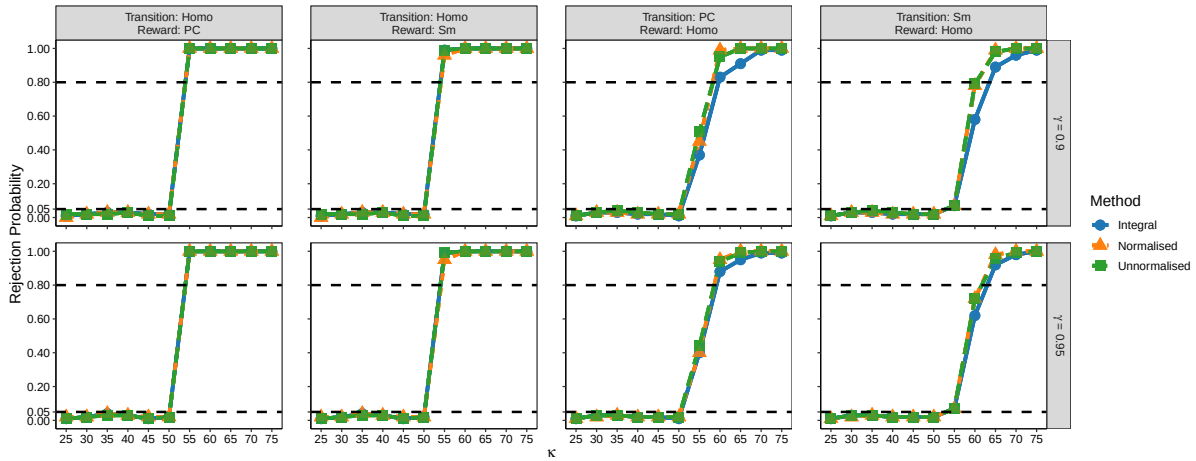
Figure 3 shows the empirical rejection probabilities of each proposed test. We summarise our findings here. First, in all settings, each test can properly control the type I error. Second, the power increases with κ as a result of more pre-change-point data being included into the interval $[T - \kappa, T]$. It also increases with N, demonstrating the consistency of our tests. Third,

as expected, gradual changes are more difficult to detect than abrupt changes. Specifically, it can be seen from Figure 3 that when $\kappa = 55$, the power of the proposed test with a smooth reward or state transition function is smaller than that with a piecewise constant reward or state transition function. Finally, the normalised and unnormalised type test statistics achieve slightly higher power than the integral type test statistic when $N = 25$, whereas the powers of the three tests become comparable when $N = 100$. However, the normalised and unnormalised type test statistics are more computationally expensive especially when the dimension of the state is high, since both require to search the maximum over the entire state-action space.

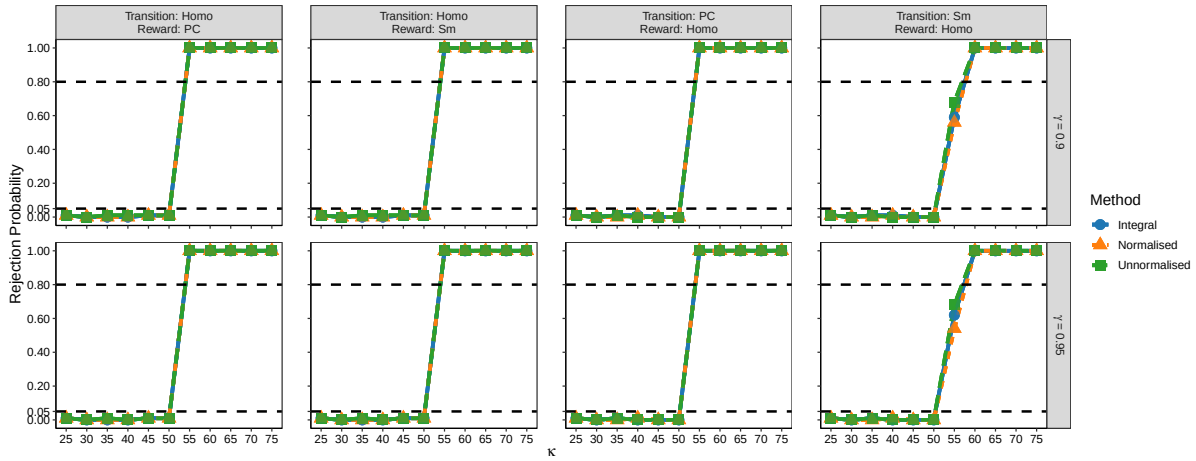
Next, we investigate the finite sample performance of the estimated change point location $\widehat{T}^* = T - \kappa_{j_0-1}$. Figure 4 depicts the histogram of $\widehat{T}^*$ in each of the simulation scenarios. It can be seen that in the first two scenarios with abrupt changes, the estimated change points concentrate at 50, which is the true change point location. In the last two scenarios with smooth changes, the estimated change points have a wider spread when $N = 25$, but are still close to 50 in most cases.

Finally, we compute the optimal policy based on the estimated change point and compare it with some baseline methods. In each simulation, after computing $\widehat{T}^*$, we estimate the optimal mean policy using the data subset $\{(S_{i,t}, R_{i,t}, A_{i,t}) : 1 \leq i \leq N, \widehat{T}^* \leq t \leq T\}$. Specifically, we adopt a decision tree model to approximate Q^{opt} to obtain interpretable policies for healthcare researchers. We couple FQI with decision tree regression to compute the Q-estimator $\widehat{Q}$. The decision tree model involves some hyperparameters such as the maximum tree depth and the minimum number of samples on each leaf node. We use 5-fold cross validation to select these hyperparameters from $\{3, 5, 6\}$ and $\{50, 60, 80\}$, respectively. See the cross-validation criterion in (20). After computing the estimated policy, we simulate 300 new subjects following such a policy for 100 time points after T^* and aggregate the discounted rewards over these subjects to estimate the expected return (e.g., value) under that policy. The proposed three tests yield similar policies. We report the results based on the integral-type test only and compare them with the following baseline methods:

Overall: Standard policy optimisation method that uses all the data;



(a) $N = 25$.



(b) $N = 100$.

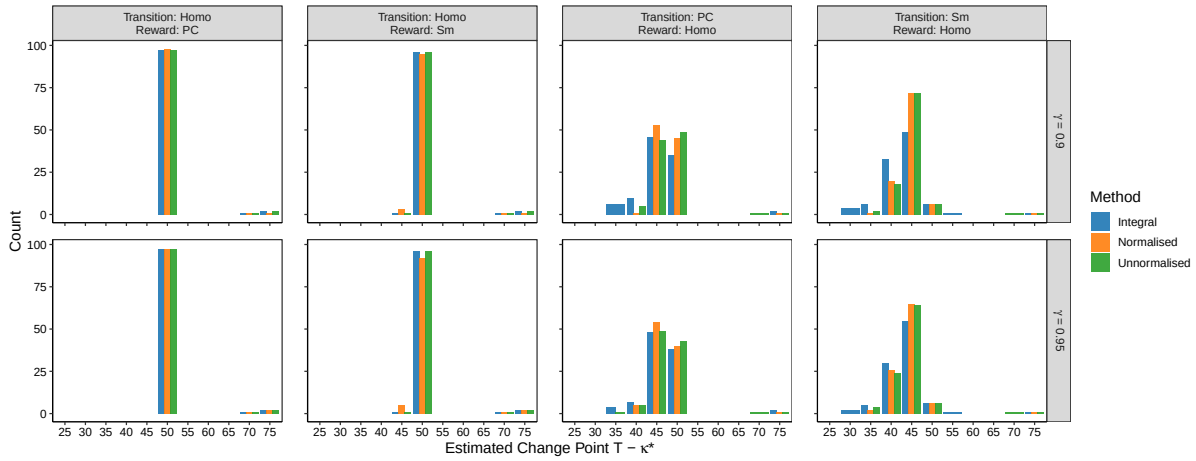
Figure 3: Empirical type I errors and powers of the proposed test and their associated 95% confidence intervals under settings described in Section 6.2. Abbreviations: Hm for homogeneous, PC for piecewise constant, and Sm for smooth.

Random: Policy optimisation with a randomly assigned change point location;

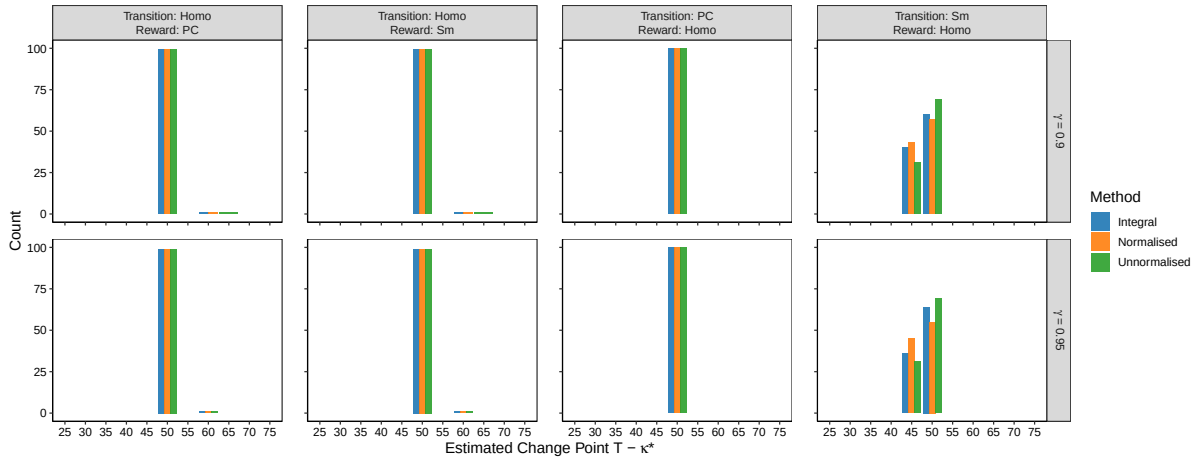
Kernel: The kernel-based approach developed by Domingues et al. (2021);

Oracle: The "oracle" policy optimisation method as well as if the oracle change point location were known in advance.

For fair comparisons, we use FQI and decision tree regression to compute the optimal Q-function for each competing method. To implement the random method, we randomly pick a



(a) $N = 25$.



(b) $N = 100$.

Figure 4: Distribution of detected change points under simulation settings in Section 6.2.

time point $\widetilde{T}^*$ uniformly from the interval $[0, T]$ and compute the optimal Q-function based on the observations that occur after time $\widetilde{T}^*$. We repeat the procedure for 20 times and take the average values of the twenty policies as the values of the random method. To implement the kernel-based method, at the kth FQI iteration, we consider the following objective function,

$$Q^{(k+1)} = \arg\min_{Q} \sum_{i,t} K\left(\frac{T-t}{Th}\right)\left\{R_{i,t} + \gamma \max_{a} Q^{(k)}(a, S_{i,t+1}) - Q(A_{i,t}, S_{i,t})\right\}^2, \quad (22)$$

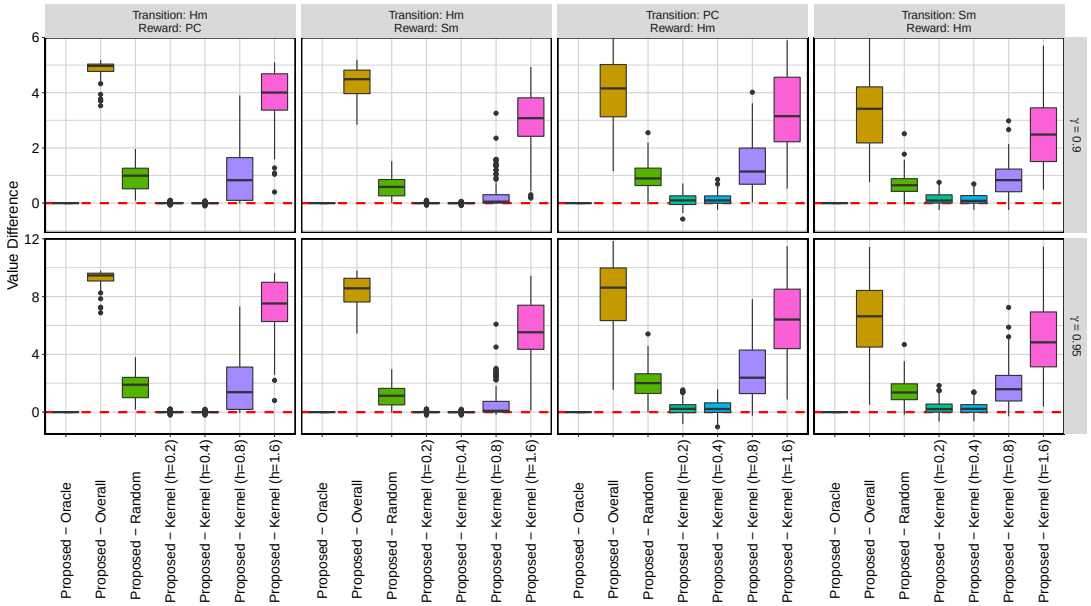
where K denotes the Gaussian RBF basis and h denotes the associated bandwidth parameter taken from the set $\{0.2, 0.4, 0.8, 1.6\}$. According to (22), the kernel-based method assigns

larger weights to more recent observations to deal with nonstationarity. To solve (22), we sample $B \gg T$ data slices across all individuals from $\{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1}; 1 \leq i \leq N)\}_{0 \leq t < T}$ with weights proportional to $K((T-t)/(Th))$ and apply the decision tree regression to these samples to compute $Q^{(k+1)}$. To implement the oracle method, we use observations that occur after the oracle change point to compute the optimal Q-function.

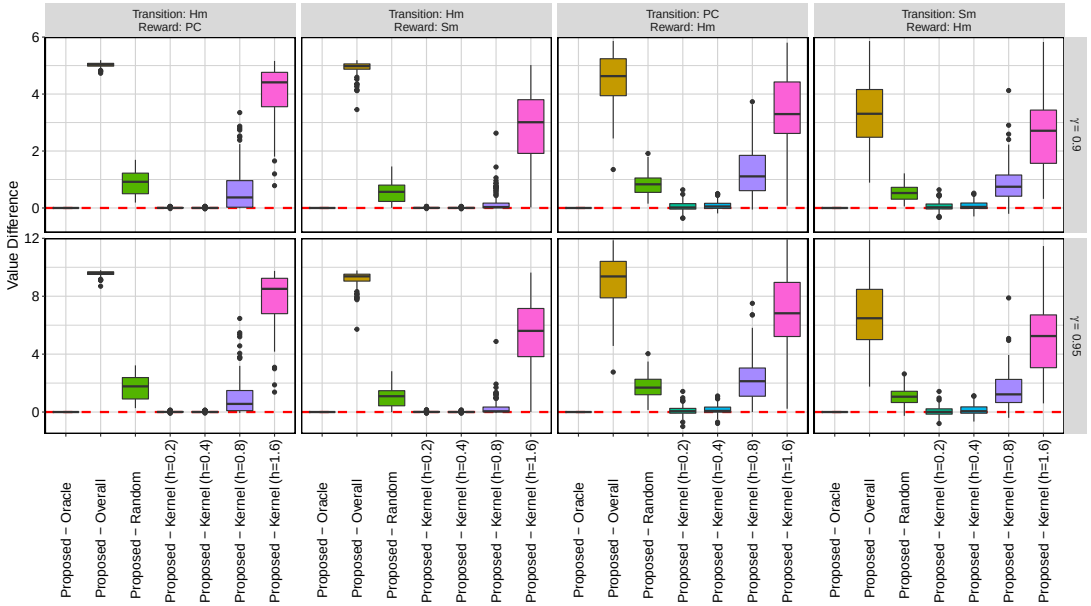
Figure 5 reports the difference between the proposed policy's value and values of policies estimated based on these baseline methods with $N = 25$ and 100. We briefly summarise a few notable findings. First, the proposed method achieves much larger policy values compared to the "overall" method, demonstrating detrimental consequences of ignoring the nonstationarity. Second, the proposed method is comparable to the oracle method, and outperforms the "random" method in all cases. This implies that correctly identifying the change point location is essential to policy optimisation in nonstationary environment. Third, the proposed method is no worse (when $h = 0.2$ or 0.4) and better than (when $h = 0.8$ or 1.6) kernel-based approaches in our cases. As shown in Figure 5, kernel-based method is sensitive to the choice of the kernel bandwidth. A poor choice of h would yield a poor policy, and it remains unclear how to determine this tuning parameter in practice.

6.3 Simulation II

To mimic the IHS study, we simulate $N = 100$ subjects, each observed over $T = 50$ time points. Our aim is to estimate an optimal treatment policy to improve these interns' long-term physical activity levels. See Section 7 for more details about the study background. At time t, the state vector $S_{i,t}$ comprises four variables to mimic the actual IHS study: the square root of step count at time t ($S_{i,t,1}$), cubic root of sleep minutes at time t ($S_{i,t,2}$), mood score at time t ($S_{i,t,3}$), and the square root of step count at time $t-1$ ($S_{i,t,4} = S_{i,t-1,1}$), i.e., the state transition is designed to follow an AR(2) process. See Supplementary C.2 for the true parameter values that govern the dynamics. The actions are binary with $P(A_{it} = 1) = 0.25$; $A_{it} = 1$ means the subject is randomised to receive activity messages at time t, and $A_{it} = 0$ means any other types of messages or no message at all. Reward $R_{i,t} = S_{i,t,1}$ is defined as the step count at time



(a) $N = 25$.



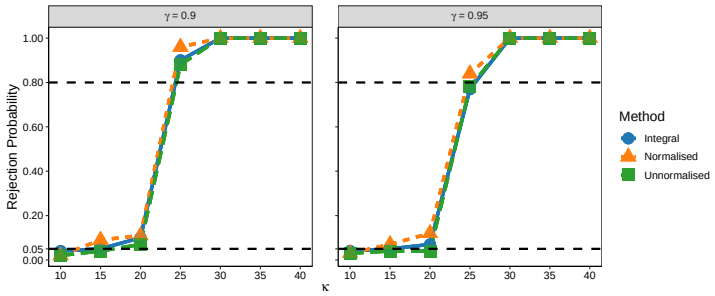
(b) $N = 100$.

Figure 5: Distribution of the difference between the expected return under the proposed policy and those under policies computed by other baseline methods, under settings in Section 6.2. The proposed policy is based on the change point detected by the integral type test statistic. In all scenarios, we find the value results based on the normalised or unnormalised test statistics are similar to those of the integral test statistic.

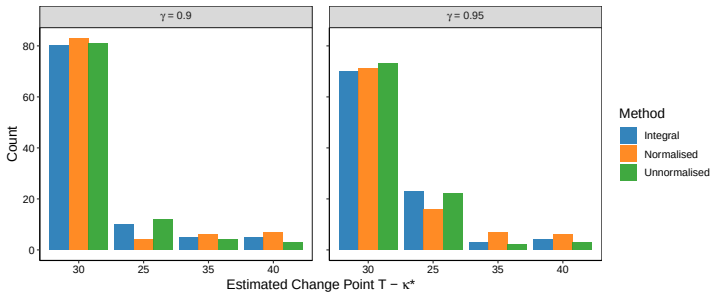
t. We assume that the state transition function has a change point at time $T^* = 25$. Under this setting, the state transition function is nonstationary whereas the reward is a stationary function of the state. In addition, the data follow the null hypothesis when $\kappa = 1, \ldots, 25$ and follow the alternative hypothesis for $\kappa = 26, \ldots, 49$. The discount factor is set to $\gamma = 0.9$ or 0.95. We test the null hypothesis along a sequence of $\kappa = 10, 15, \ldots, 40$ for every five time points. The number of basis M is chosen among $\{10, 15, 20, 25, 30\}$.

Figure 6 shows the empirical rejection rates of the proposed tests as well as the distribution of the estimated change point location. Similar to the results in Section 6.2, our proposed tests control the type I error at the nominal level (see $\kappa < 25$) and is powerful to detect the alternative hypothesis (see $\kappa > 25$). At the boundary where $\kappa = 25$ however, the proposed test fails to control the type-I error. Nonetheless, the proposed procedure yields a much better policy compared to the overall and random methods, as we show below. We also remark that the reason the proposed test fails at the boundary is because the marginal distribution of the first few states after the change point is very different from the stationary state distribution. After an initial burn-in period of 5 points, the proposed test is able to control the type-I error at $\kappa = 20$. In addition, the distribution of the estimated change point concentrates on 30, which is very close to the oracle change point location 25, implying the consistency of the proposed change point detection procedure. We remark that consistency here requires $T^{-1}|\widehat{T}^* - T^*| \xrightarrow{P} 0$ instead of $\mathbb{P}(\widehat{T}^* = T^*) \to 1$, the latter being usually impossible to achieve in change-point settings.

We also compare the proposed policy (based on the integral test) with the policies identified by the oracle, overall and random method. Table 2 reports the value difference between the proposed method and the aforementioned methods. It can be seen that the proposed method produces higher values than methods that do not properly address nonstationarity and is comparable to values produced with an oracle change point location. These results again highlight the necessity to identify a change point before applying Q-leaning.



(a) Type I error ($\kappa \leq 25$) and power ($\kappa > 25$) .



(b) Estimated change point $\widehat{T}^* = T - \kappa_{j_0-1}$ where the first rejection based on isonotic regression fit occurred at κ_{j_0}.

Figure 6: Simulation II: Empirical rejection rates of the proposed tests (integeral, normalised, and unnormalised) and the distribution of the estimated change points.

7 Application to Intern Health Study

The 2018 Intern Health Study (IHS) is a micro-randomised trial (MRT) that seeks to evaluate the efficacy of different push notifications sent via a customised study App upon proximal physical and mental health outcomes (NeCamp et al., 2020), a critical first step for designing effective just-in-time adaptive interventions. Over the $T = 26$ study period, each study subject was re-randomised weekly to receive activity suggestions or not; daily self-reported mood scores were assessed via ecological momentary assessments; step count and sleep duration in minutes were measured by wearables (Fitbit). In this paper, we focus on policy optimisation for improving time-discounted cumulative step counts under the infinite horizon setting. However, as have been shown by previous studies (Klasnja et al., 2019; Qian et al., 2022), the longer a person is under intervention, the more they may habituate to the prompts or become

Difference (in value)	Mean value (s.e) γ =0.9	γ =0.95
Proposed - Oracle	-0.06 (0.04)	-0.03 (0.03)
Proposed - Overall	22.92 (1.04)	35.92 (2.42)
Proposed - Random	5.75 (0.86)	9.62 (1.49)

Table 2: Simulation II: Value difference between the proposed method and the oracle, overall and random method. Positive numbers indicate higher values based on the proposed method.

overburdened, resulting in subjects being less responsive to the contents of the suggestions. The treatment effect of activity suggestions may transition from positive to negative, suggesting treatment policies may benefit from adaptation over time. Failure to recognise potential nonstationarity in treatment effects over time may lead to suboptimal policies that overburden subjects, resulting in app deletion and study dropouts. Here we demonstrate how to use the proposed method to detect change point and perform optimal policy estimation in the presence of potential temporal nonstationarity.

7.1 Data and method: Setup

Let the state vector S_t be comprised of the following: square root of average step count in week t, cubic root of average sleep minute in week t, average mood score in week t, and square root of average step count in week $t-1$; all state variables are normalised after respective transformations (NeCamp et al., 2020). The reward R_t is defined as the average step count in week t. The binary action $A_t = 1$ (0) corresponds to pushing (not pushing) an activity message in week t. The randomisation probabilities are known under MRT: $\mathbb{P}(A_t = 1) = 1 - \mathbb{P}(A_t = 0) = 1/4$. In the change point detection procedure, we set $\epsilon = 0.08$ and search for change points for $t \in [5, 22]$. The number of RBF basis functions $M \in \{3, 5, 8, 10\}$ is selected through 5-fold cross validation (see Section 6.1 for implementation details). We focus on three specialties: emergency ($N = 141$), pediatrics ($N = 211$), and family practice ($N = 125$). One consideration is that work schedules and activity levels vary greatly across different specialties, and thus medical interns might experience distinct change points. Stratification by specialty may

improve homogeneity of the study groups so that the assumption of a common change point is more plausible; see Section 8.4 for discussions on potential extensions to heterogeneous change points.

7.2 Results

Figure 7 shows the trajectories of p-values and the isotonic regression results using integral type test statistic; the results are similar when normalised and unnormalised tests were applied to the data (not reported here). We consider $\gamma = 0.9$ or 0.95, which produce similar results. Notice that when κ is small, many p-values are close to 1. This is due to the use of the aggregation method in (21), which tends to increase insignificant p-values and reduce the type-I error. The emergency specialty displays roughly monotonic p-values over time, indicating a single change point at $\kappa = 12$ for $\gamma = 0.9$, and $\kappa = 11$ for $\gamma = 0.95$. On the other hand, the U-shaped p-value trajectory of the pediatrics specialty shows evidence for multiple change points, one at $\kappa = 9$ and another one around $\kappa = 15$. As we have discussed earlier, when only a single change point exists, the significant p-values are likely to decrease with κ. The U-shaped p-value trajectory can occur only when the data interval contains at least two change points and the system dynamics after the second change point is similar to that before the first change occurs, yielding a small CUSUM statistics. Because we focus on the latest detected change point (first κ_{j_0-1} where κ_{j_0} results in a rejection of the null) to inform the latest data segment to use for optimal policy estimation, we perform isotonic regression on p-values up to $\kappa = 15$, which yields $\kappa_{j_0-1} = 9$. In addition, the p-value trajectories of the family practice specialty are mostly flat and are close to 1, indicating the stationarity assumption is compatible with this data subset, for which estimate the optimal policy using data from all time points.

We next compare the proposed policy optimisation method with two other methods: 1) overall, which was described in Section 6.2) and 2) behavior, which is the treatment policy used in the completed MRT. We first split all the subjects into training and evaluation data sets with a ratio of 3/2. The training data are used to learn an optimal policy $\widehat{\pi}^{opt}$ through FQI and decision tree regression, based on the estimated change point location. Hyperparameters of

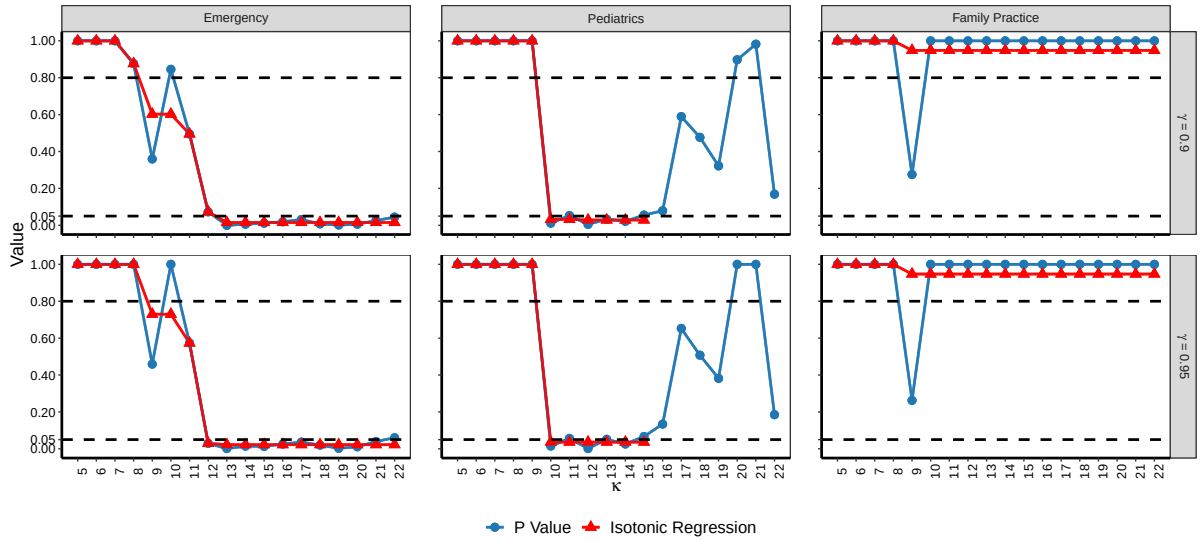


Figure 7: P-value over different values of κ (the number of time points from the last time point T). $\gamma = 0.9$ and 0.95, from top plots to bottom plots. The specialties corresponds to emergency, pediatrics and family practice, from left plots to right plots.

the decision tree regression are selected via 5-fold cross-validation, similar to the simulation sections. Next, based on the evaluation data, we employed fitted-Q evaluation (FQE, Le et al., 2019), which is designed for off-policy value evaluation. We normalise the value estimates by $1-\gamma$ to reflect value per time point, as shown in Table 3. In the emergency specialty, the optimal policy estimated using data after the detected change point improves weekly average step count per day by about $170 \sim 200$ steps relative to the estimated policy using data from all time points ("Overall"). The proposed and overall methods have equal values in the family practice specialty because no change point is identified. These results show that policy improvement can be achieved by identifying a change point and applying standard reinforcement learning to data after the detected change point that is compatible with the stationarity assumption.

# Change points	Specialty	Method	Value	
			$\gamma = 0.9$	$\gamma = 0.95$
1	Emergency	Proposed	8073.27	8003.38
		Overall	7902.39	7794.77
		Behavior	7823.75	7777.32
≥ 2	Pediatrics	Proposed	7783.86	7762.81
		Overall	7680.04	7686.46
		Behavior	7730.98	7721.29
0	Family Practice	Proposed	8087.15	8072.78
		Overall	8087.15	8072.78
		Behavior	7967.67	7957.24

Table 3: Mean value estimates using decision tree in anaylsis of IHS. Values are normalised by multiplying $1 - \gamma$. All values are evaluated over 10 splits of data.

8 Discussion

8.1 Online Change Point Detection

We consider testing nonstationarity and policy optimisation based on a pre-collected offline dataset. Meanwhile, the proposed methodology can be implemented online as data accumulate. In practice, we may choose to update the policy and the change point in batches rather than at every time point. Then we can repeatedly apply the proposed offline method to moderately large batches of observations for change point detection and policy optimisation. We also notice that there are some recent works on online nonstationary RL in the computer science literature (see e.g., Lecarpentier and Rachelson, 2019; Cheung et al., 2020; Padakandla et al., 2020; Fei et al., 2020; Xie et al., 2021; Zhong et al., 2021). In particular, Padakandla et al. (2020) proposed to apply the online change point detection algorithm developed by KJ et al. (2022) to the state-reward-next state triples to identify changes in the environment. When the behavior policy varies over time, the marginal distribution of the triplet can be nonstationary despite the state transition and reward functions being stationary. As such, their method would not apply in this case.

8.2 Alternative Change Point Detection Methods

In this paper, we focus on detecting change points of the optimal Q-function. Alternatively, one could directly detect changes in the reward and transition functions. The estimated reward and transition functions are more easily computable than the estimated optimal Q-function. However, changes in the reward and transition functions do not necessarily cause changes in the optimal Q-function. In other words, it is possible that the reward and transition functions are nonstationary over time whereas the optimal Q-function is stationary. In addition, notice that the transition function is a multi-output function. It remains challenging to detect its change points in high-dimensional settings. We leave it for future research.

Moreover, our procedure first estimates the optimal Q-function based on fitted Q-iteration and then constructs test statistics based on the Q-estimator for change point detection. Alternatively, one could develop a hybrid procedure that couples fitted Q-iteration with change point detection. That is, at each Q-iteration, we construct CUSUM-type statistics based on the estimated Q-function for change point detection. If not detected, then we proceed with the next Q-iteration. It would be interesting to investigate the performance of the resulting algorithm. However, this is beyond the scope of the current paper. We leave it for future research.

8.3 Nonregular Settings

As discussed in Section 5.1, the proposed test requires the optimal policy to be uniquely defined. Such an assumption essentially requires that $Q^{opt}(a_1, S_t) - Q^{opt}(a_2, S_t)$ is nonzero almost surely for any $a_1 \neq a_2$ and any t. It is violated when some treatment is neither beneficial nor harmful for a subset of patients in the population compared to the standard control (see e.g. Luedtke and van der Laan, 2016). In that case, we can couple the proposed test with data splitting to ensure its validity. Specifically, we first divide all trajectories into two disjoint subsets. We next learn the optimal policy using each data subset, evaluate its Q-value on the other dataset based on linear sieves (see e.g. Shi et al., 2021) and construct a CUSUM-type test statistic based on these estimated Q-values. This yields two test statistics, for each of the data subsets. Finally, we use bootstrap to compute the p-value for each test statistic and combine

them based on the Bonferroni correction.

8.4 Heterogeneous Change Points

The proposed method relies on a common change point assumption. That is, all the subjects share the same change point location. This assumption can be violated in practice, due to the subject heterogeneity. In our data application, we stratify the medical interns by their specialties. This helps improve homogeneity of the study groups to some extent. However, interns within the same specialty may also have different change points. Detecting the change point using each individual intern's trajectory is impossible, due to data scarcity. It remains unclear how to efficiently detect the change points without the homogeneity assumption. We leave it for future research.

Acknowledgement

This work is partly supported by the National Institute of Mental Health (R01 MH101459) (to Z.W.), an investigator grant from Precision Health Initiative at University of Michigan, Ann Arbor (to Z.W. and M.L.), and by EPSRC grants EP/V053639/1 (to P.F.), EP/W014971/1 (to C.S.). We thank the interns and residency programs who took part in this study and the study PI (Dr. Srijan Sen) for data access.

References

Alquier, P., Doukhan, P., and Fan, X. (2019). Exponential inequalities for nonstationary Markov chains. *Dependence Modeling*, 7(1):150–168.

Aminikhanghahi, S. and Cook, D. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51:339–367.

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., and Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet*, 395(10228):931–934.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

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

Brunk, H., Barlow, R. E., Bartholomew, D. J., and Bremner, J. M. (1972). Statistical inference under order restrictions: the theory and application of isotonic regression. Technical report, Missouri Univ Columbia Dept of Statistics.

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

Cazelles, B., Champagne, C., and Dureau, J. (2018). Accounting for non-stationarity in epidemiology by embedding time-varying parameters in stochastic models. *PLoS Computational Biology*, 14(8):e1006211.

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723.

Chakraborty, B., Murphy, S., and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3):317–343.

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

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632.

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

Chen, X. and Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2020). Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, pages 1843–1854. PMLR.

Cho, H. and Fryzlewicz, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, pages 207–229.

Cho, H. and Fryzlewicz, P. (2015). Multiple change-point detection for high-dimensional time series via Sparsified Binary Segmentation. *Journal of the Royal Statistical Society: Series B*, 77:475–507.

Csörgö, M., Csörgö, M., and Horváth, L. (1997). *Limit theorems in change-point analysis*. John Wiley & Sons.

de Haan, P., Jayaraman, D., and Levine, S. (2019). Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 11698–11709.

Dedecker, J. and Fan, X. (2015). Deviation inequalities for separately Lipschitz functionals of iterated random functions. *Stochastic Processes and their Applications*, 125(1):60–90.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*, volume 303. Springer Science & Business Media.

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. (2021). A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 3538–3546. PMLR.

Eftekhari, H., Mukherjee, D., Banerjee, M., and Ritov, Y. (2020). Markovian and non-Markovian processes with active decision making strategies for addressing the COVID-19 pandemic. *arXiv preprint arXiv:2008.00375*.

Eichenbaum, M. S., Rebelo, S., and Trabandt, M. (2020). The macroeconomics of epidemics. Technical report, National Bureau of Economic Research.

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

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

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR.

Fang, E. X., Wang, Z., and Wang, L. (2022). Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, just-accepted:1–14.

Fei, Y., Yang, Z., Wang, Z., and Xie, Q. (2020). Dynamic regret of policy optimization in non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 6743–6754.

Fryzlewicz, P. (2014). Wild Binary Segmentation for multiple change-point detection. *The Annals of Statistics*, 42:2243–2281.

Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*.

Hasselt, H. V. (2010). Double Q-learning. In *Advances in Neural Information Processing Systems*, pages 2613–2621.

Hu, X., Qian, M., Cheng, B., and Cheung, Y. K. (2021a). Personalized policy learning using longitudinal mobile health data. *Journal of the American Statistical Association*, 116(533):410–420.

Hu, Y., Kallus, N., and Uehara, M. (2021b). Fast rates for the regret of offline reinforcement learning. *arXiv preprint arXiv:2102.00479*.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, 26(1):242–272.

Jin, J., Song, C., Li, H., Gai, K., Wang, J., and Zhang, W. (2018). Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proc. CIKM*, pages 2193–2201.

Judd, K. L. (1998). *Numerical methods in economics*. MIT press.

Killick, R., Fearnhead, P., and Eckley, I. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

KJ, P., Singh, N., Dayama, P., Agarwal, A., and Pandit, V. (2022). Change point detection for compositional multivariate data. *Applied Intelligence*, 52(2):1930–1955.

Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B., and Murphy, S. A. (2019). Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of Heartsteps. *Annals of Behavioral Medicine*, 53(6):573–582.

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

Kompella, V., Capobianco, R., Jong, S., Browne, J., Fox, S., Meyers, L., Wurman, P., and Stone, P. (2020). Reinforcement learning for optimization of COVID-19 mitigation policies. *arXiv preprint arXiv:2010.10560*.

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

Kosorok, M. R. and Laber, E. B. (2019). Precision medicine. *Annual review of statistics and its application*, 6:263–286.

Le, H., Voloshin, C., and Yue, Y. (2019). Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712.

Lecarpentier, E. and Rachelson, E. (2019). Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 32.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. (2020a). Personalized Heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22.

Liao, P., Klasnja, P., and Murphy, S. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391.

Liao, P., Qi, Z., and Murphy, S. (2020b). Batch policy learning in average reward Markov decision processes. *arXiv preprint arXiv:2007.11771*.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020). Estimating dynamic treatment regimes in mobile health using V-learning. *Journal of the American Statistical Association*, 115(530):692–706.

Luedtke, A. R. and van der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44(2):713–742.

Maei, H. R., Szepesvári, C., Bhatnagar, S., and Sutton, R. S. (2010). Toward off-policy learning control with function approximation. In *ICML*, pages 719–726.

Mahase, E. (2021). COVID-19: Pfizer vaccine's efficacy declined from 96% to 84% four months after second dose, company reports.

Marling, C. and Bunescu, R. (2020). The OhioT1DM dataset for blood glucose level prediction: Update 2020. In *CEUR Workshop Proceedings*, volume 2675, page 71.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probability*, 2:620–628.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

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. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Mukerjee, H. (1988). Monotone nonparametric regression. *The Annals of Statistics*, pages 741–750.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5).

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 65(2):331–366.

NeCamp, T., Sen, S., Frank, E., Walton, M. A., Ionides, E. L., Fang, Y., Tewari, A., and Wu, Z. (2020). Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: Micro-randomized trial. *Journal of Medical Internet Research*, 22(3):e15033.

Nie, X., Brunskill, E., and Wager, S. (2021). Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409.

Niroui, F., Zhang, K., Kashino, Z., and Nejat, G. (2019). Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments. *IEEE Robotics and Automation Letters*, 4(2):610–617.

Padakandla, S., Prabuchandran, K., and Bhatnagar, S. (2020). Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, 50(11):3590–3606.

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

Qi, Z., Liu, D., Fu, H., and Liu, Y. (2020). Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. *Journal of the American Statistical Association*, 115(530):678–691.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210.

Qian, T., Walton, A. E., Collins, L. M., Klasnja, P., Lanza, S. T., Nahum-Shani, I., Rabbi, M., Russell, M. A., Walton, M. A., Yoo, H., et al. (2022). The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations. *Psychological Methods*.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Ramprasad, P., Li, Y., Yang, Z., Wang, Z., Sun, W. W., and Cheng, G. (2021). Online bootstrap inference for policy evaluation in reinforcement learning. *arXiv preprint arXiv:2108.03706*.

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

Shi, C., Lu, W., and Song, R. (2020a). Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *Journal of Machine Learning Research*, 21(176):1–67.

Shi, C., Song, R., Lu, W., and Fu, B. (2018). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of Royal Statistical Society: Series B*, 80(4):681–702.

Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. (2020b). Does the Markov decision process fit the data: Testing for the Markov property in sequential decision making. In *International Conference on Machine Learning*, pages 8807–8817. PMLR.

Shi, C., Zhang, S., Lu, W., and Song, R. (2021). Statistical inference of the value function for reinforcement learning in infinite horizon settings. *Journal of Royal Statistical Society: Series B*, accepted.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Song, R., Wang, W., Zeng, D., and Kosorok, M. R. (2015). Penalized Q-learning for dynamic treatment regimens. *Statistica Sinica*, 25(3):901–920.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition.

Sutton, R. S., Szepesvári, C., and Maei, H. R. (2008). A convergent O(n) algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in Neural Information Processing Systems*, 21(21):1609–1616.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167. 107299.

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2019). *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC press.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. (2021). Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York.

Wallace, M. P. and Moodie, E. E. M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, 71(3):636–644.

Wan, R., Zhang, X., and Song, R. (2020). Multi-objective reinforcement learning for infectious disease control with application to covid-19 spread. *arXiv preprint arXiv:2009.04607*.

Wang, L., Zhou, Y., Song, R., and Sherwood, B. (2018). Quantile-optimal treatment regimes. *J. Amer. Stat. Assoc.*, 113(523):1243–1254.

Wang, T. and Samworth, R. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B*, 80:57–83.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Xie, A., Harrison, J., and Finn, C. (2021). Deep reinforcement learning amidst continual structured non-stationarity. In *International Conference on Machine Learning*, pages 11393–11403. PMLR.

Xu, Z., Li, Z., Guan, Q., Zhang, D., Li, Q., Nan, J., Liu, C., Bian, W., and Ye, J. (2018). Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proc. ACM KDD*, pages 905–913.

Yu, M. and Chen, X. (2021). Finite sample change point inference and identification for high-dimensional mean vectors. *Journal of the Royal Statistical Society: Series B*, 83(2):247–270.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694.

Zhang, Y., Laber, E. B., Davidian, M., and Tsiatis, A. A. (2018). Estimation of optimal treatment regimes using lists. *Journal of American Statistical Association*, 113(524):1541–1549.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of American Statistical Association*, 110(510):583–598.

Zhong, H., Yang, Z., and Szepesvári, Z. W. C. (2021). Optimistic policy optimization is provably efficient in non-stationary MDPs. *arXiv preprint arXiv:2110.08984*.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S., and Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, 73(2):391–400.