

Reinforcement Learning under External Influence: Guarantees, Algorithms, and Sample Complexity

Ranga Shaarad Ayyagari

RANGAA@IISC.AC.IN

Revanth Raj Eega

REVANTHRAJE@IISC.AC.IN

Ambedkar Dukkipati

AMBEDKAR@IISC.AC.IN

*Department of Computer Science and Automation
Indian Institute of Science
Bengaluru, India*

Abstract

In this paper, we study the problem of reinforcement learning under the influence of external events. For this, we consider Markov decision processes with continuous state and action spaces whose transition dynamics are perturbed by an external process in a non-Markovian manner. First, we establish the conditions under which the problem becomes tractable, allowing it to be addressed by considering only a finite history of events, based on the properties of the perturbations introduced by the exogenous process. We propose and theoretically analyze a policy iteration algorithm to tackle this problem that learns policies contingent on the current state of the environment and a finite history of prior exogenous events. Since this algorithm is not guaranteed to converge, we provide a guarantee for policy improvement in regions of the state space determined by the approximation error induced by considering tractable policies and value functions. We also establish the sample complexity of least-squares policy evaluation and policy improvement algorithms that account for approximations arising from considering only a finite history of temporal events. While our results are applicable to general discrete-time processes satisfying certain conditions on the rate of decay of the influence of their events, we further analyze the case of discrete-time Hawkes processes with Gaussian marks. We performed experiments to demonstrate our findings for policy evaluation and deployment in traditional control environments.

Keywords: Markov Decision Processes, Temporal Point Processes, Policy Iteration

1 Introduction

In the mathematical framework of Markov Decision Processes (MDP), which forms the core of reinforcement learning, an agent interacts with an ‘environment’, and at each time step, the agent is required to make a decision and take an action. The state of the environment changes stochastically according to the Markov property, depending solely on the current state and the action taken by the agent. When solving various sequential decision-making problems, most RL algorithms assume that although the state of an MDP may be constantly changing, the rules governing state transitions remain unchanged. However, in practical applications, external events may influence the agent’s environment and change its dynamics.

Consider portfolio management in finance, where an agent has access to a state comprising various fundamental and technical indicators for a select set of financial instru-



Figure 1: Sustained effect of transient nonstationarity. The plot shows the prices of the two assets following a Geometric Brownian motion. A few exogenous Gaussian shocks to one asset force it to diverge significantly from the other over the long term, with a markedly different pattern of evolution long after the effect of the exogenous event decays.

ments and must make sequential decisions about whether to hold, sell, or short each instrument. However, the price movements of these instruments may be affected by external forces, such as government decisions, such as rate hikes, or sudden market movements caused by events occurring in a completely different sector of the economy that are not being considered by the agent. Such an external event might exert a lasting influence over different time scales, depending on the type of event that occurred. A rate cut might affect stock prices for a few months, whereas a flash crash might cease its effect in a few hours. An example illustrating the persistent nonstationarity effect of perturbations due to exogenous events is shown in Figure 1.

As another example, consider a navigation problem in which an agent has control of a robot and must travel from one point to another. Although the agent can learn to perform this task under ideal conditions, its optimal actions might depend on external factors, such as rain, sudden traffic changes, or human movement patterns. Hence, while an agent can learn to achieve goals in an ideal environment, it cannot be oblivious to external changes that influence environmental dynamics.

In the literature, nonstationarity in RL settings has been studied by considering many special cases, mostly providing practical solutions. Some works model nonstationarity as piece-wise stationarity (Alegre et al., 2021; Hadoux et al., 2014; Li et al., 2025), while a few consider drifting environments (Lecarpentier and Rachelson, 2019; Cheung et al., 2020). Hallak et al. (2015) study Contextual MDPs, in which the environment dynamics change based on externally specified “contexts” that are episodic and possibly unknown, chosen from among a finite set of known values. Tennenholtz et al. (2023) extend this setting to handle dynamic contexts known to the agent that change at each time step and influence the state transition distribution. However, their contexts are not exogenous and originate from a finite set of possible vectors.

In this study, we examine Markov Decision Processes (MDPs) with continuous state and action spaces whose transition dynamics are perturbed by an external process in a non-Markovian manner. This setting is very general, as one would only know that there is nonstationarity without explicit information about how the rewards and the probability transition matrix are affected by it. This compels one to consider the history of events at each time step to make optimal decisions, increasing the computational complexity of any algorithms considered. Furthermore, external influence mandates an additional approximation for which we establish sample complexity results.

Contributions. Our contributions are as follows.

- (1) We outline the necessary conditions that ensure the existence of a well-defined solution for this problem. Then, we establish the criteria under which an approximate solution can be determined using only the current state and finite history of past events. We show that this is possible when the perturbations caused by events older than t time steps on the MDP transition dynamics and the event process itself are bounded in total variation by M_t and N_t respectively, and $\sum_t M_t, \sum_t N_t$ are convergent series. We analyze the trade-off between the agent’s performance and the length of history considered.
- (2) We propose a policy iteration algorithm and theoretically analyze its behavior. We show that policy improvement occurs in regions of the state space where the Bellman error is “not too low” based on M_t and N_t . Based on our analysis, we observe that the higher the approximation error due to nonstationarity, the smaller the region in which the policy is guaranteed to improve.
- (3) We then analyze the sample complexity of pathwise least-squares temporal difference policy evaluation and approximate least-squares policy iteration in this setting. We study the impact of nonstationarity on the expected error of the learned value function (Lazaric et al., 2012) and the expected suboptimality of the resultant policy after K iterations. For policy evaluation, we show that, with high probability, the expected error of the value function estimate decomposes into an approximation error arising from discarding old events, an inherent error due to linear function approximation, and standard error terms due to stochasticity and mixing present in the stationary case. For the overall policy iteration algorithm, we quantify the effect of the additional error that occurs due to the consideration of approximately greedy policies that only depend on the current state and a finite history of events.
- (4) We further explain our theoretical results using the discrete-time Hawkes process and conduct experiments for policy evaluation and policy deployment in nonstationary pendulum and point maze environments to validate our results.

2 Preliminaries and Notation

Consider a Markov Decision Process (MDP) defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, Q, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, Q is the transition probability kernel, r is the reward function, and γ is the discount factor. At each time step t , the agent in state $s \in \mathcal{S}$ takes an action $a \in \mathcal{A}$ obtaining a reward $r(s, a)$, and the environment transitions to state $s' \sim Q(\cdot | s, a)$.

In general, $\mathcal{S}$ and $\mathcal{A}$ are Borel spaces, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is a measurable function, and $\gamma \in (0, 1)$. Q is a stochastic kernel on $\mathcal{S}$ given $\mathcal{S} \times \mathcal{A}$. That is, $Q(\cdot | y)$ is a probability

measure on $\mathcal{S}$ for each $y \in \mathcal{S} \times \mathcal{A}$, and $Q(B|\cdot)$ is a measurable function on $\mathcal{S} \times \mathcal{A}$ for each $B \in \mathcal{B}(\mathcal{S})$, the Borel σ -algebra on $\mathcal{S}$.

Given a class of possible policies Π , the goal of the agent is to determine a policy $\pi \in \Pi$ that achieves the maximum possible value function V^π ,

$$V^\pi(s) = \mathbb{E}_{s_t, a_t}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right], \quad \pi \in \Pi, s \in \mathcal{S},$$

the expected sum of discounted rewards obtained by an agent when it starts from a state s and follows a policy π resulting in a trajectory of states $s_0 = s, s_1, s_2, \dots$ by taking actions $a_0, a_1, a_2, \dots$. The expectation is over the states and actions in the agent's trajectory owing to the stochasticity of the environment and possibly the policy itself.

Let π^* be an optimal policy that achieves a corresponding value function V^{π^*} , denoted V^* . A measurable function $v : \mathcal{S} \rightarrow \mathbb{R}$ is said to be a solution to the Bellman optimality equation if it satisfies $v = \mathcal{T}v$, where $\mathcal{T}$ is the optimal Bellman operator defined as

$$[\mathcal{T}v](s) = \max_{a \in \mathcal{A}} \left[r(s, a) + \gamma \int_{\mathcal{S}} v(s') Q(ds'|s, a) \right], \quad \text{for all } s \in \mathcal{S}. \quad (1)$$

A function that satisfies the optimal Bellman equation is a fixed point of the operator $\mathcal{T}$. Under suitable regularity conditions on the rewards and transition kernel, the optimal value function V^* satisfies (1).

In this paper, we consider the setting in which the agent receives rewards from the environment, defined by the reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, and its goal is to maximize expected return. Equivalently, many studies consider an agent incurring "costs", defined by an analogous cost function $c : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. In this case, the objective of the agent is to minimize the expected discounted sum of the costs incurred; thus, the Bellman operator is defined as the minimization of the right-hand side in (1).

For this cost setting, under the assumptions of (i) the one-stage cost c is lower semi-continuous, non-negative, and inf-compact on $\mathcal{S} \times \mathcal{A}$, (ii) Q is strongly continuous, and (iii) there exists a policy π such that $V^\pi(s) < \infty$ for all $s \in \mathcal{S}$, one can establish the existence and uniqueness of the optimal policy and value function, given by the following result.

Theorem 1 (Theorem 4.2.3 of Hernández-Lerma and Lasserre (1996)). *If the above two assumptions hold, then:*

1. *The optimal value function V^* is the pointwise minimal solution to the above Bellman equation. If $u(\cdot)$ is any other solution to the above equation, then $u(\cdot) \geq V^*(\cdot)$.*
2. *There exists a function $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$ such that $\pi^*(s) \in \mathcal{A}$ attains the minimum in the Bellman equation, i.e.,*

$$V^*(s) = c(s, \pi^*(s)) + \gamma \int_{\mathcal{S}} V^*(s') Q(ds'|s, \pi^*(s)), \quad \text{for all } s \in \mathcal{S},$$

The deterministic stationary policy π^ is optimal. Conversely, any deterministic stationary policy that is optimal satisfies the above equation.*

3. *If an optimal policy exists, then there exists one that is deterministic and stationary.*

3 MDP influenced by a Temporal Process

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, Q, r, \gamma)$ be an MDP with state and action spaces $\mathcal{S}$ and $\mathcal{A}$ respectively, transition kernel Q , discount factor γ , and reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$, which defines the reward $r(s, a, s')$ as a function of the state-action pair (s, a) and the next state s' . Consider an external discrete-time temporal process $X = (t_i, X_i)_{i \in \mathbb{N}}$ that influences the transition probabilities of $\mathcal{M}$. Here, X is exogenous; hence, it is not affected by $\mathcal{M}$. The probability of an event occurring in the external process at time t is a function of all past (external) events until that time. Associated with each event is a probabilistic mark $x \in \mathcal{X}$, where $\mathcal{X}$ is a closed subset of $\mathbb{R}^d$.

Let $H_t = \{(t_1, x_1), (t_2, x_2), \dots, (t_j, x_j)\}$ be the history of external events until time t , where the ordered pair (t_i, x_i) denotes an event with mark x_i that occurred at time t_i . History H_t perturbs the stochastic kernel Q of the MDP to give rise to a new transition kernel Q_{H_t} on $\mathcal{S}$ given $\mathcal{S} \times \mathcal{A}$. At the same time, this history determines the next event in the temporal process. We denote the distribution of the event mark at time t by $Q_{H_t}^X$, which is a probability distribution on $\mathcal{X}$. Because this describes the event distribution external to the MDP, it does not depend on the state of the MDP or the action taken by the agent.

In this paper, we consider a discrete-time event process as an external influence and make the following assumptions.

(A1) There exists an event with a mark, say $X = 0$ (zero), that is equivalent to a non-event.

(A2) Events that occurred in the distant past have a reduced influence on the current probability transition kernel of the MDP. Specifically, there exists a convergent series $\sum_T M_T$ of real numbers, such that at any time t , for any event histories H_t and H'_t , if the MDP is in state s and action a is taken, and if H_t and H'_t differ by only one event at time t' , then

$$\text{TV}(Q_{H_t}(\cdot|s, a), Q_{H'_t}(\cdot|s, a)) < M_T, \quad \text{for all } s \in \mathcal{S}, a \in \mathcal{A},$$

provided $t - t' \geq T$. That is, if an event is older than T , its influence on the current probability distribution has a Total Variation distance upper bounded by M_T . Furthermore, Q_{H_t} is independent of the previous states conditioned on the current state. That is, Q_{H_t} is Markovian with respect to the state. Furthermore, s_{t+1} and x_{t+1} are conditionally independent given the history (H_t) , current state (s_t) , and action (a_t) .

(A3) Similarly, events that occurred in the distant past also have a reduced influence on the probability distribution of the current event mark. Specifically, there exists a convergent series $\sum_T N_T$ of real numbers such that for event histories H_t and H'_t at time t that differ at only one event at time t' , if $t - t' \geq T$, then

$$\text{TV}(Q_{H_t}^X, Q_{H'_t}^X) < N_T.$$

Along with the above assumptions (A1-A3), the MDP $\mathcal{M}$ under the influence of the exogenous event process must satisfy a set of regularity conditions to guarantee the existence and uniqueness of optimal solutions.

Define a new expected reward function $r_P : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ induced from r by P as

$$r_P(s, a) = \mathbb{E}_{s' \sim P(\cdot|s, a)} [r(s, a, s')],$$

where P is any stochastic kernel on $\mathcal{S}$ given $\mathcal{S} \times \mathcal{A}$. We assume the following regularity conditions on r and Q .

(B1) r_P is a bounded measurable function. Without any loss of generality, we assume that its co-domain is $[0, 1] \subset \mathbb{R}$.

(B2) The function $r_P(s, \cdot)$ is upper semi-continuous for each state $s \in \mathcal{S}$ and any probability distribution $P = Q_H$ induced as a transition kernel on the MDP by any possible external event history H of the temporal process.

(B3) r_P is sup-compact on $\mathcal{S} \times \mathcal{A}$, i.e., for every $s \in \mathcal{S}, r \in \mathbb{R}$ and any probability distribution $P = Q_H$ induced as a transition kernel on the MDP by any possible external event history H of the temporal process, the set $\{a \in \mathcal{A} : r_P(s, a) \geq r\} (\subseteq \mathcal{A})$ is compact.

(B4) Q is strongly continuous.

(B5) Q_H and Q_H^X are strongly continuous for any feasible event history H .

Note that a transition kernel P is said to be strongly continuous if $v(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}[v(s')]$ is continuous and bounded on $\mathcal{S} \times \mathcal{A}$ for every measurable bounded function v on $\mathcal{S}$.

3.1 Example

Non-Markov self-exciting event processes are studied extensively in finance, economics, epidemiology, etc., in the form of variants of Hawkes and jump processes. Although they are studied predominantly in continuous time, we consider discrete-time versions to fit into the traditional discrete-time reinforcement learning framework.

Consider a discrete-time marked Hawkes process $(B_t, X_t)_{t \in \mathbb{N}}$ defined as follows: $(B_t)_{t \in \mathbb{N}}$ is a sequence of random variables taking values in $\{0, 1\}$, with B_t sampled from Bernoulli(p_t), where the intensity p_t at time t depends on the realizations of B_t at the previous time steps as

$$p_1 = \alpha_0, \text{ and } p_t = \alpha_0 + \sum_{t'=1}^{t-1} \alpha_{t-t'} B_{t'}$$

and the marks X_t are real-valued Gaussian random variables clipped to $[-b, b]$, satisfying

$$\begin{aligned} X_1 & \begin{cases} \sim \mathcal{N}_{[-b,b]}(0, 1) & \text{if } B_1 = 1, \\ = 0 & \text{otherwise,} \end{cases} \text{ for } t > 1, \\ X_t & \begin{cases} \sim \mathcal{N}_{[-b,b]} \left(\sum_{t'=1}^{t-1} \beta_{t-t'} B_{t'} X_{t'}, 1 \right) & \text{if } B_t = 1, \\ = 0 & \text{otherwise,} \end{cases} \text{ for } t > 1, \end{aligned}$$

where $\mathcal{N}_{[-b,b]}$ is the normal distribution clipped to be within the interval $[-b, b]$, and (α_t) and (β_t) are sequences that converge sufficiently quickly to zero. The sum $S_t = \sum_{t'=1}^t B_{t'}$ is a self-exciting counting process that is the discrete-time counterpart of the continuous-time Hawkes process.

The above-defined sequence of random variables satisfies Assumptions **(A1)** and **(A2)**, which are required for the external process. X_t is zero when no event occurs, and old events have a reduced and decaying influence on the current events. More precisely, given

two historical sequences $(X_{t'})_{t' < t}$ and $(X'_{t'})_{t' < t}$ that only differ by one event at some time $t' \leq t - T$ with $B_{t'} = X'_{t'} = 0$ and $B_{t'} = 1$, $X_{t'} = x$, the difference in intensities is $p_t - p'_t = \alpha_{t-t'}$, and hence,

$$\begin{aligned} \text{TV}(X_t, X'_t) &\leq \frac{\alpha_{t-t'}}{2} + p_t \text{TV}\left(\mathcal{N}_{[-b,b]}\left(\sum_{t''=1}^{t-1} \beta_{t-t''} B_{t''} X_{t''}, 1\right), \mathcal{N}_{[-b,b]}\left(\sum_{t''=1}^{t-1} \beta_{t-t''} B_{t''} X'_{t''}, 1\right)\right) \\ &\leq \frac{\alpha_{t-t'}}{2} + p_t \text{TV}\left(\mathcal{N}\left(\sum_{t''=1}^{t-1} \beta_{t-t''} B_{t''} X_{t''}, 1\right), \mathcal{N}\left(\sum_{t''=1}^{t-1} \beta_{t-t''} B_{t''} X'_{t''}, 1\right)\right) \\ &= \frac{\alpha_{t-t'}}{2} + p_t \text{erf}\left(\frac{\beta_{t-t'} x}{2\sqrt{2}}\right) < \frac{\alpha_T}{2} + \text{erf}\left(\frac{\beta_T b}{2\sqrt{2}}\right) = N_T. \end{aligned} \quad (2)$$

One can find the details of this calculation in Appendix B. That is, for sufficiently fast converging α_T and β_T , the total variation perturbation N_T due to events older than T time steps approaches zero, satisfying Assumption (A2).

While (α_n) and (β_n) are any general sequences that need to satisfy some regularity conditions (Seol, 2015) and the above bound, they could also decay in a more structured manner, as follows. For instance, for some $c, \lambda > 0$, letting $\alpha_n = ce^{-\lambda n}$ gives a process with exponentially decaying intensity due to past events. In the terminology employed to describe the standard continuous-time Hawkes process, $\alpha_0 = c$ is like the base or background intensity, and $e^{-\lambda x}$ is the excitation function. Similarly, β_n , which determines the distribution of the event, could also be a parametric sequence that decays exponentially or polynomially.

4 Guarantees for the Existence of Almost-Optimal Solutions

Consider the MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, Q, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are closed subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, and $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the reward as a function of the state, action, and resultant next state. Given an external discrete-time temporal process $X = \{(t_i, X_{t_i})\}_i$ influences MDP $\mathcal{M}$, we construct a new augmented MDP $\mathcal{M}_X$, where the marks of all the events that occurred till time t are appended to the state $s \in \mathcal{S}$ to form a new augmented state $\bar{s} \in \bar{\mathcal{S}}$, where

$$\bar{s} = (s, x_t, x_{t-1}, x_{t-2}, x_{t-3}, \dots), \text{ and } \bar{\mathcal{S}} = \mathcal{S} \times \prod_{t=0}^{\infty} \mathcal{X} \left(\subseteq \mathbb{R}^m \times \prod_{t=0}^{\infty} \mathbb{R}^d \right).$$

Here, x_t is the mark of an external event that occurred at time t . If no event occurs at time t , then $x_t = 0$.

Due to the influence of external events, the decision process $(\mathcal{S}, \mathcal{A}, Q, r, \gamma)$ does not remain an MDP because the transition function is perturbed by external events. However, the corresponding decision process $\mathcal{M}_X = (\bar{\mathcal{S}}, \mathcal{A}, \bar{Q}, r, \gamma)$ is an MDP, where $\bar{Q}$ is the transition kernel on the augmented state $\bar{s}$, because all factors that affect the transitions are included in the augmented state $\bar{s} \in \bar{\mathcal{S}}$.

Furthermore, the state space of $\mathcal{M}_X$ is infinite-dimensional. However, Assumptions (A2) and (A3) in Section 3 allow one to study the possibility of finding a suitable policy that only depends on a finite horizon of past events. In this context, we present the following results.

Theorem 2. Suppose the MDP $\mathcal{M}$ and the external process satisfy the assumptions (A1-A3) and the regularity conditions (B1-B5). Then, we have the following.

- (1) There exists a deterministic optimal policy for the augmented MDP $\mathcal{M}_X$, with corresponding optimal value function V^* , that satisfies the optimal Bellman equation.
- (2) For any $\epsilon > 0$, there exists a time horizon T and a policy $\pi^{(T)}$ that depends only on the current state and past T events such that

$$V^{\pi^{(T)}} \geq V^* - \epsilon.$$

That is, for every required maximum suboptimality ϵ , there exists a corresponding “finite-horizon” policy that achieves that ϵ -suboptimal value. Further, the past event horizon T required for ϵ -suboptimality is T that satisfies

$$\sum_{t=T+1}^{\infty} M_t < \frac{\epsilon(1-\gamma)^2}{4} \text{ and } \sum_{t=T+1}^{\infty} N_t < \frac{\epsilon(1-\gamma)^2}{4}. \quad (3)$$

This result guarantees that the formulated problem is well-posed, with an optimal solution that can be approximated by a tractable policy that depends only on the current state and a finite history of events. This approximation can be made as accurate as necessary, depending on the properties of the exogenous event process and the size of the event history. To prove Theorem 2, we establish the following result, whose proof is given in Appendix C.

Lemma 3. Consider a new auxiliary MDP $\mathcal{M}_X^{(T)}$ that has the same underlying stochastic process as the augmented MDP $\mathcal{M}_X$, with one difference: only events in the past T time steps affect the current state transition and event distribution, with events before that having no effect and being effectively zero. Let π be an arbitrary policy as a function of the augmented state $\bar{s} \in \bar{\mathcal{S}}$. That is, π can be either stochastic or deterministic and can depend on the current state $s \in \mathcal{S}$ and any number of past events in $\mathcal{X}$. Then,

$$\left\| V_{\mathcal{M}_X^{(T)}}^{\pi} - V_{\mathcal{M}_X}^{\pi} \right\|_{\infty} \leq \frac{1}{1-\gamma} \left(\|r\|_{\infty} + \gamma \left\| V_{\mathcal{M}_X^{(T)}}^{\pi} \right\|_{\infty} \right) \sum_{t=T+1}^{\infty} (M_t + N_t),$$

where r is the reward function, and $V_{\mathcal{M}}^{\pi}$ denotes the value function of policy π in MDP $\mathcal{M}$.

Proof of Theorem 2 The state space is Borel because it is a countable product of Borel sets. Therefore, part (1) follows from the regularity conditions (B1-B5) satisfying the assumptions of Theorem 4.2.3 in Hernández-Lerma and Lasserre (1996), coupled with the boundedness of the reward function. For part (2), let π^* be the deterministic optimal stationary policy for $\mathcal{M}_X$ and $\pi^{*(T)}$ be the deterministic optimal stationary policy for $\mathcal{M}_X^{(T)}$. From Lemma 3, since $\|r\|_{\infty} \leq 1$ and $\|V\|_{\infty} \leq \frac{1}{1-\gamma}$,

$$V_{\mathcal{M}_X}^{\pi^{*(T)}} \geq V_{\mathcal{M}_X^{(T)}}^{\pi^{*(T)}} - \frac{1}{(1-\gamma)^2} \left(\sum_{t=T+1}^{\infty} (M_t + N_t) \right) \geq V_{\mathcal{M}_X}^{\pi^*} - \frac{1}{(1-\gamma)^2} \left(\sum_{t=T+1}^{\infty} (M_t + N_t) \right)$$

$$\geq V_{\mathcal{M}_X}^{\pi^*} - \frac{2}{(1-\gamma)^2} \left(\sum_{t=T+1}^{\infty} (M_t + N_t) \right).$$

Since both series $\sum_t M_t$ and $\sum_t N_t$ converge, there exists $T \in \mathbb{N}$ satisfying (3). Choosing such an T proves the result. $\blacksquare$

5 A Policy Iteration Algorithm

Considering that we have established the existence of an ϵ -optimal policy that is a function of only a finite event horizon, the aim is to find such a policy. To this end, we propose a policy iteration algorithm that alternates between approximate policy evaluation and approximate policy improvement. This procedure is listed in Algorithm 1.

The policy evaluation step considers candidate value functions that are functions of a finite event horizon. We define the value function in $\mathcal{M}_X$ at an infinitely augmented state as a function of the state augmented by just a finite event horizon, by sampling the previous older events from the posterior distribution μ_1 of the event process. In practice, μ_1 is the actual event process. Thus, the value function can be approximated using Monte Carlo methods. For the purpose of analysis, we consider the exact evaluation of the approximate value function.

In the policy improvement step, the policy is improved based on the reward and value function resulting from one transition. This transition can be due to past events from any arbitrary distribution μ_2 . The deterministic policy at an augmented state depends only on the finite event horizon and is the action that maximizes the right-hand side, which depends on the approximate value function.

Solving the optimization problem in the policy improvement step requires knowledge of models $r(s, a, s')$, Q_H , and Q_H^X for arbitrary histories H . In this work, we ignore any approximation errors that arise due to estimating the models and analyze the algorithm by assuming the exact step is taken as defined.

Algorithm 1 Policy Iteration

Start with arbitrary deterministic policy $\pi_0 : \mathcal{S} \times \mathcal{X}^{T+1} \rightarrow \mathcal{A}$

for each $k \in \{0, 1, \dots\}$ **till** termination **do**

 // Approximate Policy Evaluation

$$\hat{V}_k((s, x_{0:\infty})) = \mathbb{E}_{x'_{T+1:\infty} \sim \mu_1} V^{\pi_k}((s, x_{0:T}, x'_{T+1:\infty})).$$

 // Approximate Policy Improvement

$$\pi_{k+1}((s, x_{0:\infty})) = \underset{a \in A}{\operatorname{argmax}} \mathbb{E}_{\substack{s' \sim Q_{x_{0:T}, x'_{T+1:\infty}}(\cdot | s, a) \\ x' \sim Q_{x_{0:T}, x'_{T+1:\infty}}^X(\cdot | s) \\ x'_{T+1:\infty} \sim \mu_2}} \left[r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1})) \right].$$

end for

5.1 Analysis

For the analysis of this algorithm, we establish the following results, which bound the difference in the value function at two different augmented states when they differ in events that are older than T time steps.

Lemma 4. *Let π be a policy that depends only on the current state and past T events. Then,*

$$\sup_{\bar{s}=(s, x_{0:\infty})} \left| V^\pi(\bar{s}) - V^\pi((s, x_{0:T}, (0)_{T+1:\infty})) \right| < \frac{1}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t).$$

Here, $(s, x_{0:T}, (0)_{T+1:\infty}) = (s, x_0, x_1, \dots, x_T, 0, 0, \dots) \in \bar{S}$ is the extended state that has been essentially “truncated” and made finite by replacing events older than T time steps with zero. The proof of Lemma 4 is provided in Section C.1.

Corollary 5. *For a given policy π that is a function of events only in the past T time steps, and for any two augmented states $\bar{s}_1$ and $\bar{s}_2$ that have the same state and differ in any number of events that occurred before the previous T time steps,*

$$\left| V^\pi(\bar{s}_1) - V^\pi(\bar{s}_2) \right| \leq \frac{2}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t).$$

Essentially, this means that if a policy considers only events in the past T time steps, its value at states that differ at older events varies by an amount proportional to the extent of nonstationarity induced by events older than T time steps. This helps characterize the behavior of the policy iteration procedure described in Algorithm 1. While exact policy iteration results in a sequence of policies that converge to the optimal policy, our approximate policy iteration produces policies whose value functions satisfy the following result, the proof of which is provided later in this section.

Lemma 6.

$$V^{\pi_{k+1}} \geq V^{\pi_k} - \frac{2(1+\gamma)}{(1-\gamma)^3} \sum_{t=T+1}^{\infty} (M_t + N_t).$$

That is, while there is no guarantee that the sequence of value functions $\{V^{\pi_k}\}_k$ is non-decreasing, Lemma 6 guarantees that they do not decrease by more than a specific amount.

This is due to the approximation error $\epsilon = \frac{2}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t)$ at each time step induced by the nonstationarity due to exogenous events.

In general, approximate policy iteration algorithms converge under the assumption that the approximation errors gradually vanish over the course of the algorithm. However, our approximation error of ϵ remains constant throughout, leading to the above-mentioned situation. Intuitively, such a small approximation error ϵ should only disturb the convergence of the algorithm when we are ϵ -close to the optimal solution. During the

initial iterations, when far from the solution, such small approximation errors are insignificant.

Next, we establish that the value function can degrade only in regions of the state space where the Bellman error is close to zero. A Bellman error of zero generally, although not always, corresponds to an optimal policy. This interplay between the Bellman error and approximation error in our algorithm is formalized by the following result.

Theorem 7. *At iteration k of the policy iteration procedure given in Algorithm 1, for any augmented state $\bar{s} \in \bar{\mathcal{S}} = \mathcal{S} \times \prod_{t=0}^{\infty} \mathcal{X}$, at least one of the following holds:*

$$V^{\pi_{k+1}}(\bar{s}) \geq V^{\pi_k}(\bar{s}), \text{ or} \quad (4)$$

$$|\mathcal{F} V^{\pi_k}(\bar{s}) - V^{\pi_k}(\bar{s})| < \frac{4(1+\gamma)}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t), \quad (5)$$

where V^{π} denotes the value function of policy π in MDP $\mathcal{M}_X$.

That is, the policy's performance improves everywhere except in the region of the state space where the Bellman error is very small. The faster the decay of the exogenous event influence, the smaller the approximation error and the smaller the region of the state space in which the policy may not improve.

Corollary 8. *Consider an MDP with the exogenous process being a discrete-time Hawkes process as described in Section 3.1, with an exponentially decaying excitation function $c_{\alpha} e^{-\lambda_{\alpha} t}$, along with an exponentially decaying $\beta_t = c_{\beta} e^{-\lambda_{\beta} t}$ and state transition perturbations that decay as $M_t = \bar{c} e^{-\bar{\lambda} t}$. Then, each step k of the policy iteration algorithm is guaranteed to keep the value function of the policy non-decreasing everywhere in the state space except regions $\bar{s} \in \bar{\mathcal{S}}$ satisfying*

$$|\mathcal{F} V^{\pi_k}(\bar{s}) - V^{\pi_k}(\bar{s})| < \frac{4(1+\gamma)}{(1-\gamma)^2} \left(\frac{\bar{c}}{\bar{\lambda}} e^{-\bar{\lambda} T} + \frac{c_{\alpha}}{\lambda_{\alpha}} e^{-\lambda_{\alpha} T} + \frac{c_{\beta} b}{\sqrt{2\pi} \lambda_{\beta}} e^{-\lambda_{\beta} T} \right).$$

As the time window T considered increases, the error reduces as $e^{-\{\bar{\lambda}, \lambda_{\alpha}, \lambda_{\beta}\}T}$. Hence, depending on the rate of influence decay, increasing the time window beyond a certain point results in diminishing returns. The error also has a proportional dependence on c_{α} , the base intensity of the Hawkes process, and an inverse dependency on λ_{α} , the decay of the excitation function of the Hawkes process.

Proof of Theorem 7 Whether (4) or (5) holds depends on the state $\bar{s}$ falls in the “sub-optimality” set B or not, which is the set of all states $\bar{s} = (s, x_{0:\infty}) \in \bar{\mathcal{S}}$ that satisfy

$$\begin{aligned} & \mathbb{E}_{s' \sim Q_{x_{0:T}, x'_{T+1:\infty}}(\cdot | s, a)} \left[r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1})) \right] \\ & x' \sim Q_{x_{0:T}, x'_{T+1:\infty}}^X(\cdot) \\ & x'_{T+1:\infty} \sim \mu_2 \\ & a \sim \pi_k(\bar{s}) \end{aligned}$$

$$\begin{aligned}
 &< \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim Q_{x_{0:T}, x'_{T+1:\infty}}(\cdot|s, a)} \left[r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1})) \right] - \delta \\
 &\quad x' \sim Q_{x_{0:T}, x'_{T+1:\infty}}^X(\cdot) \\
 &\quad x'_{T+1:\infty} \sim \mu_2
 \end{aligned}$$

for some suitable $\delta > 0$ that is yet to be chosen. It is to be noted that whether or not this condition holds, the above inequality or the corresponding equality holds for $\delta = 0$ by the definition of π_{k+1} .

By definition of $\hat{V}_k$ and using corollary 5, we have

$$\hat{V}_k((s, x_{0:T})) = \mathbb{E}_{x'_{T+1:\infty} \sim \mu_2} [V^{\pi_k}((s, x_{0:T}, x'_{T+1:\infty}))] \quad (6)$$

$$\begin{aligned}
 |V^{\pi_k}((s, x_{0:\infty})) - \hat{V}_k((s, x_{0:T}))| &= |\mathbb{E}_{x'_{T+1:\infty} \sim \mu_2} [V^{\pi_k}((s, x_{0:\infty})) - V^{\pi_k}((s, x_{0:T}, x'_{T+1:\infty}))]| \\
 &\leq \mathbb{E}_{x'_{T+1:\infty} \sim \mu_2} \left[\frac{2}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) \right] \quad (7)
 \end{aligned}$$

Using the Bellman equation, we have

$$\begin{aligned}
 V^{\pi_k}((s, x_{0:\infty})) &= \mathbb{E}_{\substack{s' \sim Q_{x_{0:\infty}}(\cdot|s, a) \\ x' \sim Q_{x_{0:\infty}}^X(\cdot) \\ a \sim \pi_k(\bar{s})}} [r(s, a, s') + \gamma V^{\pi_k}((s', x', x_{0:\infty}))] \\
 V^{\pi_k}((s, x_{0:\infty})) &\leq \mathbb{E}_{\substack{s' \sim Q_{x_{0:\infty}}(\cdot|s, a) \\ x' \sim Q_{x_{0:\infty}}^X(\cdot) \\ a \sim \pi_k(\bar{s})}} [r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1}))] \\
 &\quad + \frac{2\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) \quad (8)
 \end{aligned}$$

$$\begin{aligned}
 &\leq \mathbb{E}_{\substack{s' \sim Q_{x_{0:T}, x'_{T+1:\infty}}(\cdot|s, a) \\ x' \sim Q_{x_{0:T}, x'_{T+1:\infty}}^X(\cdot) \\ x'_{T+1:\infty} \sim \mu_2 \\ a \sim \pi_k(\bar{s})}} [r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1}))] \\
 &\quad + \frac{1}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t) + \frac{2\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) \quad (9)
 \end{aligned}$$

The inequality in (8) follows from (7) and the inequality in (9) follows from (31). Suppose that $\bar{s} \in B$. Then,

$$\begin{aligned}
 V^{\pi_k}((s, x_{0:\infty})) &\leq \mathbb{E}_{\substack{s' \sim Q_{x_{0:T}, x'_{T+1:\infty}}(\cdot|s, a) \\ x' \sim Q_{x_{0:T}, x'_{T+1:\infty}}^X(\cdot) \\ x'_{T+1:\infty} \sim \mu_2 \\ a \sim \pi_{k+1}(\bar{s})}} [r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1}))] \\
 &\quad + \frac{1+\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) - \delta \quad (10)
 \end{aligned}$$

$$\begin{aligned}
 &\leq \mathbb{E}_{\substack{s' \sim Q_{x_0:T, x'_{T+1:\infty}}(\cdot|s, a) \\ x' \sim Q_{x_0:T, x'_{T+1:\infty}}^X(\cdot) \\ x'_{T+1:\infty} \sim \mu_2 \\ a \sim \pi_{k+1}(\bar{s})}} [r(s, a, s') + \gamma V^{\pi_k}((s', x', x_{0:\infty}))] \\
 &\quad + \frac{2\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) + \frac{1+\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) - \delta \tag{11}
 \end{aligned}$$

$$\begin{aligned}
 &\leq \mathbb{E}_{\substack{s' \sim Q_{x_0:\infty}(\cdot|s, a) \\ x' \sim Q_{x_0:\infty}^X(\cdot) \\ a \sim \pi_{k+1}(\bar{s})}} [r(s, a, s') + \gamma V^{\pi_k}((s', x', x_{0:\infty}))] \\
 &\quad + \frac{1}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t) + \frac{1+3\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) - \delta \tag{12}
 \end{aligned}$$

$$= \mathbb{E}_{\pi_{k+1}} [r(s, a, s') + \gamma V^{\pi_k}(\bar{s}')] + \frac{2(1+\gamma)}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) - \delta \tag{13}$$

$$\leq \mathbb{E}_{\pi_{k+1}} \left[\sum_{t=0}^{\infty} \gamma^t r(s, a, s') \right] + \frac{1}{1-\gamma} \left[\frac{2(1+\gamma)}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) - \delta \right] \tag{14}$$

$$\begin{aligned}
 &= V^{\pi_{k+1}}(\bar{s}) + \frac{1}{1-\gamma} \left[\frac{2(1+\gamma)}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) - \delta \right] \tag{15} \\
 &= V^{\pi_{k+1}}(\bar{s}),
 \end{aligned}$$

$$\text{for } \delta = \frac{2(1+\gamma)}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t).$$

The inequality in (10) follows from the fact that $\bar{s} \in B$. The inequality in (11) follows from (7) and the inequality in (12) follows from (31). Recursively expanding V^{π_k} in (13) gives (14)

Now, suppose $\bar{s} \notin B$. We need to show that this implies (5), that is, the Bellman error should be very small. It is always true that $\mathcal{T}V \geq V$. In the other direction, for any $\bar{s} = (s, x_{0:\infty}) \in \bar{\mathcal{S}}$,

Since $\bar{s} \notin B$,

$$\begin{aligned}
 &\max_{a \in \mathcal{A}} \mathbb{E}_{\substack{s' \sim Q_{x_0:T, x'_{T+1:\infty}}(\cdot|s, a) \\ x' \sim Q_{x_0:T, x'_{T+1:\infty}}^X(\cdot) \\ x'_{T+1:\infty} \sim \mu_2}} [r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1}))] \\
 &\leq \mathbb{E}_{\substack{s' \sim Q_{x_0:T, x'_{T+1:\infty}}(\cdot|s, a) \\ x' \sim Q_{x_0:T, x'_{T+1:\infty}}^X(\cdot) \\ x'_{T+1:\infty} \sim \mu_2 \\ a = \pi_k(\bar{s})}} [r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1}))] + \delta \tag{16}
 \end{aligned}$$

We can lower bound bound the left hand side of (16) using (7) and (31)

$$\begin{aligned}
 & \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim Q_{x_{0:T}, x'_{T+1:\infty}}(.|s, a)} \left[r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1})) \right] \\
 & \quad x' \sim Q_{x_{0:T}, x'_{T+1:\infty}}^X(.) \\
 & \quad x'_{T+1:\infty} \sim \mu_2 \\
 & \geq \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim Q_{x_{0:T}, x'_{T+1:\infty}}(.|s, a)} \left[r(s, a, s') + \gamma V^{\pi_k}((s', x', x_{0:\infty})) \right] \\
 & \quad x' \sim Q_{x_{0:T}, x'_{T+1:\infty}}^X(.) \\
 & \quad x'_{T+1:\infty} \sim \mu_2 \\
 & \quad - \frac{2}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) \\
 & \geq \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim Q_{x_{0:\infty}}(.|s, a)} \left[r(s, a, s') + \gamma V^{\pi_k}((s', x', x_{0:\infty})) \right] \\
 & \quad x' \sim Q_{x_{0:\infty}}^X(.) \\
 & \quad - \frac{1}{(1-\gamma)} \sum_{t=T+1}^{\infty} (M_t + N_t) - \frac{2}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) \quad (17)
 \end{aligned}$$

We can upper bound the right hand side of (16) using (7) and (31)

$$\begin{aligned}
 & \mathbb{E}_{s' \sim Q_{x_{0:T}, x'_{T+1:\infty}}(.|s, a)} \left[r(s, a, s') + \gamma \hat{V}_k((s', x', x_{0:T-1})) \right] \\
 & \quad x' \sim Q_{x_{0:T}, x'_{T+1:\infty}}^X(.) \\
 & \quad x'_{T+1:\infty} \sim \mu_2 \\
 & \quad a = \pi_k(\bar{s}) \\
 & \leq \mathbb{E}_{s' \sim Q_{x_{0:T}, x'_{T+1:\infty}}(.|s, a)} \left[r(s, a, s') + \gamma V^{\pi_k}((s', x', x_{0:\infty})) \right] \\
 & \quad x' \sim Q_{x_{0:T}, x'_{T+1:\infty}}^X(.) \\
 & \quad x'_{T+1:\infty} \sim \mu_2 \\
 & \quad a = \pi_k(\bar{s}) \\
 & \quad + \frac{2\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) \\
 & \leq \mathbb{E}_{s' \sim Q_{x_{0:\infty}}(.|s, a)} \left[r(s, a, s') + \gamma V^{\pi_k}((s', x', x_{0:\infty})) \right] \\
 & \quad x' \sim Q_{x_{0:\infty}}^X(.) \\
 & \quad a = \pi_k(\bar{s}) \\
 & \quad + \frac{1}{(1-\gamma)} \sum_{t=T+1}^{\infty} (M_t + N_t) + \frac{2\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) \quad (18)
 \end{aligned}$$

Therefore, from (17) and (18), and taking $\delta = \frac{2(1+\gamma)}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t)$ we get

$$\begin{aligned}
 & \frac{-2\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) - \frac{1}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t) \\
 & \quad + \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim Q_{x_{0:\infty}}(.|s, a)} \left[r(s, a, s') + \gamma V^{\pi_k}((s', x', x_{0:\infty})) \right] \\
 & \quad x' \sim Q_{x_{0:\infty}}^X(.)
 \end{aligned}$$

$$\begin{aligned}
 &\leq \mathbb{E}_{\substack{s' \sim Q_{x_{0:\infty}}(\cdot|s,a) \\ x' \sim Q_{x_{0:\infty}}^X(\cdot) \\ a = \pi_k(\bar{s})}} [r(s, a, s') + \gamma V^{\pi_k}((s', x', x_{0:\infty}))] \\
 &\quad + \frac{2\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) + \frac{1}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t) + \delta \\
 (TV^{\pi_k})(\bar{s}) &- \frac{1+\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) \\
 &\leq V^{\pi_k}(\bar{s}) + \frac{1+\gamma}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) + \delta \\
 (TV^{\pi_k})(\bar{s}) - V^{\pi_k}(\bar{s}) &\leq \frac{4(1+\gamma)}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t)
 \end{aligned}$$

$$\text{Therefore, } |(TV^{\pi_k})(\bar{s}) - V^{\pi_k}(\bar{s})| \leq \frac{4(1+\gamma)}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t)$$

Proof of Lemma 6 Setting $\delta = 0$ in the first part of the above proof gives us the Lemma, for which there is no assumption on the extent of policy improvement.

$$\begin{aligned}
 V^{\pi_k}(\bar{s}) &\leq V^{\pi_{k+1}}(\bar{s}) + \frac{1}{1-\gamma} \left[\frac{2(1+\gamma)}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t) \right] \\
 V^{\pi_k}(\bar{s}) &\leq V^{\pi_{k+1}}(\bar{s}) + \frac{2(1+\gamma)}{(1-\gamma)^3} \sum_{t=T+1}^{\infty} (M_t + N_t)
 \end{aligned}$$

■

In Theorem 7, in regions of the augmented state space without guaranteed policy improvement, the Bellman error depends on the approximation error due to the ignoring old events. It is the cumulative effect of all events older than T on the current events, as well as the state transition kernel. Specifically, it is proportional to $\sum_{t=T+1}^{\infty} (M_t + N_t)$, where T is the time window beyond which events are discarded, M_t is an upper bound on the total variation disturbance on the state transition kernel caused by an event older than t time steps, and N_t bounds the total variation effect due to old events on the event process itself. The faster the decay of the exogenous event influence, the smaller the truncation error due to nonstationarity.

Proof of Corollary 8 From (2), the extra error introduced is proportional to

$$\begin{aligned}
 \sum_{t=T+1}^{\infty} (M_t + N_t) &\leq \sum_{t=T+1}^{\infty} \left(\bar{c}e^{-\bar{\lambda}t} + c_{\alpha}e^{-\lambda_{\alpha}t} + \operatorname{erf}\left(\frac{c_{\beta}be^{-\lambda_{\beta}t}}{2\sqrt{2}}\right) \right) \\
 &\leq \sum_{t=T+1}^{\infty} \left(\bar{c}e^{-\bar{\lambda}t} + c_{\alpha}e^{-\lambda_{\alpha}t} + \sqrt{1 - \exp\left(\frac{-c_{\beta}^2b^2}{2\pi}e^{-2\lambda_{\beta}t}\right)} \right)
 \end{aligned}$$

$$\begin{aligned}
 &\leq \sum_{t=T+1}^{\infty} \left(\bar{c}e^{-\bar{\lambda}t} + c_{\alpha}e^{-\lambda_{\alpha}t} + \frac{c_{\beta}b}{\sqrt{2\pi}}e^{-\lambda_{\beta}t} \right) \\
 &\leq \int_{t=T}^{\infty} \left(\bar{c}e^{-\bar{\lambda}t} + c_{\alpha}e^{-\lambda_{\alpha}t} + \frac{c_{\beta}b}{\sqrt{2\pi}}e^{-\lambda_{\beta}t} \right) dt \\
 &= \frac{\bar{c}}{\bar{\lambda}}e^{-\bar{\lambda}T} + \frac{c_{\alpha}}{\lambda_{\alpha}}e^{-\lambda_{\alpha}T} + \frac{c_{\beta}b}{\sqrt{2\pi}\lambda_{\beta}}e^{-\lambda_{\beta}T}.
 \end{aligned}$$

■

6 Least-Squares Policy Iteration

Although the algorithm presented in Section 5 specifies the value function and its update, in practice, it is essential to note that, in practice, this function must be learned through samples acquired from the environment and is typically approximated using a selected class of functions.

In this section, we analyze the sample complexity associated with policy iteration, in which the policy evaluation step is achieved by employing pathwise Least-Squares Temporal Difference (LSTD) learning (Lazaric et al., 2012). This method utilizes samples derived from a singular sample path generated by the policy. We begin by establishing bounds for the evaluation error within the linear function space case, along with certain regularity conditions imposed on the MDP. Subsequently, we employ these results to bound the sub-optimality of the least-squares policy iteration.

In contrast to Lazaric et al. (2012), our study deals with an MDP characterized by an infinite-dimensional augmented state space, a stochastic reward function, and, more importantly, an additional error due to the use of tractable policies and value functions on a finite-dimensional domain. This results in additional sample complexity, even for evaluating a given policy, depending on the extent to which exogenous events affect the evolution of the states. In this section, we quantify the precise nature of these errors.

Consider a function class $\mathcal{F}$ consisting of functions that can approximate the value function on the augmented MDP, $\mathcal{F} \subset \{f : \bar{\mathcal{S}} \rightarrow \mathbb{R}\}$. Since the domain of the functions in this class is infinite-dimensional, this can be further broken down into two approximations for dealing with these functions practically: a truncation operation to make the domain finite-dimensional, and then another class of functions that are then used to cover possible functions on this domain, i.e.,

$$\mathcal{F} = \left\{ f \circ f_{\text{trunc}}^{(T)} : f \in \mathcal{F}^{(T)} \right\}, \text{ and } f_{\text{trunc}}^{(T)} : \bar{\mathcal{S}} \rightarrow \mathcal{S} \times \prod_{t=0}^T \mathcal{X},$$

where

$$\mathcal{F}^{(T)} = \left\{ f : \mathcal{S} \times \prod_{t=0}^T \mathcal{X} \rightarrow \mathbb{R} \right\}$$

Here, $f_{\text{trunc}}^{(T)}$ is a fixed truncation function that removes extra events that are older than T time steps while preserving the current state and new events, and $\mathcal{F}^{(T)}$ is a function

class defined on this new truncated domain. For a fixed T , finding the best function in $\mathcal{F}$ is the same as finding the best function in $\mathcal{F}^{(T)}$.

The standard approximating function space considered is an arbitrary finite-dimensional function space. Specifically, assume that there are d basis functions $\varphi_i : \mathcal{S} \times \mathcal{X}^{T+1} \rightarrow \mathbb{R}$ that linearly span $\mathcal{F}^{(T)}$ and are bounded by L . That is, for each $f \in \mathcal{F}^{(T)}$, there exists $\alpha \in \mathbb{R}^d$ such that $f = \sum_{i \in [d]} \alpha_i \varphi_i$. These basis functions together form a feature representation function $\phi : \mathcal{S} \times \mathcal{X}^{T+1} \rightarrow \mathbb{R}^d$, defined by $\phi(x) = (\varphi_1(x), \dots, \varphi_d(x))$. Further, we define $\widetilde{\mathcal{F}}$ as the class of functions obtained by truncating the functions in $\mathcal{F}$ to be bounded by L .

We now consider the adaptation of the standard Least-Squares Policy Iteration to our setting. The policy evaluation step is described in Section 6.1, wherein a value function is learned as a function of the current state and a finite history of events, with the previous event values set to zero for simplicity. In the policy improvement step, a greedy policy is defined as a function of the state and a finite history of events. This induces an additional approximation error in the model. The overall procedure is listed in Algorithm 2.

Algorithm 2 Approximate Least-Squares Policy Iteration

Given: Basis function set $\phi = (\varphi_1, \dots, \varphi_d) : \mathcal{S} \times \mathcal{X}^{T+1} \rightarrow \mathbb{R}^d$

Start with arbitrary deterministic policy $\pi_0 : \mathcal{S} \times \mathcal{X}^{T+1} \rightarrow \mathcal{A}$

for each $k \in \{0, 1, \dots\}$ **till** termination **do**

 // Least-Squares Temporal Difference Learning for policy evaluation

 Obtain samples $\bar{s}_1, \dots, \bar{s}_n$ and rewards $r_1, \dots, r_n$ using policy π_k

 Construct the feature vectors as

$$\Phi = \left(\phi \left(\bar{s}_1^{(T)} \right), \dots, \phi \left(\bar{s}_n^{(T)} \right) \right). \quad (19)$$

 Obtain value function $\widetilde{V} = \text{Trunc} \left(\sum_{i \in [d]} \hat{\alpha}_i \varphi_i \right)$, where

$$\hat{\alpha} = \left[\Phi^\top (I - \gamma \hat{P}) \right]^\dagger \Phi^\top r.$$

 // Approximate Greedy Policy Improvement

 Define new policy as the best approximating greedy policy that is a function of the state and just a finite history of events, as

$$\pi_{k+1}((s, x_{0:\infty})) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \mathbb{E}_{\substack{s' \sim Q_{x_{0:T},(0)}(\cdot | s, a) \\ x' \sim Q_{x_{0:T},(0)}^X(\cdot | s)}} \left[r(s, a, s') + \gamma \widetilde{V}_k((s', x', x_{0:T-1})) \right].$$

end for

6.1 Approximate Policy Evaluation

The error between the true value function and the learned truncated value function can be broken down in terms of the errors introduced by truncating the infinite-dimensional domain and employing linear function approximation, as well as the stochasticity of the samples used to learn the approximate function. The aim is to provide a bound to $\|V - \tilde{V}\|_\rho$, where V is the true value function on $\mathcal{S}$, $\tilde{V}$ is the learned value function in $\mathcal{F}$, ρ is some distribution over $\mathcal{S}$, and the norm $\|\cdot\|_\rho$ is the $l^2(\rho)$ -norm, which is the expected l^2 norm of the value of the function w.r.t the measure ρ , i.e,

$$\|f\|_\rho^2 = \int_{\mathcal{S}} f(x)^2 \rho(dx).$$

While the $f_{\text{trunc}}^{(T)}$ operator ignores old events, we also define a projection operator Π^{trunc} onto the function space defined on the truncated domain, as

$$\Pi^{\text{trunc}} V = \underset{f: \mathcal{S} \times \prod_{t=0}^T \mathcal{X} \rightarrow \mathbb{R}}{\text{argmin}} \|f - V\|_\rho.$$

which finds the best approximation that does not depend on the old events, where ‘best’ is defined in expectation with respect to ρ . The best approximate value function $\tilde{V}$ is that function in $\mathcal{F}$ (or equivalently $\mathcal{F}^{(T)}$) that minimizes the above-expected norm difference.

Suppose we are given N samples $\bar{s}_1, \dots, \bar{s}_N$ from a Markov chain induced by policy π in the augmented MDP $\mathcal{M}_X$, and the corresponding rewards $r_1, \dots, r_N$. We minimize the empirical norm difference at these points, defined as

$$\|f\|_N = \left(\frac{1}{N} \sum_{t=1}^N f(x_t)^2 \right)^{\frac{1}{2}}.$$

The above function defines a norm, along with a corresponding inner product, in a new inner product space $\mathcal{F}_N$, which is the space of all (N) values of the functions at the given sample points. This can be considered a subset of $\mathbb{R}^N$ with an inner product. Here, $\mathcal{F}_N = \{(f(\bar{s}_1), \dots, f(\bar{s}_N)) : f \in \mathcal{F}\} = \{(f(s_1^{(T)}), \dots, f(s_N^{(T)})) : f \in \mathcal{F}^{(T)}\} = \{\Phi\alpha : \alpha \in \mathbb{R}^d\}$, where $\Phi = (\phi(\bar{s}_1^{(T)}), \dots, \phi(\bar{s}_N^{(T)}))_{N \times d}$ and $\bar{s}_i^{(T)} = f_{\text{trunc}}^{(T)}(\bar{s}_i)$. In pathwise LSTD, given a sequence of states $\bar{s}_1, \dots, \bar{s}_n$ and corresponding rewards $r_1, \dots, r_n$ obtained by following policy π , the value function is approximated as

$$\hat{V} = \sum_{i \in [d]} \hat{\alpha}_i \phi_i, \text{ for } \hat{\alpha} = [\Phi^\top (I - \gamma \hat{P})]^\dagger \Phi^\top r,$$

where $\Phi = (\phi(\bar{s}_1^{(T)}), \dots, \phi(\bar{s}_N^{(T)}))_{N \times d}$ is the feature matrix, $\bar{s}_i^{(T)} = f_{\text{trunc}}^{(T)}(\bar{s}_i)$, and $\hat{P}_{n \times n} = (\mathbb{I}\{j = i + 1\})_{i,j}$ is the pathwise empirical transition matrix. Since the reward function takes values in $[0, 1]$, the true value function is bounded by $1/(1 - \gamma)$. Therefore, to obtain the best approximation, the above obtained $\hat{V}$ can be truncated to take values in $[0, 1/(1 - \gamma)]$ to obtain the final estimate $\tilde{V}$.

6.2 Sample complexity of Policy Evaluation

The aim is to provide a bound to the expected error $\|V - \tilde{V}\|_\rho$, where V is the true value function on $\mathcal{S}$, $\tilde{V}$ is the learned value function in $\mathcal{F}$, ρ is a distribution over $\mathcal{S}$, and the norm $\|\cdot\|_\rho$ is the $l^2(\rho)$ -norm.

To determine the expected error of the value estimates in the specified norm, the intermediate task is to bound the empirical error, which is the difference between the true solution and the estimate on the given data points that have been used to determine the estimate. We derive a bound on $\|V - \hat{V}\|_N$ as a function of the number of data points and the complexity of the function class under consideration.

Lemma 9. *Let $v = (V(\tilde{s}_t))_{t \in [N]}$ and $\hat{v} = (\hat{V}(\tilde{s}_t))_{t \in [N]}$. Then, with probability at least $1 - \delta$,*

$$\|v - \hat{v}\|_N \leq \frac{1}{\sqrt{1 - \gamma^2}} \|v - \hat{\Pi}v\|_N + \frac{L}{(1 - \gamma)^2} \sqrt{\frac{d}{v_N}} \left(\sqrt{\frac{2 \log(2d/\delta)}{N}} + \frac{1}{N} \right), \quad (20)$$

where N is the number of samples used, $\hat{\Pi}$ is the projection onto $\mathcal{F}_N$, d is the dimensionality of the linear function space considered, L is the upper bound of the basis functions, and v_N is the smallest positive eigenvalue of the Gram matrix $\Phi^\top \Phi$.

Proof See Appendix D.1.1. ■

The inequality (20) is almost identical to the one in Lazaric et al. (2012) because we are still operating in the range space of the basis functions restricted to a finite set of states; therefore, the structure of the MDP in our setting does not affect the analysis. Because the proof follows a similar path, we have included in the Appendix those steps that differ from the original proof, which are mainly due to the noise terms taken in the martingale difference sequence. In our setting, these noise terms also include stochasticity due to the reward, in addition to the transition function, leading to a difference in the final expression.

While this final inequality is quite similar, we shall see that when considering the expected error in the augmented state space, we can bring out the properties of our setting by splitting the expected error in terms of the error due to function truncation and the inherent error due to linear function approximation.

The above inequality bounds the average error of the estimated value function, which is evaluated only at the points obtained as samples from the environment. It is desirable to study the error in the approximate value function learned as an expectation over the states with respect to a distribution ρ . Generally, ρ is taken to be the stationary distribution of the Markov chain induced in the MDP by the policy π . To obtain such a result, additional assumptions are imposed on this Markov chain, such as the speed of convergence to its stationary distribution.

For linear function spaces, assuming that the policy induces a β -mixing Markov chain gives the following generalization bound for policy evaluation.

Theorem 10. *Assume that the policy π induces on the MDP $\mathcal{M}_X$ a β -mixing Markov chain with parameters $\bar{\beta}$, b , κ and stationary distribution ρ . Let $\tilde{s}_1, \dots, \tilde{s}_{N+\tilde{N}}$ be a sample path obtained*

using the policy π . Suppose the first $\tilde{N}$ samples are discarded for Markov chain burn-in, and the remaining N samples are used to compute the truncated least-squares path-wise estimate $\tilde{V}$ of the true value function V . Let ν be a lower bound on the eigenvalues of the sample-based Gram matrix that holds with a probability $1 - \delta/4$. Then, provided the number of discarded samples is $\tilde{N} = \left(\frac{1}{b} \log \frac{2e\tilde{\beta}n}{\delta}\right)^{\frac{1}{\kappa}}$, with probability at least $1 - \delta$,

$$\begin{aligned} \|\tilde{V} - V\|_\rho &\leq \frac{4\sqrt{2}}{\sqrt{1-\gamma^2}} \left(\frac{3}{(1-\gamma^2)} \sum_{t=T+1}^{\infty} (M_t + N_t) + \|\Pi^{\text{trunc}} V - \Pi V\|_\rho \right) \\ &\quad + \frac{2L}{(1-\gamma)^2} \sqrt{\frac{d}{\nu}} \left(\sqrt{\frac{2 \log(8d/\delta)}{N}} + \frac{1}{N} \right) + \epsilon_1(N, \tilde{\beta}, b, \kappa, d, \delta) + 2\sqrt{2}\epsilon_2(N, \tilde{\beta}, b, \kappa, \delta), \end{aligned}$$

$$\begin{aligned} \text{where } \epsilon_1(N, d, \delta) &= \frac{24}{1-\gamma} \sqrt{\frac{2\Lambda_1(N, \tilde{\beta}, d, \delta/4)}{N} \max \left\{ \frac{\Lambda_1(N, \tilde{\beta}, d, \delta/4)}{b}, 1 \right\}^{\frac{1}{\kappa}}}, \\ \epsilon_2(N, \delta) &= 12 \left(\frac{1}{1-\gamma} + L \|\alpha^*\| \right) \sqrt{\frac{2\Lambda_2(N, \tilde{\beta}, \delta/4)}{N} \max \left\{ \frac{\Lambda_2(N, \tilde{\beta}, \delta/4)}{b}, 1 \right\}^{\frac{1}{\kappa}}}, \end{aligned}$$

V is the true value function, $\tilde{V}$ is the learned value function clipped at $\frac{1}{1-\gamma}$, $\Pi^{\text{trunc}} V$ is the best possible value function on the truncated state space, ΠV is the best approximating value function in the linear function space considered, M_t and N_t are the upper bounds on the total variation due to exogenous events older than t time steps induced in the transition dynamics and event mark distribution, respectively, $\Lambda_1(N, d, \delta)$ and $\Lambda_2(N, \delta)$ are as defined in Lemma 13 in Section D, and α^* is such that $\Pi V = \sum_i \alpha_i^* \varphi_i$.

This result breaks down the overall error into individual components, namely, the approximation error due to nonstationarity, the inherent error due to linear function approximation, the error due to the stochasticity of the samples that reduces with the number of samples, and other errors due to mixing of the Markov chain, the dimensionality of the function space, etc.

The effect of nonstationarity is reflected in the first term proportional to $\sum_t (M_t + N_t)$ for a general exogenous process whose events have an effect that decays as M_t and N_t on the state transition and next event distribution, respectively. For a specific temporal process, such as the discrete Hawkes process described in Corollary 8, this term is replaced by its corresponding upper bound $\left(\frac{\bar{c}}{\lambda} e^{-\bar{\lambda}T} + \frac{c_\alpha}{\lambda_\alpha} e^{-\lambda_\alpha T} + \frac{c_\beta}{\sqrt{2\pi}\lambda_\beta} e^{-\lambda_\beta T} \right)$ that depends on the parameters of the process.

The detailed background for the conditions on the MDP under which the results hold, the exact expressions for Λ_1 and Λ_2 , and the supporting generalization Lemmas needed for the proof are provided in Section D.

Proof of Theorem 10 Following proof of Theorem 5 in Lazaric et al. (2012),

$$2 \|\hat{V} - V\|_N \geq 2 \|\tilde{V} - V\|_N \geq \|\tilde{V} - V\|_\rho - \epsilon_1, \quad (21)$$

where the first inequality results from truncation, and the second inequality is a generalization Lemma for Markov chains stated in Section D, and so

$$\|\tilde{V} - V\|_\rho \leq 2\|\hat{V} - V\|_N + \epsilon_1 \quad (22)$$

$$\leq \frac{2}{\sqrt{1-\gamma^2}} \|V - \hat{\Pi}V\|_N + \frac{2L}{(1-\gamma)^2} \sqrt{\frac{d}{v}} \left(\sqrt{\frac{2\log(2d/\delta')}{N}} + \frac{1}{N} \right) + \epsilon_1 \quad (23)$$

$$\leq \frac{2}{\sqrt{1-\gamma^2}} \|V - \Pi V\|_N + \frac{2L}{(1-\gamma)^2} \sqrt{\frac{d}{v}} \left(\sqrt{\frac{2\log(2d/\delta')}{N}} + \frac{1}{N} \right) + \epsilon_1 \quad (24)$$

$$\leq \frac{4\sqrt{2}}{\sqrt{1-\gamma^2}} \|V - \Pi V\|_\rho + \frac{2L}{(1-\gamma)^2} \sqrt{\frac{d}{v}} \left(\sqrt{\frac{2\log(2d/\delta')}{N}} + \frac{1}{N} \right) + \epsilon_1 + 2\sqrt{2}\epsilon_2 \quad (25)$$

$$\begin{aligned} &\leq \frac{4\sqrt{2}}{\sqrt{1-\gamma^2}} \left(\|V - \Pi^{\text{trunc}}V\|_\rho + \|\Pi^{\text{trunc}}V - \Pi V\|_\rho \right) \\ &\quad + \frac{2L}{(1-\gamma)^2} \sqrt{\frac{d}{v}} \left(\sqrt{\frac{2\log(2d/\delta')}{N}} + \frac{1}{N} \right) + \epsilon_1 + 2\sqrt{2}\epsilon_2 \end{aligned} \quad (26)$$

$$\begin{aligned} &\leq \frac{4\sqrt{2}}{\sqrt{1-\gamma^2}} \left(\|V - V^\epsilon(s, x_{0:T}, \bar{0})\|_\rho + \|\Pi^{\text{trunc}}V - \Pi V\|_\rho \right) \\ &\quad + \frac{2L}{(1-\gamma)^2} \sqrt{\frac{d}{v}} \left(\sqrt{\frac{2\log(2d/\delta')}{N}} + \frac{1}{N} \right) + \epsilon_1 + 2\sqrt{2}\epsilon_2 \end{aligned} \quad (27)$$

$$\begin{aligned} &\leq \frac{4\sqrt{2}}{\sqrt{1-\gamma^2}} \left(\|V - V^\epsilon\|_\rho + \|V^\epsilon - V^\epsilon(s, x_{0:T}, \bar{0})\|_\rho + \|\Pi^{\text{trunc}}V - \Pi V\|_\rho \right) \\ &\quad + \frac{2L}{(1-\gamma)^2} \sqrt{\frac{d}{v}} \left(\sqrt{\frac{2\log(2d/\delta')}{N}} + \frac{1}{N} \right) + \epsilon_1 + 2\sqrt{2}\epsilon_2 \end{aligned} \quad (28)$$

$$\begin{aligned} &\leq \frac{4\sqrt{2}}{\sqrt{1-\gamma^2}} \left(\frac{3}{(1-\gamma^2)} \sum_{t=T+1}^{\infty} (M_t + N_t) + \|\Pi^{\text{trunc}}V - \Pi V\|_\rho \right) \\ &\quad + \frac{2L}{(1-\gamma)^2} \sqrt{\frac{d}{v}} \left(\sqrt{\frac{2\log(2d/\delta')}{N}} + \frac{1}{N} \right) + \epsilon_1 + 2\sqrt{2}\epsilon_2. \end{aligned} \quad (29)$$

Inequality (23) is due to Lemma 9, (24) is by the definition of $\hat{\Pi}$, and (25) is the generalization bound similar to (21) in the opposite direction for upper bounding the empirical error in terms of the expected error. (26) is just the triangle inequality.

(27) is by definition of the Π^{trunc} operator, where V^ϵ is the ‘truncated’ value function of an ϵ -suboptimal policy that depends only on the past T events. This is guaranteed from the proof of Theorem 2 for $\epsilon = \frac{2}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t)$.

(28) is due to the triangle inequality by adding and subtracting the actual value function V^ϵ of the ϵ -suboptimal policy. The final inequality (29) is due to Lemma 4 and the value of ϵ .

Both these generalization bounds happen with probability at least $1 - \delta'$ each. Furthermore, with the lower bound ν on ν_N holding with probability at least $1 - \delta'$, the final bound holds with probability at least $1 - \delta = 1 - 4\delta'$. ■

The generalization terms due to linear function approximation using Markov samples remain essentially the same as in Lazaric et al. (2012). The main difference is the first extra term in the error, which is induced by truncating the augmented state space by discarding the features of old events. The amount of error introduced owing to this approximation depends on the amount of influence exerted by old events on the current events and the state transition kernel.

The results required for the proof can be taken from Lazaric et al. (2012) without any modifications because, even if some of the results, specifically the results on the bound on the covering numbers for linear function spaces taken from Györfi et al. (2002), require the domain of the basis functions to be finite-dimensional, the analysis takes place in the co-domain and can be directly applied to our setting. We state the complete theorem along with the Lemmas in Section D.

6.3 Sample Complexity of Least Squares Policy Iteration

To analyze Algorithm 1, we need to consider the following standard regularity conditions on the stationary distribution of greedy policies, future-state distribution of arbitrary sequences of policies, and linear independence of the features.

(C1) (Lower-bounding distribution) There exists a distribution ν and corresponding constant $C_\nu < \infty$ such that for any policy π that is greedy with respect to a function in the truncated space $\mathcal{F}$, $\nu \leq C_\nu \rho^\pi$, where ρ^π is the stationary distribution of policy π .

(C2) (Discounted-average concentrability of future-state distribution (Antos et al., 2008)) Given the target distribution ρ and an arbitrary sequence of policies $\{\pi_m\}_{m \geq 1}$, the concentrability coefficient

$$c_{\rho, \nu}(m) = \sup_{\pi_1 \dots \pi_m} \left\| \frac{d\nu P^{\pi_1} P^{\pi_2} \dots P^{\pi_m}}{d\rho} \right\|_\infty,$$

and the second-order discounted-average concentrability of future-state distributions, as defined below, is finite.

$$C_{\rho, \nu} = (1 - \gamma)^2 \sum_{m \geq 1} m \gamma^{m-1} c_{\rho, \nu}(m).$$

(C3) (Linearly independent features) Let ν be the lower-bounding distribution from Assumption (C1). The features $\phi(\cdot)$ of the function space $\mathcal{F}$ are linearly independent w.r.t. ν , and the smallest eigenvalue ω_ν of the Gram matrix $G_\nu \in \mathbb{R}^{d \times d}$ w.r.t. ν is strictly positive.

(C4) (Slower β -mixing process) There exists a stationary β -mixing process with parameters $\bar{\beta}, b, \kappa$, such that for any policy π that is greedy w.r.t. a function in the truncated space $\mathcal{F}^{(T)}$, it is slower than the stationary β -mixing process with stationary distribution ρ^π (with parameters $\bar{\beta}_\pi, b_\pi, \kappa_\pi$). This means that $\bar{\beta}$ is larger, and b and κ are smaller than their counterparts $\bar{\beta}_\pi, b_\pi$ and κ_π .

Theorem 11. *Under Assumptions (C1-C4), the value function of policy π_K obtained using Algorithm 2 for K iterations satisfies, with probability at least $1 - \delta$,*

$$\begin{aligned} \|V^* - V^{\pi_K}\|_\rho &\leq \frac{\sqrt{3}\gamma}{(1-\gamma)^2} \sqrt{C_{\rho,\nu}} \left[(1+\gamma) \sqrt{2C_\nu} \left(\frac{4\sqrt{2}}{\sqrt{1-\gamma^2}} \left(\frac{3}{(1-\gamma^2)} \sum_{t=T+1}^{\infty} (M_t + N_t) + \mathcal{E}_0(\mathcal{F}) \right) \right. \right. \\ &\quad \left. \left. + \frac{2L}{(1-\gamma)^2} \sqrt{\frac{d}{\nu}} \left(\sqrt{\frac{2\log(8d/\delta)}{N}} + \frac{1}{N} \right) + \mathcal{E}_1 + 2\sqrt{2}\mathcal{E}_2 \right) + \frac{2}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t) \right] \\ &\quad + \frac{\sqrt{6}\gamma^{K/2}}{(1-\gamma)^2}, \end{aligned}$$

where,

- $\mathcal{E}_0(\mathcal{F}) = \sup_{\pi \in \mathcal{G}(\mathcal{F})} \|\Pi^{\text{trunc}} V^\pi - \Pi V^\pi\|_{\rho^\pi}$
- $\mathcal{E}_1$ is ϵ_1 from Theorem 10 written for the slower β -mixing process defined in Assumption 4,
- $\mathcal{E}_2$ is ϵ_2 from Theorem 10 written for the slower β -mixing process defined in Assumption 4, and $\|\alpha^*\|$ replaced by $\sqrt{\frac{C}{\omega_\mu}} \frac{1}{1-\gamma}$, and
- ν_μ is ν from Lemma 14 in which ω is replaced by ω_ν defined in Assumption 3, and the second term is written for the slower β -mixing process defined in Assumption 4.

The error due to the finite history approximation of nonstationarity induced by external influence is captured by the first term and the second-to-last term in the bound, that is, $\sum_{t=T+1}^{\infty} (M_t + N_t)$. For instance, when we consider the external process as a discrete-time Hawkes process (see Section 3.1), this additional error due to external influence can be written as $\left(\frac{\bar{c}}{\lambda} e^{-\bar{\lambda}T} + \frac{c_\alpha}{\lambda_\alpha} e^{-\lambda_\alpha T} + \frac{c_\beta}{\sqrt{2\pi}\lambda_\beta} e^{-\lambda_\beta T} \right)$. One can notice that this term decays exponentially with T .

Proof of Theorem 11 For the sequence $\tilde{V}_k$ of approximate value functions of the policies π_k obtained by Algorithm 2, Lemma 15 provides the following recursion for the suboptimality of the policies.

$$\begin{aligned} V^* - V^{\pi_{k+1}} &\leq \gamma P^{\pi^*} (V^* - V^{\pi_k}) + \gamma E_k b_k + E'_k h_k, \\ \text{where } E_k &= P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} - P^{\pi^*} (I - \gamma P^{\pi_k})^{-1}, \\ E'_k &= \gamma P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} + I, \\ b_k &= \tilde{V}_k - T^{\pi_k} \tilde{V}_k, \\ \text{and } h_k &= T^{\pi_{k+1}} \tilde{V}_k - T^{\pi_{k+1}} \tilde{V}_k. \end{aligned}$$

Here, P^π is an operator that provides the expected next-step value function when following policy π , and is defined as $P^\pi V(\bar{s}) = \mathbb{E}_{s' \sim Q(\bar{s}, \pi(s))} V(s')$.

This bound on the difference between the optimal and current approximate value function has an extra term $E'_k h_k$ because our algorithm uses an approximate greedy policy that only depends on the current state and past T events instead of the true greedy policy, which can be a function of an infinite history of events. This is reflected in h_k , which is the

difference due to applying the $T^{\tilde{\pi}_{k+1}}$ on the previous value function $\tilde{V}_k$ using true greedy policy $\tilde{\pi}$, instead of $T^{\pi_{k+1}}$ with approximate greedy policy π_{k+1} .

By induction and then taking absolute value, we obtain

$$\begin{aligned} |V^* - V^{\tilde{\pi}_K}| &\leq \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} (\gamma F_k |b_k| + F'_k |h_k|) + (\gamma P^{\pi^*})^K |V^* - V^{\pi_0}|, \\ \text{where } F_k &= P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} + P^{\pi^*} (I - \gamma P^{\pi_k})^{-1} \\ \text{and } F'_k &= \gamma P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} + I. \end{aligned}$$

Using the fact that $V^* - V^{\pi_0} \leq \frac{2}{1-\gamma} R_{\max} \mathbf{1}$, where $R_{\max}$ is the maximum possible reward in any transition, one can rewrite the above bound as

$$|V^* - V^{\pi_K}| \leq \frac{1 + 2\gamma + \gamma^K - 4\gamma^{K+1}}{(1-\gamma)^2} \left[\sum_{k=0}^{K-1} (\alpha_k A_k |b_k| + \beta_k B_k |h_k|) + \alpha_K A_K R_{\max} \mathbf{1} \right],$$

where A_k and B_k represent the operators defined as

$$\begin{aligned} A_k &= \begin{cases} \frac{1-\gamma}{2} (P^{\pi^*})^{K-k-1} F_k, & 0 \leq k < K \\ (P^{\pi^*})^K & k = K. \end{cases} \\ B_k &= (1-\gamma) (P^{\pi^*})^{K-k-1} F'_k, \end{aligned}$$

and corresponding coefficients

$$\begin{aligned} \alpha_k &= \begin{cases} \frac{2(1-\gamma)\gamma^{K-k}}{1+2\gamma+\gamma^K-4\gamma^{K+1}}, & 0 \leq k < K, \\ \frac{2(1-\gamma)\gamma^K}{1+2\gamma+\gamma^K-4\gamma^{K+1}}, & k = K, \end{cases} \\ \beta_k &= \frac{(1-\gamma)\gamma^{K-k-1}}{1+2\gamma+\gamma^K-4\gamma^{K+1}}. \end{aligned}$$

The operators A_k and B_k are positive, that is, $A_k V \geq 0$ and $B_k V \geq 0$ whenever $V \geq 0$, and leave unity invariance, that is, $A_k \mathbf{1} = \mathbf{1}$ and $B_k \mathbf{1} = \mathbf{1}$, and the corresponding coefficients sum to 1. Therefore, after raising both sides of the inequality to the power p and integrating with respect to an arbitrary distribution ρ , we can use the Jensen inequality to obtain

$$\|V^* - V^{\pi_K}\|_{p,\rho}^p \leq \lambda_K \cdot \rho \left[\sum_{k=0}^{K-1} (\alpha_k A_k |b_k|^p + \beta_k B_k |h_k|^p) + \alpha_K A_K R_{\max}^p \mathbf{1} \right],$$

where the $\rho \cdot$ operator gives the expectation with respect to ρ , i.e., $\rho V = \mathbb{E}_{\tilde{s} \sim \rho} V(\tilde{s})$, and $\lambda_K = \left(\frac{1+2\gamma+\gamma^K-4\gamma^{K+1}}{(1-\gamma)^2} \right)^p$. ■

From definitions of coefficients $c_{\rho,v}(m)$,

$$\rho A_k \leq (1-\gamma) \sum_{m \geq 0} \gamma^m c_{\rho,v}(m+K-k)v, \text{ and}$$

$$\rho B_k \leq (1 - \gamma) \sum_{m \geq 0} \gamma^m c_{\rho, v}(m + K - k - 1) v.$$

This gives us

$$\begin{aligned} \|V^* - V^{\pi_K}\|_{p, \rho}^p &\leq \lambda_K \left[\sum_{k=0}^{K-1} \left(\alpha_k (1 - \gamma) \sum_{m \geq 0} \gamma^m c_{\rho, v}(m + K - k) \|b_k\|_{p, v}^p \right. \right. \\ &\quad \left. \left. + \beta_k (1 - \gamma) \sum_{m \geq 0} \gamma^m c_{\rho, v}(m + K - k - 1) \|h_k\|_{p, v}^p \right) + \alpha_K R_{\max}^p \right]. \end{aligned}$$

$$\text{Let } \epsilon_1 \stackrel{\text{def}}{=} \max_{0 \leq k < K} \|b_k\|_{p, v} = \max_{0 \leq k < K} \|\tilde{V}_k - T^{\pi_K} \tilde{V}_k\|_{p, v}, \text{ and}$$

$$\epsilon_2 \stackrel{\text{def}}{=} \frac{2R_{\max}}{1 - \gamma} \sum_{t=T+1}^{\infty} (M_t + N_t).$$

Here, ϵ_1 is the maximum Bellman error for all value function approximations over all iterations, and ϵ_2 is the approximation error term due to the use of a truncated and finite representation of the augmented state space, which consists of just the current state and events in the past T time steps. Using Lemma 16, $|h_k|$ can be bounded and, one can write,

$$\begin{aligned} \|V^* - V^{\pi_K}\|_{p, \rho}^p &\leq \lambda_K \left[\frac{2\gamma}{1 + 2\gamma + \gamma^K - 4\gamma^{K+1}} C_{\rho, v} \epsilon_1^p + \frac{\gamma}{1 + 2\gamma + \gamma^K - 4\gamma^{K+1}} C_{\rho, v} \epsilon_2^p \right. \\ &\quad \left. + \frac{2(1 - \gamma)\gamma^K}{1 + 2\gamma + \gamma^K - 4\gamma^{K+1}} R_{\max}^p \right] \\ &\leq \left(\frac{1}{(1 - \gamma)^2} \right)^p \left[\gamma(1 + 2\gamma + \gamma^K - 4\gamma^{K+1})^{p-1} C_{\rho, v} (2\epsilon_1^p + \epsilon_2^p) \right. \\ &\quad \left. + 2(1 - \gamma)\gamma^K (1 + 2\gamma + \gamma^K - 4\gamma^{K+1})^{p-1} R_{\max}^p \right]. \end{aligned}$$

Noting that $1 + 2\gamma + \gamma^K - 4\gamma^{K+1} < 3$, and considering the l_2 norm by setting $p = 2$, we obtain

$$\begin{aligned} \|V^* - V^{\pi_K}\|_{\rho} &\leq \frac{\sqrt{3\gamma}}{(1 - \gamma)^2} \sqrt{C_{\rho, v}} (\sqrt{2}\epsilon_1 + \epsilon_2) + \frac{\sqrt{6}\gamma^{K/2}}{(1 - \gamma)^2} R_{\max} \\ &= \frac{\sqrt{3\gamma}}{(1 - \gamma)^2} \sqrt{C_{\rho, v}} \left(\sqrt{2} \max_{0 \leq k < K} \|V_k - T^{\hat{\pi}_k} V_k\|_v + \frac{2R_{\max}}{1 - \gamma} \sum_{t=T+1}^{\infty} (M_t + N_t) \right) \\ &\quad + \frac{\sqrt{6}\gamma^{K/2}}{(1 - \gamma)^2} R_{\max}. \end{aligned}$$

Based on Assumption (C1) of ν being the lower bounding distribution of all ρ_k with corresponding constant C_ν , we have

$$\begin{aligned} \|V^* - V^{\pi_K}\|_\rho &\leq \frac{\sqrt{3\gamma}}{(1-\gamma)^2} \sqrt{C_{\rho,\nu}} \left(\sqrt{2C_\nu} \max_{0 \leq k < K} \|V_k - T^{\pi_k} V_k\|_{\rho_k} + \frac{2R_{\max}}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t) \right) \\ &\quad + \frac{\sqrt{6}\gamma^{K/2}}{(1-\gamma)^2} R_{\max}. \end{aligned}$$

By applying Lemma 9 of Lazaric et al. (2012) to the approximate greedy policy π_k , one can bound the Bellman error using the evaluation error as

$$\begin{aligned} \|V^* - V^{\pi_K}\|_\rho &\leq \frac{\sqrt{3\gamma}}{(1-\gamma)^2} \sqrt{C_{\rho,\nu}} \left((1+\gamma) \sqrt{2C_\nu} \max_{0 \leq k < K} \|V_k - V^{\pi_k}\|_{\rho_k} + \frac{2R_{\max}}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t) \right) \\ &\quad + \frac{\sqrt{6}\gamma^{K/2}}{(1-\gamma)^2} R_{\max}. \end{aligned}$$

The first term, which is the maximum evaluation error over K iterations, can be bounded using Theorem 10 to obtain the final result.

7 Experiments

7.1 Policy Evaluation

We conducted experiments to analyze policy evaluation in a nonstationary variant of the classic Pendulum-v1 environment within the OpenAI Gym control suite. The task involves applying torque to a pendulum to swing it and keep it upright. We utilized a fixed neural network policy that was partially pretrained using DDPG (Lillicrap et al., 2016). Subsequently, we employed pathwise LSTD with linear function approximation to evaluate this policy, as described in Section 6. The features considered are the standard cosine and sine of the angle and the angular velocity, with additional features being the angle itself, along with the squares of all these features, resulting in an 8-dimensional state.

To address nonstationarity, we consider the discrete-time Gaussian-marked Hawkes process outlined in Section 3.1. The intensity resulting from due to an event decays as $\alpha = e^{-\lambda_\alpha t}$ with $\lambda_\alpha = 1.0$, and the effect on the event mark decays as $\beta_t = 1/(1+t^2)$. The events were added to the torque applied to the pendulum. Figure 2 shows the expected error of the learned value function as a function of the number of samples, event horizon T , and rate λ_α of decay of the Hawkes process.

For a fixed event horizon 5, the error decreases as the number of samples increases, which is intuitive. For a fixed number of samples 10,000, the average error initially decreases as the event horizon increases, corresponding to the first term in the sample complexity of the total variation induced by events older than the event horizon. After a certain stage, increasing the event horizon results in a gradual increase in error. This is because events older than the horizon now contribute a negligible influence on the current dynamics, while the dimensionality of the features increases owing to the inclusion of more past events in the features, resulting in a slightly greater expected error.

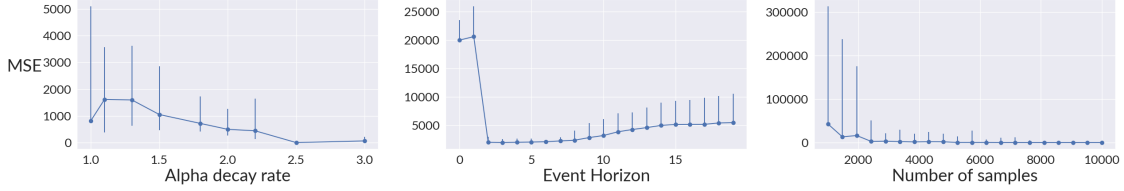


Figure 2: Performance of pathwise LSTD. The plots show the dependence of the Mean Squared Error of the value function learned using pathwise LSTD as a function of the rate of decay of the exogenous Hawkes process, the horizon of past events considered, and the number of samples used.

The first plot shows the dependence on λ_α , the rate of decay of α_t . Faster decay results in less nonstationarity and a lower approximation error owing to the first term in the upper bound of the sample complexity.

We conducted 20 trials and plotted the median of the Mean Squared Errors for the experiments with respect to the decay rate and number of samples. The error bars indicate the range of 40-60 percentile versus the decay rate and 20-80 percentile for the other two.

These results empirically demonstrate the sample complexity bound for policy evaluation with linear function approximation in Theorem 10. The average error decreases with increasing N and decreasing N_t .

Increasing the event horizon T increases d while reducing $\sum_{t=T+1}^{\infty} \{M_t, N_t\}$, resulting in a trade-off between the corresponding two terms in the bound, with the minimal error being achieved for $T = 2$.

We conducted experiments for policy improvement using the car racing environment of the gymnasium library. We modified the environment to make it nonstationary. The external event in this case was the occurrence of rain, which reduced road friction. This reduction in road friction is temporary, and the friction increases exponentially over time back to the original value as the road dries. The agent should learn to account for the reduced friction, especially in the corners, and reduce its speed. Otherwise, the agent goes off the course. In a car racing scenario, where even a slight mistake will end up in losing the race, the agent needs to drive at the right speed to lead in the race while ensuring that the car does not skid.

We trained the original environment with the maximum number of steps set to 800. The agent was trained almost perfectly, completing the lap without going off course. This score serves as the upper bound for the nonstationary case. Next, we trained an agent for the nonstationary case, where the agent had no information about external events. As expected, the agent's performance was not as good as that in the stationary case. Because the agent has no information about the external event, it has learned to navigate the course at a lower speed to avoid going off course. We then experimented by providing the agent with information regarding the past T events for $T=10, 15$, and 25 . With the additional information of T , the agent was able to learn a better policy. As shown in Figure 3, as

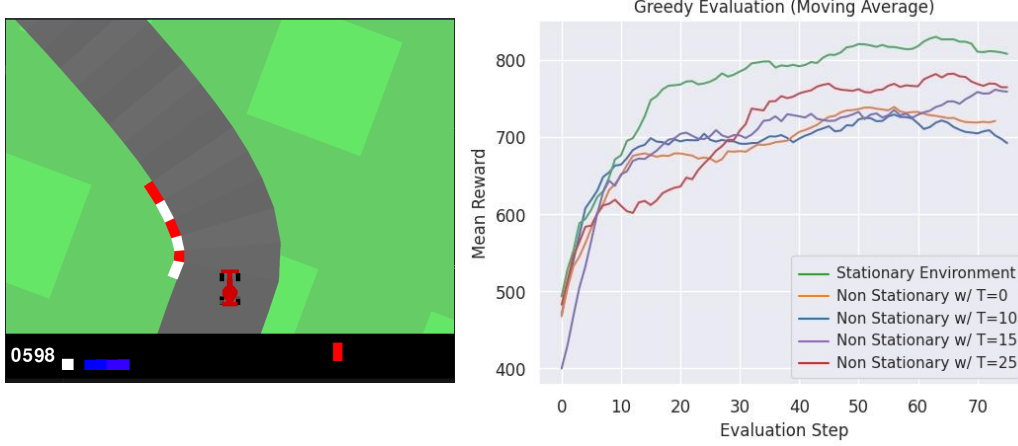


Figure 3: Left: Car racing environment in action. The friction between the road and car tires changes due to rain, which acts as an external event. The road friction will increase to the original value over time. Right: Policy improvement for nonstationary environment with different time horizons T .

we increase horizon T , the agent can learn a better policy. However, the effect of passing information related to external events decreases as T increases, as the effect of older events diminishes exponentially.

7.2 Policy Deployment

In this paper, we considered a setting in which the environment containing an agent is perturbed owing to the presence of exogenous factors. Suppose a trained policy is available to the agent that is optimal in the absence of these external influences. Now, we wish to deploy this agent in a nonstationary environment where these external events affect the MDP dynamics. In this case, it would be more efficient to modify the existing policy to deal with nonstationarity directly than to train a new policy from the ground up.

We consider a simple model-based planning strategy, wherein the agent learns the dynamics model and uses it to simulate a few trajectories based on the current policy. More precisely, given a pretrained policy π , when at an augmented state $\bar{s} \in \mathcal{S}$ during deployment, d actions $a_1, \dots, a_d$ are sampled by perturbing the action $a = \pi(\bar{s})$. From each action a_i , a trajectory $(s, a_i, s_{i,1}, a_{i,1}, \dots, s_{i,H}, a_{i,H})$ is obtained, and its corresponding sum of rewards $r_i = \sum_{t=1}^H r(s_{i,t}, a_{i,t})$ is calculated. The action a_i that gives rise to the highest reward r_i is chosen and taken in the environment. This process was repeated at each time step during the deployment.

This procedure can be performed in two different ways. A naive way to do this is to learn a dynamics model $\hat{T} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta\mathcal{S}$ directly in the augmented state space and use it to simulate the trajectories. However, a more intelligent method is to learn the event process separately, and use it to simulate events $e_{i,t}$ that are fed into a dynamic

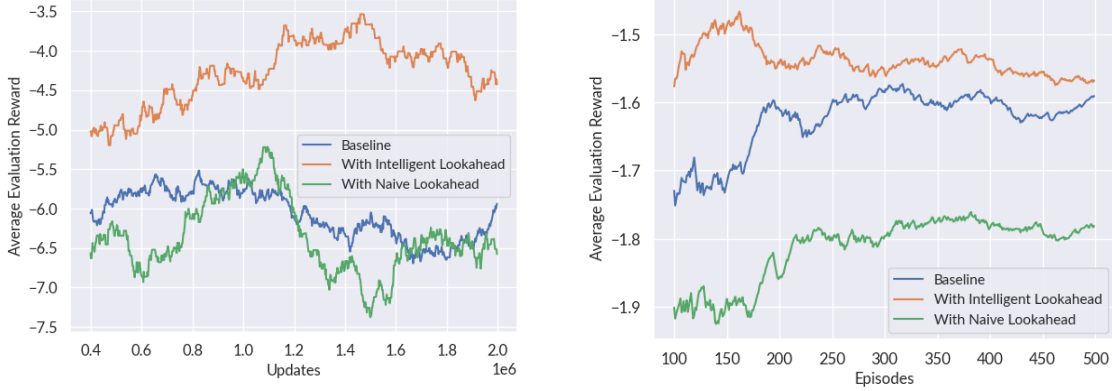


Figure 4: Results on nonstationary Pendulum and Point Maze environments. Learning state dynamics and exogenous events separately and planning intelligently outperforms a stationary policy or a model-based policy that tries to learn the dynamics model in the augmented state space directly.

model $\bar{T} : \bar{\mathcal{S}} \times \mathcal{A} \rightarrow \mathcal{S}$ that only predicts the states. Figure 4 depicts the results obtained using these two methods, along with those of a baseline that is the stationary policy learned in the stationary version of the environment. The first plot shows the results of the nonstationary pendulum task described previously. The second plot is for a nonstationary version of the Gymnasium Point Maze environment (Fu et al., 2020), in which a 2DoF ball is force-actuated in the Cartesian x - y directions to reach a randomly specified target goal in a closed, continuous 2D maze. Nonstationarity is induced by independent Hawkes processes, as described previously, that occur at 5 randomly sampled and fixed points in the maze, each of which exerts a pulling force on the agent towards itself that is inversely proportional to the squared distance. From the plots, it is evident that an intelligent model-based look-ahead policy that mimics the structure of the nonstationarity performs better than naive baselines.

The algorithms theoretically analyzed in the previous sections are extensions of standard reinforcement learning algorithms that operate in the augmented state space. This is a valid strategy in many cases, as observed in the policy evaluation experiments described earlier. However, such a strategy ignores the structure of nonstationarity, sacrificing performance for generality. When there is additional knowledge on the nature of the exogenous event process, a valid question is whether using this knowledge helps solve the problem better. This experiment showed that, for the case of deploying a stationary policy in a nonstationary environment using model-based planning, considering the nature of the event process and incorporating it intelligently with current algorithms will lead to better results than ignoring the structure of the nonstationarity and solely operating in the augmented state space.

This discrepancy is due to the nature of the neural network function approximation primarily used in practical scenarios. The approximate policy improvement step in Algo-

rithm 1 attempts to maximize the one-step returns obtained from each augmented state $\bar{s}$, which requires knowledge of the dynamics model. When this is unknown, an approximate model is used, which can be constructed in two ways, as described in the previous sections. Standard neural network dynamics models used in the literature are ill-equipped to model event processes, compelling one to model them separately from the original state $s \in \mathcal{S}$ when possible. Further research is needed on how to precisely model and incorporate event processes in reinforcement learning policies in a way that utilizes the structure of nonstationarity and the nature of the induced perturbations.

8 Related Work

Numerous prior studies addressing nonstationary RL problems predominantly offer practical solutions that lack theoretical backing. In this context, we refrain from discussing such works and instead aim to position our research within the framework of existing theoretical findings.

Lecarpentier and Rachelson (2019) considers a finite horizon MDP where the transition dynamics and reward function change in a manner that is Lipschitz continuous over time. Cheung et al. (2020) analyzes a similar scenario within a finite MDP, where the total variation in the environment, termed the variation budget, is known to the agent. This work proposes an optimistic value iteration-based algorithm that guarantees an upper bound on its dynamic regret. Wei and Luo (2021) introduces a black-box reduction that transforms any appropriate algorithm with optimal performance in a near-stationary environment into an algorithm that achieves optimal dynamic regret in a nonstationary environment. Guo et al. (2024) investigates average reward nonstationary MDPs in Borel spaces and proposes a rolling-horizon algorithm to obtain arbitrarily almost-optimal solutions based on a suitably changing horizon. At each stage, a finite horizon problem starting from that stage is solved to obtain the decision rule. In contrast, our algorithm considers a horizon of past external events instead of choosing the current action. Furthermore, along with standard regularity conditions of compactness, continuity, etc., Condition 1 in Guo et al. (2024) contains global conditions on the sequence of transition kernels at each stage of the MDP to guarantee the existence of optimal policies and solutions to the Average Optimality Equation. In contrast, our assumptions (A2) and (A3) are more local in nature, with bounds on the perturbations of the transition kernel.

Hadoux et al. (2014) introduced a class of piecewise stationary MDPs known as Hidden-Semi-Markov-Model MDPs, characterized by semi-Markov transitions between hidden stationary MDP models. To solve this problem, this work adapts the partially observable Monte Carlo planning algorithm used for solving partially observable MDPs.

Feng et al. (2022) propose factored nonstationary MDPs, where the transition dynamics are defined in terms of a dynamic Bayesian network over the states, actions, rewards, and latent change factors that induce nonstationarity. Here, both the states and the latent change factors have transition dynamics obeying this causal graph, and the latent factors evolve in a Markovian fashion. This work suggests employing variational autoencoders to learn the transition dynamics.

A factorization setting called exogenous MDP is presented in (Efroni et al., 2022), wherein the state is divided into two parts: an endogenous state, which is controllable by

the agent’s actions, and an exogenous state, which evolves on its own and does not affect the agent’s reward. The precise manner in which this division exists is unknown to the agent. While this may appear similar to our interpretation of states and external events, in our scenario, external events affect the dynamics of the state and, consequently, the rewards received by the agent. It is important to note that this work only considered the tabular setting and proposed algorithms that can only deal with finite horizon episodes. A similar decomposition of the state space into endogenous and exogenous subspaces was studied by Dietterich et al. (2018). In this study, it was posited that the exogenous states evolve in a Markov fashion and that the reward function can be additively decomposed into exogenous and endogenous rewards. This study proposed algorithms to determine the exogenous component of the state as projections that maximize the partial correlation coefficient as a proxy for the mutual information between the exogenous next state and the endogenous current state and action conditioned on the exogenous current state.

While previous studies, such as Efroni et al. (2022); Dietterich et al. (2018), focused on the decomposition of states into endogenous and exogenous components, our work addresses a fundamentally different problem: the impact of integrating exogenous event information into reinforcement learning.

Changing dynamics of environments based on externally specified “contexts” have also been studied under Contextual MDPs (Hallak et al., 2015). However, these studies considered contexts that are constant for each episode and have a finite number of possible values. The aforementioned study proposed the CECE algorithm, which learns a set of policies corresponding to all contexts and determines the latent context for an episode (if unknown) to choose the policy corresponding to the current context.

More specifically, the CECE clusters an initial set of trajectories into groups corresponding to a fixed set of latent contexts. For each new episode, a partial realization is used to classify it into one of these clusters, and a model learned on that cluster is used to exploit the rest of the trajectory. This algorithm cannot handle the setting considered in this study because each new time step results in a change in the context, and knowledge of a latent context for the current time step does not guarantee the same context for the rest of the episode, except in a trivial special case.

Modi et al. (2018) consider a relatively closer setting to ours, where the context is known and possibly continuous, with their algorithm being more amenable to being adapted to handle dynamic contexts. However, it works by keeping count of the number of occurrences of (s, a) in each of a set of representative contexts to learn a set of explicit models, which is only possible in the setting of finite states and action spaces.

Tennenholtz et al. (2023) extend Contextual MDPs to handle dynamic contexts that are known to the agent, change at each time step, and influence the state transition distribution. However, their work differs from ours in several significant aspects. This study considers the contexts at each time step to depend on the history of states, actions, and contexts, whereas the next state depends only on the current state, action, and context. This means that causality exists in both directions, from states to contexts and vice versa. In contrast, in our work, the contexts are exogenous events that cannot be controlled by the agent, and the causality flows only in one direction, from contexts to states. Furthermore, their work considers finite state spaces, with the analysis explicitly geared towards such spaces and bounds that depend on the size of the state and action sets.

In Tennenholtz et al. (2023), the context space is a fixed finite set of M vectors, and a latent map is learned from the (state, action, context) space to $\mathbb{R}^M$, with the range of the map still being a finite, albeit exponentially increasing, set. This study also considers a finite-horizon MDP. This relatively simple setting allows one to tractably find the latent map and learn an optimal policy. In this sense, our setting is much more general and challenging. In addition, the work of Tennenholtz et al. (2023) presents regret analysis, whereas our study focuses on the convergence results of algorithms.

Our analysis addresses the nonstationarity induced in the MDP by the influence of non-Markovian events, necessitating the learning of policies that are contingent upon both the current state and a history of prior events. Although this approach bears resemblance to the strategies employed in Partially Observable MDPs (POMDPs) (Sunberg and Kochenderfer, 2018), the underlying rationale diverges significantly. In POMDPs, histories of observations are utilized despite the Markovian nature of state transition and observation functions, primarily to mitigate an agent’s inability to perceive the true state. Conversely, in our framework, both the state and external events are fully observable, and the necessity for an augmented space that incorporates event histories arises from the intrinsic non-Markovian characteristics of a nonstationary environment. Moreover, the causal framework of our scenario is distinctly different. The events exhibit non-Markovian properties, and the state at any given time directly affects the distribution of the subsequent state, regardless of the knowledge of an event. In POMDPs, while beliefs and sufficient statistics are contingent upon a history of observations, the states themselves maintain a Markovian nature, leading to observations that are causally independent.

Our work makes significant inroads into understanding how the characteristics of exogenous event processes affect the tractability of the problem and the convergence and sample complexity of RL algorithms. To the best of our knowledge, no existing studies in the literature offer these findings.

9 Outlook

This study lays the groundwork for understanding sequential decision-making problems under the influence of external temporal events, offering several theoretical insights within the reinforcement learning framework. The results are established in a general setting that relies on learning functions in an augmented state space. The challenge here is that the augmented state space can be exponentially larger due to the long-term dependencies of external events; hence, it is essential to develop provable approximate methods to mitigate this.

The “factorization” of the environment into Markovian states and non-Markovian events that impact these states, compressing the histories of events into latent features that are independent of the states, may provide a strategic advantage. This approach could facilitate the learning of representations of the expanded states by developing a distinct event distribution model that can be scaled up, as it does not rely on the agent, along with a separate state transition model that necessitates agent interaction with the environment. A compressed representation of events derived from extensive event data can be used to learn a feasible state transition kernel. This can be achieved either implicitly through function approximators, such as recurrent neural networks (or LSTMs), or explicitly by learning

compressed state representations using mutual information bottlenecks. Although these are practical methods, the theoretical analysis of such learned compressed state representations remains an open problem.

In this work, we used an augmented state representation obtained by considering the current state and a fixed time horizon T of past events. The choice of T is a hyperparameter that controls the extent to which information about the environment is preserved, with larger values of T resulting in better approximation and thereby better performance of the corresponding algorithms, as described in Theorems 7, 10 and 11. Although this is a fixed choice made by the algorithm designer, the next logical step would be to learn the optimal value of T using the data collected from the environment and adapt it based on the performance of the agent. We leave this as a future research problem.

The analysis of sample complexity in Section 6 hinges on generalizing the sample error of policy evaluation to an expected error over the entire state space. Such generalizations typically necessitate assumptions regarding the distribution of states. In our results, the assumption pertains to the rate of convergence of the Markov chain induced from the augmented MDP by the policy under evaluation, a notion frequently encountered in the reinforcement learning literature. However, in our case, the specific structure of the state space, specifically, the separation of the augmented state space into states and events with different characteristics, may necessitate alternate assumptions. For instance, while the recurrence and aperiodicity of the Markov chain may not hold, it may be reasonable to assume stationarity solely for the event sequences. Therefore, the generalizability of the value function in our setting can be analyzed in a more specific way that considers this exogenous structure.

The policy improvement step involves optimizing the action space in accordance with the expected one-step reward and value function for the subsequent time step, necessitating an understanding of the model structure. Therefore, when faced with an unknown transition and reward model, it is crucial to explore the sample complexity associated with learning these models from samples and the impact this has on the effectiveness of the resulting policy.

Appendix A. List of notations and abbreviations

Standard mathematical notations are used.

Notation	Description
$\mathbb{N}$	Set of natural numbers, $\{1, 2, 3, \dots\}$.
$\mathbb{R}$	Set of real numbers.
$ \mathcal{X} $	Cardinality of set $\mathcal{X}$.
$\inf$	Infimum of a given set
$\sup$	Supremum of a given set
$\ \cdot\ _2$	l_2 norm of a vector.
$[N]$	The set $\{1, 2, \dots, N\}$ for some $N \in \mathbb{N}$.
$\mathbb{E}[\cdot]$	Expectation with respect to a certain distribution.
$\mathbb{P}(\cdot)$	Probability of a given event with respect to some distribution.

$\Delta(\cdot)$	Class of probability distributions over a given set
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2

Notation related to reinforcement learning.

Notation	Description
$\mathcal{M}$	A Markov Decision Process (MDP)
$\mathcal{S}$	State space of an MDP
$\mathcal{A}$	Action space of an MDP
r	Reward function
Q	Transition kernel of an MDP
γ	Discount factor
π	A policy that acts in an MDP
π^*	Optimal policy
V^π	Value function of policy π
V^*	Optimal value function
Q^π	Action-value function of policy π
Q^*	Optimal Q function
$\mathcal{T}$	Bellman operator

Abbreviations.

Notation	Description
MDP	Markov Decision Process
PPO	Proximal Policy Optimization
DOF	Degrees Of Freedom

Appendix B. Proof of inequality 2

First we show that $\text{TV}(\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)) = \text{erf}\left(\frac{|\mu_2 - \mu_1|}{2\sqrt{2}}\right)$. Without loss of generality, we assume $\mu_2 \geq \mu_1$. Now,

$$\begin{aligned}
 \text{TV}(\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)) &= \frac{1}{2} \int_{-\infty}^{\infty} \left| \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_1)^2}{2}\right) - \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_2)^2}{2}\right) \right| dx \\
 &= \int_{\frac{\mu_1 + \mu_2}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} \left(\exp\left(-\frac{(x - \mu_2)^2}{2}\right) - \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_1)^2}{2}\right) \right) dx \\
 &= \frac{1}{\sqrt{2\pi}} \left(\int_{\frac{\mu_1 - \mu_2}{2\sqrt{2}}}^{\infty} \sqrt{2} \exp(-y^2) dy - \int_{\frac{\mu_2 - \mu_1}{2\sqrt{2}}}^{\infty} \sqrt{2} \exp(-z^2) dz \right) \\
 &= \frac{2\sqrt{2}}{\sqrt{2\pi}} \int_0^{\frac{\mu_2 - \mu_1}{2\sqrt{2}}} \exp(-y^2) dy = \text{erf}\left(\frac{|\mu_1 - \mu_2|}{2\sqrt{2}}\right) \tag{30}
 \end{aligned}$$

Now, we proceed to show the steps involved in establishing (2). Given two historical sequences $(X_{t'})_{t' < t}$ and $(X'_{t'})_{t' < t}$ that only differ by one event at some time $t' \leq t - T$ with $B'_{t'} = X'_{t'} = 0$ and $B_{t'} = 1$, $X_{t'} = x$, the difference in intensities is $p_t - p'_t = \alpha_{t-t'}$

Let q_1 and q_2 be the density function of X_t and X'_t

$$\begin{aligned}
 \text{TV}(X_t, X'_t) &= \frac{1}{2} \int_{-\infty}^{\infty} |q_1(x) - q_2(x)| dx \\
 &= \frac{1}{2} \int_{-\infty}^{\infty} \left| q_1(x | E_t = 0) \Pr(E_t = 0) + q_1(x | E_t = 1) \Pr(E_t = 1) \right. \\
 &\quad \left. - q_2(x | E'_t = 0) \Pr(E'_t = 0) - q_2(x | E'_t = 1) \Pr(E'_t = 1) \right| dx \\
 &\leq \frac{1}{2} \int_{-\infty}^{\infty} |q_1(x | E_t = 0)(1 - p_t) - q_2(x | E'_t = 0)(1 - p'_t)| dx \\
 &\quad + \frac{1}{2} \int_{-\infty}^{\infty} |q_1(x | E_t = 1)p_t - q_2(x | E'_t = 1)p'_t| dx \\
 &\leq \frac{1}{2} \int_{-\infty}^{\infty} |\delta(x)(1 - p_t) - \delta(x)(1 - p'_t)| dx \\
 &\quad + \frac{p_t}{2} \int_{-\infty}^{\infty} |q_1(x | E_t = 1) - q_2(x | E'_t = 1)| dx \\
 &= \frac{|p'_t - p_t|}{2} \int_{-\infty}^{\infty} \delta(x) dx + p_t \text{TV}(\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)) \\
 &= \frac{\alpha_{t-t'}}{2} + p_t \text{TV}(\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)) \\
 &= \frac{\alpha_{t-t'}}{2} + p_t \text{erf}\left(\frac{|\mu_1 - \mu_2|}{2\sqrt{2}}\right) = \frac{\alpha_{t-t'}}{2} + p_t \text{erf}\left(\frac{|\beta_{t-t'} x|}{2\sqrt{2}}\right) \\
 &\leq \frac{\alpha_T}{2} + \text{erf}\left(\frac{\beta_T b}{2\sqrt{2}}\right) = N_T.
 \end{aligned}$$

Appendix C. Proof of Lemma 3

The aim is to study the value function of the policy in two MDPs $\mathcal{M}_X$ and $\mathcal{M}_X^{(T)}$, where $\mathcal{M}_X$ is the nonstationary MDP with augmented states consisting of the actual state and the entire history of events until then, and $\mathcal{M}_X^{(T)}$ is an approximate MDP where only events in the past T time steps affect the state transition and event distribution. For notational convenience, one can extend the state space of $\mathcal{M}_X^{(T)}$ to that of $\mathcal{M}_X$ by including the information of every past event in the state, without changing the transition function. Define a class of auxiliary MDPs $\mathcal{M}_h^{(T)}$, $h \in \{0, 1, 2, \dots\}$ (similar to that of Theorem 1 from Xiao et al. (2019)) such that when starting from the initial (augmented) state $\bar{s} \in \bar{\mathcal{S}}$, the transitions follow the transition kernel of $\mathcal{M}_X$ for h time steps and then the transition kernel of $\mathcal{M}_X^{(T)}$ for all transitions that follow. The value functions induced by π in $\mathcal{M}_X$ and $\mathcal{M}_X^{(T)}$ can be related using these new intermediate MDPs, which interpolate between the two environments.

Since only a single policy is under discussion, the π in V^π is omitted for the rest of this proof for the sake of simplicity. V denotes the value function of the policy in $\mathcal{M}_X$, $V^{(T)}$ denotes its value in $\mathcal{M}_X^{(T)}$, and $V_h^{(T)}$ denotes its value in $\mathcal{M}_h^{(T)}$.

From the definition of $\mathcal{M}_h^{(T)}$ and the boundedness of r , we have

$$V^{(T)} = V_0^{(T)}, \text{ and } V = \lim_{h \rightarrow \infty} V_h^{(T)},$$

and hence,

$$\begin{aligned} V - V^{(T)} &= \lim_{H \rightarrow \infty} V_H^{(T)} - V_0 = \lim_{H \rightarrow \infty} (V_H^{(T)} - V_0) = \lim_{H \rightarrow \infty} \sum_{h=0}^{H-1} (V_{h+1}^{(T)} - V_h^{(T)}) \\ &= \sum_{h=0}^{\infty} (V_{h+1}^{(T)} - V_h^{(T)}). \end{aligned}$$

That is, the difference between V and $V^{(T)}$ can be characterized in terms of the differences between $V_h^{(T)}$ and $V_{h+1}^{(T)}$ for each $h \in \{0, 1, 2, \dots\}$. Now, for any augmented state $\bar{s} \in \bar{S}$,

$$\begin{aligned} V_h^{(T)}(\bar{s}) &= \sum_{t=0}^{h-1} \gamma^t \mathbb{E}_{\substack{\bar{s}_t \sim p_t(\cdot|\bar{s}) \\ a_t \sim \pi(\cdot|\bar{s}_t) \\ \bar{s}_{t+1} \sim p(\cdot|\bar{s}_t, a_t)}} [r(\bar{s}_t, a_t, \bar{s}_{t+1})] + \gamma^h \mathbb{E}_{\substack{\bar{s}_h \sim p_h(\cdot|\bar{s}) \\ a_h \sim \pi(\cdot|\bar{s}_h) \\ \bar{s}_{h+1} \sim p^{(T)}(\cdot|\bar{s}_h, a_h)}} [r(\bar{s}_h, a_h, \bar{s}_{h+1})] \\ &\quad + \sum_{t=h+1}^{\infty} \gamma^t \mathbb{E}_{\substack{\bar{s}_t \sim p_{t-h}^{(T)} \circ p_h(\cdot|\bar{s}) \\ a_t \sim \pi(\cdot|\bar{s}_t) \\ \bar{s}_{t+1} \sim p^{(T)}(\cdot|\bar{s}_t, a_t)}} [r(\bar{s}_t, a_t, \bar{s}_{t+1})], \end{aligned}$$

where $p_t(\cdot|\bar{s})$ denotes a transition of t steps in the environment $\mathcal{M}_X$ starting from $\bar{s}$, a superscript $^{(T)}$ denotes the corresponding object in $\mathcal{M}_X^{(T)}$, and the policy π is implicit whenever it is not explicitly mentioned in any expression. The reward function does not depend on the external events, so $r(\bar{s}_t, a_t, \bar{s}_{t+1})$ is just $r(s_t, a_t, s_{t+1})$, where $\bar{s} = (s, x)$ and s is the actual state.

This can also be written as

$$V_h^{(T)}(\bar{s}) = \sum_{t=0}^{h-1} \gamma^t \mathbb{E}_{\substack{\bar{s}_t \sim p_t(\cdot|\bar{s}) \\ a_t \sim \pi(\cdot|\bar{s}_t) \\ \bar{s}_{t+1} \sim p(\cdot|\bar{s}_t, a_t)}} [r(\bar{s}_t, a_t, \bar{s}_{t+1})] + \gamma^h \mathbb{E}_{\bar{s}_h \sim p_h(\cdot|\bar{s})} V^{(T)}(\bar{s}_h).$$

Using these two representations for $V_h^{(T)}$ and $V_{h+1}^{(T)}$ respectively gives

$$\begin{aligned} V_{h+1}^{(T)}(\bar{s}) - V_h^{(T)}(\bar{s}) &= \gamma^h \mathbb{E}_{\substack{\bar{s}_h \sim p_h(\cdot|\bar{s}) \\ a_h \sim \pi(\cdot|\bar{s}_h) \\ \bar{s}_{h+1} \sim p(\cdot|\bar{s}_h, a_h)}} [r(\bar{s}_h, a_h, \bar{s}_{h+1})] + \gamma^{h+1} \mathbb{E}_{\bar{s}_{h+1} \sim p_{h+1}(\cdot|\bar{s})} V^{(T)}(\bar{s}_{h+1}) \end{aligned}$$

$$\begin{aligned}
 & - \gamma^h \mathbb{E}_{\substack{\bar{s}_h \sim p_h(\cdot|\bar{s}) \\ a_h \sim \pi(\cdot|\bar{s}_h) \\ \bar{s}_{t+1} \sim p^{(T)}(\cdot|\bar{s}_h, a_h)}} [r(\bar{s}_h, a_h, \bar{s}_{h+1})] - \gamma^{h+1} \mathbb{E}_{\bar{s}_{h+1} \sim p^{(T)} \circ p_h(\cdot|\bar{s})} V^{(T)}(\bar{s}_{h+1}) \\
 & = \gamma^h \mathbb{E}_{\substack{\bar{s}_h \sim p_h(\cdot|\bar{s}) \\ a_h \sim \pi(\cdot|\bar{s}_h)}} \left[\mathbb{E}_{\bar{s}_{h+1} \sim p(\cdot|\bar{s}_h, a_h)} r(\bar{s}_h, a_h, \bar{s}_{h+1}) - \mathbb{E}_{\bar{s}_{h+1} \sim p^{(T)}(\cdot|\bar{s}_h, a_h)} r(\bar{s}_h, a_h, \bar{s}_{h+1}) \right] \\
 & \quad + \gamma^{h+1} \mathbb{E}_{\substack{\bar{s}_h \sim p_h(\cdot|\bar{s}) \\ a_h \sim \pi(\cdot|\bar{s}_h, a_h)}} \left[\mathbb{E}_{\bar{s}_{h+1} \sim p(\cdot|\bar{s}_h, a_h)} V^{(T)}(\bar{s}_{h+1}) - \mathbb{E}_{\bar{s}_{h+1} \sim p^{(T)}(\cdot|\bar{s}_h, a_h)} V^{(T)}(\bar{s}_{h+1}) \right].
 \end{aligned}$$

Because the current state and current event marks are independent and depend only on the previous events, the transition kernel can be factored as two independent distributions over S and $\mathcal{X}$. Hence, replacing the augmented state $\bar{s} \in \bar{S}$ with the explicit representation (s, H) , where $s \in S$ and $H \in \mathcal{X}^\infty$ gives, for any bounded measurable function f on $S \times \mathcal{X}^\infty$,

$$\begin{aligned}
 & \mathbb{E}_{(s_{t+1}, H_{t+1}) \sim p(\cdot|s_t, H_t, a_t)} f(s_{t+1}, H_{t+1}) - \mathbb{E}_{(s_{t+1}, H_{t+1}) \sim p^{(T)}(\cdot|s_t, H_t, a_t)} f(s_{t+1}, H_{t+1}) \\
 & = \mathbb{E}_{\substack{s_{t+1} \sim Q_{H_t}(\cdot|s_t, a_t) \\ X_{t+1} \sim Q_{H_t}^X(\cdot|s_t, a_t)}} f(s_{t+1}, X_{t+1}, H_t) - \mathbb{E}_{\substack{s_{t+1} \sim Q_{H_t}^{(T)}(\cdot|s_t, a_t) \\ X_{t+1} \sim Q_{H_t}^{X(T)}(\cdot|s_t, a_t)}} f(s_{t+1}, X_{t+1}, H_t) \\
 & = \mathbb{E}_{\substack{s_{t+1} \sim Q_{H_t}(\cdot|s_t, a_t) \\ X_{t+1} \sim Q_{H_t}^X(\cdot)}} f(s_{t+1}, X_{t+1}, H_t) - \mathbb{E}_{\substack{s_{t+1} \sim Q_{H_t}^{(T)}(\cdot|s_t, a_t) \\ X_{t+1} \sim Q_{H_t}^X(\cdot)}} f(s_{t+1}, X_{t+1}, H_t) \\
 & \quad + \mathbb{E}_{\substack{s_{t+1} \sim Q_{H_t}^{(T)}(\cdot|s_t, a_t) \\ X_{t+1} \sim Q_{H_t}^X(\cdot)}} f(s_{t+1}, X_{t+1}, H_t) - \mathbb{E}_{\substack{s_{t+1} \sim Q_{H_t}^{(T)}(\cdot|s_t, a_t) \\ X_{t+1} \sim Q_{H_t}^{X(T)}(\cdot)}} f(s_{t+1}, X_{t+1}, H_t) \\
 & \leq \mathbb{E}_{X_{t+1} \sim Q_{H_t}^X(\cdot)} \left[\|f(\cdot, X_{t+1}, H_t)\|_\infty \text{TV}\left(Q_{H_t}(\cdot|s_t, a_t), Q_{H_t}^{(T)}(\cdot|s_t, a_t)\right) \right] \\
 & \quad + \mathbb{E}_{s_{t+1} \sim Q_{H_t}^{(T)}(\cdot|s_t, a_t)} \left[\|f(s_{t+1}, \cdot, H_t)\|_\infty \text{TV}\left(Q_{H_t}^X(\cdot), Q_{H_t}^{X(T)}(\cdot)\right) \right] \\
 & \leq \|f\|_\infty \left(\sum_{t=T+1}^{\infty} M_t + \sum_{t=T+1}^{\infty} N_t \right). \tag{31}
 \end{aligned}$$

The second-to-last inequality is a result of the Total Variation Distance being an integral probability metric (Sriperumbudur et al., 2012). Using this expression in the previous equation gives

$$\begin{aligned}
 V_{h+1}^{(T)}(\bar{s}) - V_h^{(T)}(\bar{s}) & \leq \gamma^h \|r\|_\infty \left(\sum_{t=T+1}^{\infty} (M_t + N_t) \right) + \gamma^{h+1} \|V^{(T)}\|_\infty \left(\sum_{t=T+1}^{\infty} (M_t + N_t) \right) \\
 & \leq \gamma^h (\|r\|_\infty + \gamma \|V^{(T)}\|_\infty) \left(\sum_{t=T+1}^{\infty} (M_t + N_t) \right).
 \end{aligned}$$

This gives the final result,

$$\begin{aligned}\|V - V^{(T)}\|_\infty &\leq \sum_{h=0}^{\infty} \gamma^h (\|r\|_\infty + \gamma \|V^{(T)}\|_\infty) \left(\sum_{t=T+1}^{\infty} (M_t + N_t) \right) \\ &= \frac{1}{1-\gamma} (\|r\|_\infty + \gamma \|V^{(T)}\|_\infty) \left(\sum_{t=T+1}^{\infty} (M_t + N_t) \right).\end{aligned}$$

■

C.1 Proof of Lemma 4

Let π be a policy that depends only on the current state and the past T events.

$$\begin{aligned}V^\pi((s, x_{0:\infty})) &= \mathbb{E}_{\substack{s' \sim Q_{x_{0:\infty}}(\cdot|s,a) \\ x' \sim Q_{x_{0:\infty}}(\cdot) \\ a = \pi((s, x_{0:T}))}} [r(s, a, s') + \gamma V^\pi((s', x', x_{0:\infty}))], \\ \text{and } V^\pi((s, x_{0:T}, 0)) &= \mathbb{E}_{\substack{s' \sim Q_{x_{0:T}}(\cdot|s,a) \\ x' \sim Q_{x_{0:T}}(\cdot) \\ a = \pi((s, x_{0:T}))}} [r(s, a, s') + \gamma V^\pi((s', x', x_{0:T}, 0))].\end{aligned}$$

Taking the difference between these two expressions and adding and subtracting the additional term

$$\mathbb{E}_{\substack{s' \sim Q_{x_{0:T}}(\cdot|s,a) \\ x' \sim Q_{x_{0:T}}(\cdot) \\ a = \pi((s, x_{0:T}))}} V^\pi((s', x', x_{0:\infty}))$$

$$\begin{aligned}\text{gives } &|V^\pi((s, x_{0:\infty})) - V^\pi((s, x_{0:T}, 0))| \\ &\leq \|r\|_\infty \left(\sum_{t=T+1}^{\infty} M_t \right) + \gamma \|V^\pi\|_\infty \left(\sum_{t=T+1}^{\infty} M_t + N_t \right) \\ &\quad + \gamma \mathbb{E}_{\substack{s' \sim Q_{x_{0:T}}(\cdot|s,a) \\ x' \sim Q_{x_{0:T}}(\cdot) \\ a = \pi((s, x_{0:T}))}} |V^\pi((s', x', x_{0:\infty})) - V^\pi((s', x', x_{0:T}, 0))|.\end{aligned}$$

The first term on the right-hand side is a result of the modification of inequality (31), where the function is not dependent on the external process. In the above inequality, the left-hand side is the difference between two value functions whose first T events are the same and differ everywhere else, but the right-hand side has a discounted term in which the first $T + 1$ terms coincide. Therefore, repeating the same procedure results in value function differences in states that increasingly coincide.

Since each repetition adds a γ factor to the term, which itself is bounded by $\frac{1}{1-\gamma}$ due to the rewards being bounded, the resulting series converges, giving us the following bound.

$$\sup_{\tilde{s}=(s, x_{0:\infty})} |V^\pi((s, x_{0:\infty})) - V^\pi((s, x_{0:T}, 0))|$$

$$\begin{aligned}
 &\leq \sum_{t=T+1}^{\infty} \gamma^{t-T-1} \left[\|r\|_{\infty} \left(\sum_{t'=t}^{\infty} M_{t'} \right) + \gamma \|V\|_{\infty} \left(\sum_{t'=t}^{\infty} M_{t'} + N_{t'} \right) \right] \\
 &\leq \left(1 + \frac{\gamma}{1-\gamma} \right) \sum_{t=T+1}^{\infty} \sum_{t'=t}^{\infty} \gamma^{t-T-1} (M_{t'} + N_{t'}) \\
 &\leq \frac{1}{1-\gamma} \sum_{t'=T+1}^{\infty} \sum_{t=0}^{T-t'-1} \gamma^t (M_{t'} + N_{t'}) \\
 &= \frac{1}{1-\gamma} \sum_{t'=T+1}^{\infty} \frac{1-\gamma^{t'-T}}{1-\gamma} (M_{t'} + N_{t'}) \\
 &\leq \frac{1}{(1-\gamma)^2} \sum_{t=T+1}^{\infty} (M_t + N_t).
 \end{aligned}$$

■

Appendix D. Sample Complexity

In this section, we provide details of the proofs of the sample complexity of policy evaluation as well as the overall policy iteration algorithm, wherein the policy evaluation step is performed using LSTD. The analysis closely follows the results in Lazaric et al. (2012). We state the preliminaries and statements of some of their lemmas and provide details of where our proof differs from theirs.

D.1 Sample Complexity of Least-Squares Policy Evaluation

This theorem states an upper bound on the expected error of the learned value function. Therefore, the proof has two parts: the first is Lemma 9, which is a bound on the empirical value function, and the second is a generalization to the entire space. The bound on the empirical error is quite similar to the original in the stationary case; therefore, we skip some steps. The final bound on the expected error, whose proof is provided in Section D.1.2, offers more insights into the trade-off between the tractability of the algorithm and the approximation error due to nonstationarity.

D.1.1 EMPIRICAL ERROR

We wish to show that with probability at least $1 - \delta$,

$$\|v - \hat{v}\|_N \leq \frac{1}{\sqrt{1-\gamma^2}} \|v - \widehat{\Pi}v\|_N + \frac{L}{(1-\gamma)^2} \sqrt{\frac{d}{v_N}} \left(\sqrt{\frac{2 \log(2d/\delta)}{N}} + \frac{1}{N} \right).$$

Following Lazaric et al. (2012),

$$\|V - \widehat{V}\|_N = \|v - \hat{v}\|_N \leq \frac{1}{\sqrt{1-\gamma^2}} \|v - \widehat{\Pi}v\|_N + \frac{1}{1-\gamma} \|\widehat{\Pi}v - \widehat{\Pi}\widehat{\mathcal{T}}v\|_N. \quad (32)$$

The first term is the approximation error, which is unavoidable owing to the restricted function space. The second term also includes the estimation error due to using the path-wise Bellman operator instead of the actual Bellman operator and is a function of the

number of samples used in the training. It can be written as $\|\widehat{\Pi}\xi\|_N$, where $\xi = \widehat{\mathcal{T}}v - v$ is the estimation error of the output. Now, for $t < N$,

$$\begin{aligned}\xi_t &= r_t + \gamma V(\bar{s}_{t+1}) - V(\bar{s}_t) \\ &= r(s_t, \pi(\bar{s}_t), s_{t+1}) + \gamma V(\bar{s}_{t+1}) - \mathbb{E}_{\bar{s}_{t+1}} \left[r(s_t, \pi(\bar{s}_t), \bar{s}_{t+1}) + \gamma V(\bar{s}_{t+1}) \right] \\ &= r(s_t, \pi(\bar{s}_t), s_{t+1}) - \mathbb{E}_{S_{t+1} \sim Q_{x,t}(\cdot | s_t, \pi(\bar{s}_t))} [r(s_t, \pi(\bar{s}_t), S_{t+1})] \\ &\quad + \gamma \left[V(\bar{s}_{t+1}) - \mathbb{E}_{\bar{s}_{t+1} \sim Q, Q_{x,t}^X(\cdot | s_t, \pi(\bar{s}_t))} V(\bar{s}_{t+1}) \right],\end{aligned}$$

and for $t = N$,

$$\begin{aligned}\xi_t &= r(s_t, \pi(\bar{s}_t), s_{t+1}) - \mathbb{E}_{S_{t+1} \sim Q_{x,t}(\cdot | s_t, \pi(\bar{s}_t))} [r(s_t, \pi(\bar{s}_t), S_{t+1})] \\ &\quad - \gamma \mathbb{E}_{\bar{s}_{t+1} \sim Q, Q_{x,t}^X(\cdot | s_t, \pi(\bar{s}_t))} [V(\bar{s}_{t+1})].\end{aligned}$$

ξ_t are functions of $\bar{s}_t$, which are samples from the Markov chain induced by policy π in the augmented MDP $\bar{M}$. Considering them as random variables, $(\xi_t \varphi_i(\bar{s}_t))_{1 \leq t \leq N}$ is a martingale difference sequence w.r.t $\bar{s}_t$, with each element being zero mean and bounded by $\frac{L}{1-\gamma}$. So, applying Azuma's inequality to $(\xi_t \varphi_i(\bar{s}_t))_{1 \leq t \leq N}$ gives, for each $i \in [d]$,

$$\mathbb{P} \left(\left| \sum_{t=1}^{N-1} \xi_t \varphi_i(s_t) \right| \geq \epsilon \right) < 2 \exp \left(-\frac{\epsilon^2 (1-\gamma)^2}{2(N-1)L^2} \right).$$

Letting the right hand side be $\frac{\delta}{d}$, with probability at least $1 - \delta$,

$$\left| \sum_{t=1}^{N-1} \xi_t \varphi_i(s_t) \right| < \frac{L}{1-\gamma} \sqrt{2(N-1) \log \left(\frac{2d}{\delta} \right)}, \quad \text{for all } i \in [d].$$

Since $|\xi_n \varphi_i(\bar{s}_t)| < \frac{L}{1-\gamma}$ always, we have,

$$\left| \sum_{t=1}^{N-1} \xi_t \varphi_i(s_t) \right| < \frac{L}{1-\gamma} \left(\sqrt{2(N-1) \log \left(\frac{2d}{\delta} \right)} + 1 \right), \quad \text{for all } i \in [d] \text{ w.p. } \geq 1 - \delta. \quad (33)$$

Following Lazaric et al. (2012), this can be used to bound $\|\hat{\Pi}\xi_n\|_N$ as

$$\left\| \widehat{\Pi} \widehat{\mathcal{T}}v - \widehat{\Pi}v \right\|_N \leq \frac{1}{N} \sqrt{\frac{d}{v_n}} \max_i \left| \sum_{t=1}^N \xi_t \varphi_i(s_t) \right|, \quad (34)$$

where v_n is the smallest positive eigenvalue of the Gram matrix $\Phi^\top \Phi$. By substituting (33) and (34) in (32) gives the desired result.

D.1.2 GENERALIZATION

Intuitively, bounding the expected error of the estimated value function based on the error at only a few samples requires the samples to be close to the stationary distribution and the function values at the samples to be close to their expected values.

The first requirement is formalized in terms of the rate of mixing of the Markov chain induced by the MDP using the policy.

Definition 12. A Markov chain $\mathcal{M} = (X_t)_{t \geq 1}$ is said to be exponentially β -mixing with parameters $\bar{\beta}, b, \kappa$ if its mixing coefficients

$$\beta_i = \sup_{t \geq 1} \left[\sup_{B \in \sigma(X_{t+i}, \dots)} |\mathbb{P}(B|X_1, \dots, X_t) - \mathbb{P}(B)| \right]$$

satisfy $\beta_i \leq \bar{\beta} \exp(-bi^\kappa)$.

When such an exponentially mixing Markov chain is also ergodic and aperiodic with stationary distribution ρ , for any initial distribution λ , there is a bound on the following total variation: (Definition 21 of Lazaric et al. (2012))

$$\left\| \int_{\mathcal{X}} \lambda(dx) \mathbb{P}(\cdot|x) - \rho(\cdot) \right\|_{\text{TV}} \leq \bar{\beta} \exp(-bi^\kappa).$$

This helps bound the difference between the norm of a function in the stationary distribution and the empirical distribution defined using a few samples.

Lemma 13 (Generalization Lemma for Markov chains, Lemma 24 of Lazaric et al. (2012)). Let $\tilde{\mathcal{F}}$ be the d -dimensional class of linear functions truncated at the threshold B . Let $(X_t)_{t \in [n]}$ be a sequence of samples from an ergodic, aperiodic, exponentially β -mixing Markov chain with parameters $\bar{\beta}, b, \kappa$, arbitrary initial distribution, and stationary distribution ρ . If the first $\tilde{n}$ samples are discarded and only the last $n - \tilde{n}$ samples are used, then with probability at least $1 - \delta$, the empirical and $l_2(\rho)$ norm of any function $\tilde{f} \in \tilde{\mathcal{F}}$ are related as,

$$\begin{aligned} \|\tilde{f}\| - 2\|\tilde{f}\|_{X_{1:n}} &\leq \epsilon(\delta), \\ \|\tilde{f}\|_{X_{1:n}} - 2\sqrt{2}\|\tilde{f}\| &\leq \epsilon(\delta), \\ \text{for } \epsilon(\delta) &= 12B \sqrt{\frac{2\Lambda(n - \tilde{n}, d, \delta)}{n - \tilde{n}}} \max \left\{ \frac{\Lambda(n - \tilde{n}, d, \delta)}{1}, 1 \right\}^{\frac{1}{\kappa}}, \\ \Lambda(n, d, \delta) &= 2(d + 1) \log n + \log \frac{\epsilon}{\delta} + \log^+ \left(\max \left\{ 16(6e)^{2(d+1)}, \bar{\beta} \right\} \right), \\ \text{and } \tilde{n} &= \left(\frac{1}{b} \log \left(\frac{2e\bar{\beta}n}{\delta} \right) \right)^{\frac{1}{\kappa}}. \end{aligned}$$

The above lemma is essentially a generalization bound for truncated linear functions operating on samples from an exponentially mixing Markov chain and is used in the proof of Theorem 10 to translate the bound on the empirical error to a bound on the expected error in the stationary distribution of the Markov chain induced by the policy under evaluation.

The above bounds hold for every member of the d -dimensional class of functions, hence the dependence of ϵ on d through $\Lambda(n, d, \delta)$. In contrast, when there is just one truncated linear function under consideration, the above bound holds, but with

$$\Lambda(n, \delta) = \log \frac{\epsilon}{\delta} + \log \left(\max \left\{ 6, n\bar{\beta} \right\} \right).$$

Another issue in generalizing the sample-based bound in Lemma 9 to the entire state space is its dependence on the eigenvalues of the sample Gram matrix. To tackle this problem, Lazaric et al. (2012) derived the following probabilistic lower bound on the smallest eigenvalue of the sample-based Gram matrix as a function of the smallest eigenvalue of the Gram matrix $G = \mathbb{E}_{x \sim \rho} [\phi(x)\phi(x)^\top]$.

Lemma 14 (Lemma 4 of Lazaric et al. (2012)). *Let $\omega > 0$ be the smallest eigenvalue of the Gram matrix, defined above. If the data generating process is an exponentially β -mixing Markov chain with parameters $\bar{\beta}, b, \kappa$, then the smallest eigenvalue v_n of the sample-based Gram matrix constructed using n samples satisfies*

$$\sqrt{v_n} \geq \sqrt{v} = \frac{\sqrt{\omega}}{2} - 6L \sqrt{\frac{2\Lambda(n, d, \delta)}{n} \max \left\{ \frac{\Lambda(n, d, \delta)}{b}, 1 \right\}^{\frac{1}{\kappa}}} > 0,$$

provided the number of samples n satisfies

$$n > \frac{288L^2 \Lambda(n, d, \delta)}{\omega} \max \left\{ \frac{\Lambda(n, d, \delta)}{b}, 1 \right\}^{\frac{1}{\kappa}},$$

$$\text{where } \Lambda(n, d, \delta) = 2(d+1) \log n + \log \frac{e}{\delta} + \log^+ \left(\max \left\{ 18(6e)^{2(d+1)}, \bar{\beta} \right\} \right)$$

comes from a simpler version of Lemma 13 without the additional initial $\tilde{n}$ samples for the burn-in of the Markov chain.

The generalization Lemma 13, combined with the lower bound on the eigenvalues and our results in Sections 4 and 5.1, give rise to Theorem 10 on the sample complexity of policy evaluation.

D.2 Sample Complexity of Approximate Least-Squares Policy Iteration (Algorithm 2)

Here, we develop a recursive bound on the approximate suboptimality $\|V^* - \tilde{V}^{\pi_{k+1}}\|$ needed to study the propagation of error and obtain an upper bound on the suboptimality of the policy after K steps of our algorithm. Denote $\ell_k = V^* - V^{\pi_k}$, $e_k = \tilde{V}_k - V^{\pi_k}$, $g_k = V^{\pi_{k+1}} - V^{\pi_k}$, $b_k = \tilde{V}_k - T^{\pi_k} \tilde{V}_k$, and $h_k = T^{\pi_{k+1}} \tilde{V}_k - T^{\pi_k} \tilde{V}_k$.

Lemma 15.

$$\begin{aligned} \ell_{k+1} &\leq \gamma P^{\pi^*} l_k + \gamma P^{\pi_{k+1}} \left[\left[I - \gamma (I - \gamma P^{\pi_{k+1}})^{-1} (P^{\pi_k} - P^{\pi_{k+1}}) \right] e_k \right. \\ &\quad \left. + [I - \gamma P^{\pi_{k+1}}]^{-1} h_k \right] - \gamma P^{\pi^*} e_k + h_k \\ &\leq \gamma P^{\pi^*} l_k + \gamma \left[P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} - P^{\pi^*} (I - \gamma P^{\pi_k})^{-1} \right] b_k \\ &\quad + \left[\gamma P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} + I \right] h_k \end{aligned}$$

Proof We adapt and modify Lemmas 2-4 of Munos (2003) to account for the fact that the policy under consideration by our algorithm is not the true greedy policy, but an approximately greedy policy that considers only a subset of the extended state, which is the current state and a finite history of events. Following Lemma 2 of Munos (2003), we obtain

$$\begin{aligned}
 \ell_{k+1} &= V^* - V^{\pi_{k+1}} \\
 &= (T^{\pi^*} V^* - T^{\pi^*} V^{\pi_k}) + (T^{\pi^*} V^{\pi_k} - T^{\pi^*} \tilde{V}_k) + (T^{\pi^*} \tilde{V}_k - T^{\tilde{\pi}_{k+1}} \tilde{V}_k) \\
 &\quad + (T^{\tilde{\pi}_{k+1}} \tilde{V}_k - T^{\pi_{k+1}} \tilde{V}_k) + (T^{\pi_{k+1}} \tilde{V}_k - T^{\pi_{k+1}} V^{\pi_k}) + (T^{\pi_{k+1}} V^{\pi_k} - T^{\pi_{k+1}} V^{\pi_{k+1}}) \\
 &\leq \gamma P^{\pi^*} l_k - \gamma P^{\pi^*} e_k + 0 + h_k + \gamma P^{\pi_{k+1}} e_k - \gamma P^{\pi_{k+1}} g_k.
 \end{aligned} \tag{35}$$

Similarly, following Lemma 3 of Munos (2003),

$$\begin{aligned}
 g_k &= V^{\pi_{k+1}} - V^{\pi_k} \\
 &= (T^{\pi_{k+1}} V^{\pi_{k+1}} - T^{\pi_{k+1}} V^{\pi_k}) + (T^{\pi_{k+1}} V^{\pi_k} - T^{\pi_{k+1}} \tilde{V}_k) + (T^{\pi_{k+1}} \tilde{V}_k - T^{\tilde{\pi}_{k+1}} \tilde{V}_k) \\
 &\quad + (T^{\tilde{\pi}_{k+1}} \tilde{V}_k - T^{\pi_k} \tilde{V}_k) + (T^{\pi_k} \tilde{V}_k - T^{\pi_k} V^{\pi_k}) \\
 &\geq \gamma P^{\pi_{k+1}} g_k - \gamma P^{\pi_{k+1}} h_k + \gamma P^{\pi_k} e_k \\
 &\geq -[I - \gamma P^{\pi_{k+1}}]^{-1} [\gamma (P^{\pi_{k+1}} - P^{\pi_k}) e_k + h_k].
 \end{aligned}$$

Finally, following Lemma 4 of Munos (2003),

$$e_k - g_k \leq [I - \gamma (I - \gamma P^{\pi_{k+1}})^{-1} (P^{\pi_k} - P^{\pi_{k+1}})] e_k + [I - \gamma P^{\pi_{k+1}}]^{-1} h_k.$$

Substituting the above in (35) yields the desired result. ■

Lemma 16.

$$|h_k(\bar{s})| \leq \frac{2R_{\max}}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t)$$

for all states $\bar{s} \in \bar{\mathcal{S}}$, the augmented state space.

Proof For $\bar{s}^{(T)} = (s, x_{0:T})$, the approximately greedy policy and true greedy policy are defined as below.

$$\begin{aligned}
 \pi_{k+1}(\bar{s}^{(T)}) &= \underset{a \in A}{\operatorname{argmax}} \mathbb{E}_{s' \sim Q_{x_{0:T}, (0)}(\cdot | s, a)} \left[r(s, a, s') + \gamma \tilde{V}_k((s', x', x_{0:T-1})) \right] \\
 &\quad x' \sim Q_{x_{0:T}, (0)}^X(\cdot | s) \\
 \tilde{\pi}_{k+1}(\bar{s}) &= \underset{a \in A}{\operatorname{argmax}} \mathbb{E}_{s' \sim Q_{x_{0:\infty}}(\cdot | s, a)} \left[r(s, a, s') + \gamma \tilde{V}_k((s', x', x_{0:T-1})) \right] \\
 &\quad x' \sim Q_{x_{0:\infty}}^X(\cdot | s)
 \end{aligned}$$

Define corresponding Bellman operators

$$(T^{\pi_{k+1}} V_k)(\bar{s}) = \mathbb{E}_{s' \sim Q_{x_{0:\infty}}(\cdot | s, \pi_{k+1}(\bar{s}^{(T)}))} \left[r(s, \pi_{k+1}(\bar{s}^{(T)}), s') + \gamma \tilde{V}_k((s', x', x_{0:T-1})) \right] \\
 x' \sim Q_{x_{0:\infty}}^X(\cdot | s)$$

$$(T^{\pi_{k+1}} V_k)(\bar{s}) = \mathbb{E}_{\substack{s' \sim Q_{x_{0:\infty}}(\cdot | s, \pi_{k+1}(\bar{s}^{(T)})) \\ x' \sim Q_{x_{0:\infty}}^X(\cdot | s)}} \left[r(s, \pi_{k+1}(\bar{s}^{(T)}), s') + \gamma \widetilde{V}_k((s', x', x_{0:T-1})) \right]$$

$$(T^{\bar{\pi}_{k+1}} V_k)(\bar{s}) = \mathbb{E}_{\substack{s' \sim Q_{x_{0:\infty}}(\cdot | s, \bar{\pi}_{k+1}(\bar{s})) \\ x' \sim Q_{x_{0:\infty}}^X(\cdot | s)}} \left[r(s, \bar{\pi}_{k+1}(\bar{s}), s') + \gamma \widetilde{V}_k((s', x', x_{0:T-1})) \right]$$

Define the approximate state-action value function $\widetilde{V}_{\text{action}}$ function as below (state-action function is also referred to as Q function, to avoid conflict of notation, we denote this by V_{action})

$$\widetilde{V}_{\text{action}(\bar{s}, a)} = \mathbb{E}_{\substack{s' \sim Q_{x_{0:\infty}}(\cdot | s, a) \\ x' \sim Q_{x_{0:\infty}}^X(\cdot | s)}} \left[r(s, a, s') + \gamma \widetilde{V}_k((s', x', x_{0:T-1})) \right]$$

Using the definition of $\widetilde{V}_{\text{action}}$, we can write policies π_{k+1} , $\bar{\pi}_{k+1}$ and their corresponding Bellman operators as below

$$\pi_{k+1}(\bar{s}^{(T)}) = \arg \max_a \widetilde{V}_{\text{action}}((\bar{s}^{(T)}, (0)_{T+1:\infty}), a)$$

$$\bar{\pi}_{k+1}(\bar{s}) = \arg \max_a \widetilde{V}_{\text{action}}(\bar{s}, a)$$

$$(T^{\pi_{k+1}} V_k)(\bar{s}) = \widetilde{V}_{\text{action}}(\bar{s}, \pi_{k+1}(\bar{s}^{(T)})) = \widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty})))$$

$$(T^{\bar{\pi}_{k+1}} V_k)(\bar{s}) = \widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}(\bar{s})) = \max_a \widetilde{V}_{\text{action}}(\bar{s}, a)$$

Using the above definitions, we can write $h_k(\bar{s})$ for all $\bar{s} \in \bar{S}$ as

$$h_k(\bar{s}) = \widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}(\bar{s})) - \widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty})))$$

$$h_k(\bar{s}) = \widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}(\bar{s})) - \widetilde{V}_{\text{action}}((\bar{s}^{(T)}, (0)_{T+1:\infty}), \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty})))$$

$$+ \widetilde{V}_{\text{action}}((\bar{s}^{(T)}, (0)_{T+1:\infty}), \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty}))) - \widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty}))) \quad (36)$$

Using (31), we can bound

$$\widetilde{V}_{\text{action}}((\bar{s}^{(T)}, (0)_{T+1:\infty}), \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty}))) - \widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty})))$$

$$\leq \|\widetilde{V}_{\text{action}}\|_{\infty} \sum_{t=T+1}^{\infty} (M_t + N_t) \leq \frac{R_{\max}}{1 - \gamma} \sum_{t=T+1}^{\infty} (M_t + N_t) \quad (37)$$

Bounding the first term of (36),

$$\widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}(\bar{s})) - \widetilde{V}_{\text{action}}((\bar{s}^{(T)}, (0)_{T+1:\infty}), \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty})))$$

$$\begin{aligned}
 &= \widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}(\bar{s})) - \widetilde{V}_{\text{action}}((\bar{s}^{(T)}, (0)_{T+1:\infty}), \bar{\pi}_{k+1}(\bar{s})) \\
 &\quad + \widetilde{V}_{\text{action}}((\bar{s}^{(T)}, (0)_{T+1:\infty}), \bar{\pi}_{k+1}(\bar{s})) - \widetilde{V}_{\text{action}}((\bar{s}^{(T)}, (0)_{T+1:\infty}), \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty})))
 \end{aligned}$$

Using (31) and by the definition of $\widetilde{V}_{\text{action}}$ and $\bar{\pi}_{k+1}$,

$$\begin{aligned}
 &\widetilde{V}_{\text{action}}(\bar{s}, \bar{\pi}_{k+1}(\bar{s})) - \widetilde{V}_{\text{action}}((\bar{s}^{(T)}, (0)_{T+1:\infty}), \bar{\pi}_{k+1}((\bar{s}^{(T)}, (0)_{T+1:\infty}))) \\
 &\leq \|\widetilde{V}_{\text{action}}\|_{\infty} \sum_{t=T+1}^{\infty} (M_t + N_t) \\
 &\leq \frac{R_{\max}}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t)
 \end{aligned} \tag{38}$$

By substituting (37) and (38) in (36), we get

$$h_k(\bar{s}) \leq \frac{2R_{\max}}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t).$$

Similarly,

$$-h_k(\bar{s}) \leq \frac{2R_{\max}}{1-\gamma} \sum_{t=T+1}^{\infty} (M_t + N_t).$$

■

References

- Lucas N. Alegre, Ana L. C. Bazzan, and Bruno C. da Silva. Minimum-delay adaptation in non-stationary reinforcement learning via online high-confidence change-point detection. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, (AAMAS), 2021.
- András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.
- Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *Proceedings of the 37th International Conference on Machine Learning*, (ICML), 2020.
- Thomas Dietterich, George Trimponias, and Zhitang Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, (ICML), 2018.
- Yonathan Efroni, Dylan J Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *35th Annual Conference on Learning Theory*, (COLT), 2022.

- Fan Feng, Biwei Huang, Kun Zhang, and Sara Magliacane. Factored adaptation for non-stationary reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35 of (*NeurIPS*), 2022.
- Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: datasets for deep data-driven reinforcement learning. *arXiv*, 2004.07219, 2020.
- Xin Guo, Yonghui Huang, and Yi Zhang. On average optimality for non-stationary markov decision processes in borel spaces. *Mathematics of Operations Research*, 2024.
- László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk, et al. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer, 2002.
- Emmanuel Hadoux, Aurélie Beynier, and Paul Weng. Solving hidden-semi-markov-mode markov decision problems. In *8th International Conference on Scalable Uncertainty Management*, (SUM), 2014.
- Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv*, 1502.02259, 2015.
- Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-time Markov control processes: basic optimality criteria*. Applications of Mathematics. Springer, 1996.
- Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 2012.
- Erwan Lecarpentier and Emmanuel Rachelson. Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32 of (*NeurIPS*), 2019.
- Ming Li, Chen Shi, Zhize Wu, and Piotr Fryzlewicz. Testing stationarity and change point detection in reinforcement learning. *Annals of Statistics*, 2025. To appear.
- Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations*, (ICLR), 2016.
- Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *29th International Conference on Algorithmic Learning Theory*, (ALT), 2018.
- Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the 20th International Conference on Machine Learning*, (ICML), 2003.
- Youngsoo Seol. Limit theorems for discrete hawkes processes. *Statistics & Probability Letters*, 99:223–229, 2015.
- Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

- Zachary Sunberg and Mykel Kochenderfer. Online algorithms for pomdps with continuous state, action, and observation spaces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, (ICAPS), 2018.
- Guy Tennenholtz, Nadav Merlis, Lior Shani, Martin Mladenov, and Craig Boutilier. Reinforcement learning with history dependent dynamic contexts. In *Proceedings of the 40th International Conference on Machine Learning*, (ICML), 2023.
- Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach. In *34th Annual Conference on Learning Theory*, (COLT), 2021.
- Chenjun Xiao, Yifan Wu, Chen Ma, Dale Schuurmans, and Martin Müller. Learning to Combat Compounding-Error in Model-Based Reinforcement Learning. In *NeurIPS 2019 Deep Reinforcement Learning Workshop*, (NeurIPS Workshop), 2019.