Square-root regret bounds for continuous-time episodic Markov decision processes

Xuefeng Gao [1], Xun Yu Zhou [2]

October 3, 2023

Abstract

We study reinforcement learning for continuous-time Markov decision processes (MDPs) in the finite-horizon episodic setting. In contrast to discrete-time MDPs, the inter-transition times of a continuous-time MDP are exponentially distributed with rate parameters depending on the state–action pair at each transition. We present a learning algorithm based on the methods of value iteration and upper confidence bound. We derive an upper bound on the worst-case expected regret for the proposed algorithm, and establish a worst-case lower bound, with both bounds of the order of square-root on the number of episodes. Finally, we conduct simulation experiments to illustrate the performance of our algorithm.

1 Introduction

Reinforcement learning (RL) studies the problem of sequential decision making in an unknown environment by carefully balancing between exploration (learning) and exploitation (optimizing) (Sutton and Barto 2018). While the RL study has a relatively long history, it has received considerable attention in the past decades due to the explosion of available data and rapid improvement of computing power. A hitherto default mathematical framework for RL is Markov decision process (MDP), where the agent does not know the transition probabilities and can observe a reward resulting from an action but does not know the reward function itself. There has been extensive research on RL for *discrete-time* MDPs (DTMDPs); see, e.g., Jaksch et al. (2010), Osband and Van Roy (2017), Azar et al. (2017), Jin et al. (2018). However, much less attention has been paid to RL for *continuous-time* MDPs, whereas there are many real-world applications where one *needs* to interact with the unknown environment and learn the optimal strategies continuously in time. Examples include autonomous driving, control of queueing systems, control of infectious diseases, preventive maintenance and robot navigation; see, e.g., Guo and Hernández-Lerma (2009), Piunovskiy and Zhang (2020), Chapter 11 of Puterman (2014) and the references therein.

In this paper we study RL for tabular continuous-time Markov decision processes (CTMDPs) in the finite-horizon, episodic setting, where an agent interacts with the unknown environment in episodes of a fixed length with finite state and action spaces. The study of model-based (i.e. the underlying models are assumed to be known) finite-horizon CTMDPs has a very long history, probably dating back to Miller (1968), with vast applications including queueing optimization (Lippman 1976), dynamic pricing (Gallego and Van Ryzin 1994), and finance and insurance (Bäuerle and

---

[1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China. Email: `xfgao@se.cuhk.edu.hk`.

[2]Department of Industrial Engineering and Operations Research and The Data Science Institute, Columbia University, New York, NY 10027, USA. Email: `xz2574@columbia.edu`.

Rieder 2011). However, the problem of learning the optimal policy of a finite-horizon *unknown* CTMDP has not been studied in the literature, to our best knowledge. The goal of this paper is to fill this gap by developing a learning algorithm for episodic RL in tabular CTMDPs with rigorous worst-case regret guarantees, as well as providing a worst-case regret lower bound. Here, the regret measures the difference between the reward collected by the algorithm during learning and the reward of an optimal policy should the model parameters be completely known.

At the outset, one may think learning CTMDPs is just a straightforward extension of its discrete-time counterpart in terms of theoretical analysis and algorithm design and thus not worthy of a separate study. This is not the case. The reason is because a CTMDP is characteristically different from a DTMDP in that, after an agent observes a state x and takes an action a, the process remains in this state for a *random* holding time period that follows an exponential distribution with rate parameter $\lambda(x, a)$. Then the process jumps to another state y with a transition probability $p(y|x, a)$. For DTMDPs, the holding times (i.e., inter-transition times) are deterministic and take value of one time step in the general framework of semi-Markov decision processes (Puterman 2014, Chapter 11). This difference, along with the continuity of the time parameter, leads to several subtle yet substantial challenges in the design and analysis of episodic RL algorithms for CTMDPs. In particular, existing derivations of state-of-the-art worst-case regret bounds for learning tabular DTMDPs (see, e.g., Azar et al. 2017, Zanette and Brunskill 2019) often rely on induction/recursion in the time parameter, which naturally fails for continuous-time problems.

In this paper, we develop a learning algorithm for CTMDPs with unknown transition probabilities and holding time rates, based on two key methodological ingredients. First, we build on the value iteration algorithms of Mamer (1986), Huang and Guo (2011) for solving finite-horizon CTMDPs when all the model parameters are known. This type of algorithms are based on an operator with a fixed point corresponding to the solution to the dynamic programming equation. A major benefit of this approach is that it naturally introduces a (discrete) number of iterations that can be used for induction arguments in regret analysis. Second, we take the idea of "optimism under uncertainty", i.e., one acts as if the environment is as nice as plausibly possible (Lattimore and Szepesvári 2020). This is a popular paradigm to address the exploration–exploitation trade-off in RL. To ensure optimism for the purpose of getting a reasonably good upper bound of the true value function, we carefully design the exploration bonus needed for efficient learning and regret analysis.

Because we are dealing with Markov chains (instead of, say, diffusion processes with continuous state space), it is only natural and indeed unavoidable that certain components of our analysis are inspired by those for discrete-time Markov chains, in particular the UCRL2 (Upper Confidence Reinforcement Learning) algorithm of Jaksch et al. (2010) and the UCBVI (Upper Confidence Bound Value Iteration) algorithm of Azar et al. (2017), both originally developed for RL in tabular DTMDPs. At the beginning of each episode, our algorithm estimates the transition probabilities and the holding time rate using the past data/samples. Based on the level of uncertainty in these estimates, we carefully design an exploration bonus for each state–action pair. We then update the policy by applying a (modified) value iteration to the estimated CTMDP with the reward augmented by the bonus.

Our main result, Theorem 1, indicates that the proposed RL algorithm has a worst case expected regret of $\tilde{O}(\sqrt{K})$ where K is the number of episodes and $\tilde{O}$ ignores logarithmic factors and hides

dependency on other constants. Moreover, Theorem 2 provides a lower bound showing that the square-root dependance of the worst case regret upper bound on K is optimal (up to a logarithmic factor). Furthermore, we conduct simulation experiments to illustrate the performance of our algorithm.

While our algorithm bears certain similarity to UCBVI of Azar et al. (2017) conceptually, there are *significant* differences in our algorithm design and regret analysis compared with the discrete time setting, which we elaborate below.

(a) First, in contrast with learning DTMDPs, we need to estimate the holding time rate $\lambda(x, a)$ of a CTMDP for each state–action pair (x, a) in our RL algorithm. In the regret analysis, we also need to analyze the estimation error of $\lambda(x, a)$ by establishing a finite sample confidence bound. Note that in the episodic setting, the random exponential holding time associated with each state–action pair may be truncated by the endpoint of the time horizon. This generates possibly non i.i.d. observations of holding times within the same episode. Hence, we can not directly apply concentration inequalities for i.i.d. exponential distributions to obtain the confidence bound. The key idea to overcome this difficulty is to 'piece together' the samples of exponential holding times at a pair (x, a) as realizations of inter-arrival times of a Poisson process with rate $\lambda(x, a)$ modelling the number of visits to the pair over time in an episode. Even if the time is eventually truncated, the number of visits to the pair (x, a) (more precisely, those visits where the next states are observed) in an episode still follows a Poisson distribution conditioned on the total amount of time staying at (x, a). Since the holding times at (x, a) across different episodes are independent, we can then apply the tail bounds for Poisson distribution to derive the desired confidence bound for the holding time rate $\lambda(x, a)$. Such an analysis is not required for learning DTMDPs.

(b) Second, because the value function of a CTMDP depends on both the transition probabilities and the holding time rates, the design of the exploration bonus in our algorithm, which is used to bound estimation errors on the value function, is more sophisticated compared with the discrete time setting. Specificaly, we need to carefully analyze how the estimation errors of holding time rates and transition probabilities *jointly* affect the computed value function in our learning algorithm. Moreover, because the "Bellman operator" in our continuous-time setting (see the operator $\mathcal{T}^a$ in Equation (6)) is more complicated due to the presence of exponential holding times, the error analysis of the learned model also become more involved.

(c) Third, our algorithm also differs from the classical UCBVI algorithm in terms of the planning oracle. In constast to solving finite-horizon DTMDPs, the value iteration algorithm for solving *finite-horizon* CTMDPs with known parameters converges generally in *infinite* number of iterations. This leads to important difference in the regret analysis, in particular in the proof of optimism (see Lemma 5). In the study of learning DTMDPs, the optimism can be proved by backward induction on the time parameter. In our setting of episodic CTMDPs, however, induction is not enough to establish optimism. Extra analysis of (modified) value iterations for CTMDPs (Algorithm 2) is required, because Algorithm 2 is stopped after a finite number of iterations and we need to quantify the error. In this paper, we show that for *finite-horizon* CTMDPs with bounded transition rates, the (modified) value iteration converges asymptotically with a linear rate $\rho \in (0, 1)$ when the number of iterations approach

infinity. This is achieved by showing that the Bellman operator for the empirical CTMDP with reward function augmented by the bonus (see the operator in Equation (15)) is a contraction mapping with the coefficient ρ. Such a technical analysis is not required for the classical UCBVI algorithm in the discrete-time setting.

(d) Fourth, in contrast to the analysis of UCBVI for DTMDPs, the standard pigeon-hole argument (see, e.g., Azar et al. (2017) or Lemma 7.5 in Chapter 7 of Agarwal et al. 2021), which is used to bound the sum of bonuses evaluated on data to obtain sublinear regret, can *not* be applied to our continuous time setting. Our exploration bonus involves both the estimation errors of transition probabilities and holding time rates. The estimation error of transition probabilities of a CTMDP is *not* inversely proportional to the square root of the *total visit counts* at each state-action pair. This is because after visiting a state–action pair, the random holding time may be truncated by the endpoint of the horizon, and hence the next transition/state may not be observed. This subtle difference, compared with the discrete-time setting, leads to substantial challenges in our regret analysis. In particular, it is possible that the sum of estimation errors of transition probabilities evaluated on data is linear in K for some sample paths (see (41)). This shows that the standard pathwise pigeon-hole argument can not be directly used for obtaining *sublinear* regret in learning CTMDPs. In our paper, we develop a new probabilistic method to overcome such a difficulty; see Proposition 1 and its proof. On the other hand, the estimation error in the holding time rate of a CTMDP is inversely proportional to the square root of the total time spent at each state after an action, and this total spending time is *real-valued* instead of *integer-valued*. Hence, the pigeon-hole principle as a type of counting argument does not apply. We overcome this difficulty by relating the total spending time with the visit counts for each state–action pair (see Proposition 2 and its proof).

(e) Finally, the transitions of a finite-horizon CTMDP do not happen according to a Poisson process in general, because the rates of exponential holding times $\lambda(\cdot,\cdot)$ depend on the state–action pairs. As a consequence, while the horizon length of each episode is fixed, the number of transitions in different episodes are random variables that are generally unbounded and statistically different from each other. This is in sharp contrast with regret minimization in episodic DTMDPs, where the number of decision steps is often fixed and common across different episodes. In our regret analysis, we overcome this difficulty by 'uniformizing' the total number of decision steps in each episode for CTMDPs with bounded transition rates, and then applying bounds for the maximum of K independent Poisson distributions with the same rate.

This paper is related to several studies on RL for CTMDPs in the *infinite-horizon* setting. Computational RL methods were developed for infinite-horizon CTMDPs as well as for the more general semi-Markov decision processes (SMDPs) very early on (Bradtke and Duff 1995, Das et al. 1999). However, available theoretical results on regret bounds for learning infinite-horizon CTMDPs are few and far between. Fruit and Lazaric (2017) study learning in continuous-time infinite-horizon average-reward SMDPs which is more general than CTMDPs in that the holding times can follow general distributions. They adapt the UCRL2 algorithm by Jaksch et al. (2010) and show that their algorithm achieves $\tilde{O}(\sqrt{n})$ worst-case regret after n decision steps. Instead of worst-case (instance-independent) performance bounds, Gao and Zhou (2022) recently establish instance-

dependent regret bounds that are logarithmic in time for learning tabular CTMDPs, again in the infinite-horizon average-reward setting. It is well recognized that the finite-horizon continuous-time decision problem is different from and indeed generally harder than its infinite-horizon counterpart in a number of aspects, including the need of considering the time variable in the value function and the truncation of the holding time discussed earlier. As such, our paper naturally differs from those on infinite-horizon problems.

We conclude this introduction by mentioning a growing body of literature on RL in continuous time with possibly continuous state and/or action spaces. Recently, Wang et al. (2020), Jia and Zhou (2022a,b, 2023) consider RL for diffusion processes and study respectively the problems of generating trial-and-error policies strategically, police evaluation, policy gradient and q-learning. However, the regret analysis remains largely untouched in that series of study. Basei et al. (2021), Guo et al. (2021) and Szpruch et al. (2021) study continuous-time RL for linear quadratic/convex models with *continuous* state spaces and propose algorithms with sublinear regret bounds in the finite-horizon episodic setting. By contrast, our work considers episodic RL for CTMDPs with a *discrete* state space.

The remainder of the paper is organized as follows. In Section 2, we formulate the problem of episodic RL for CTMDPs. In Section 3 we introduce our learning algorithm, while in Section 4 we present the main results on regret bounds. Section 5 reports the result of simulation experiments and Section 6 contains the proofs of the main results. Finally Section 7 concludes.

2   Episodic RL in a Tabular CTMDP

We consider a continuous-time Markov decision process (CTMDP) with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$ and a finite time horizon $[0, H]$. The CTMDP evolves as follows (Puterman 2014, Chapter 11). At the initial time 0, the process is in state $x_0 \in \mathcal{S}$ and an agent chooses an action $a_0 \in \mathcal{A}$. The process remains in state x_0 for a random holding time period τ_0 that follows an exponential distribution with rate parameter $\lambda(x_0, a_0)$, while rewards are continuously accrued at a rate $r(x_0, a_0)$ during the holding period τ_0. Then the process jumps to another state $x_1 \in \mathcal{S}$ at time τ_0 with a transition probability $p(x_1|x_0, a_0)$, and another action a_1 is made upon landing on x_1. This series of events is repeated until the end of the horizon, H. It is immediate to see that if action a is chosen in state x, then the joint probability that the holding time in state x is not greater than t and the next state is y is given by $\left(1 - e^{-\lambda(x,a)t}\right) \cdot p(y|x, a)$.

For $(x, a) \in \mathcal{S} \times \mathcal{A}$, define the (controlled) transition rates

$$q(y|x, a) := \lambda(x, a) \cdot p(y|x, a) \geq 0, \quad \text{for } y \in \mathcal{S}, \ y \neq x,$$

and set $q(x|x, a) = -\lambda(x, a) \leq 0$ so that $\sum_{z \in \mathcal{S}} q(z|x, a) \equiv 0$. Mathematically, a finite-horizon CTMDP model $\mathcal{M}$ is characterized by the following set:

$$\mathcal{M} := (\mathcal{S}, \mathcal{A}, r(\cdot, \cdot), q(\cdot|\cdot, \cdot), H, x_0), \tag{1}$$

where $r(\cdot, \cdot)$ is the reward function, $q(\cdot|\cdot, \cdot)$ is the transition rate, $H < \infty$ is the length of the horizon and x_0 is the initial state of the CTMDP model. Note that $q(\cdot|\cdot, \cdot)$ and $(\lambda(\cdot, \cdot), p(\cdot|\cdot, \cdot))$ are

equivalent in the sense that they determine each other. Hence, we will also occasionally use

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, r(\cdot, \cdot), \lambda(\cdot, \cdot), p(\cdot | \cdot, \cdot), H, x_0).$$

We emphasize that $(\lambda(x,a); (x,a) \in \mathcal{S} \times \mathcal{A})$ are the rate parameters of the exponential distributions of holding times in a CTMDP and they are *not* the discounting factors typically appearing in discrete-time reinforcement learning.

Throughout this paper, we make the following assumption.

Assumption 1. *There exist two constants $\lambda_{\min}, \lambda_{\max} \in (0, \infty)$, such that $\lambda(x,a) \in [\lambda_{\min}, \lambda_{\max}]$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$ and, without loss of generality, $r(x,a) \in [0,1]$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$. We also assume that the constant $\lambda_{\max}$ is known.*

This assumption states that the holding time rates and reward rates are bounded, which is fairly natural since the state and action spaces are both finite. In addition, it is assumed that we know the specific value of $\lambda_{\max}$, which is needed when we implement our proposed learning algorithm and carry out the theoretical analysis of the algorithm. For some CTMDPs, one may be able to obtain priori upper bounds on the holding time rates based on known information of the specific applied problems, although in general this may not be always the case. Finally, while our regret upper bound will depend on $\lambda_{\min}$, we do not need the knowledge of $\lambda_{\min}$ as an input for our algorithm.

Note that in our CTMDP model, the system evolves continuously in time, and the decision maker chooses actions at every transition when the system state changes/jumps. Given a control policy π, which is a map from $\mathcal{S} \times [0, H]$ to $\mathcal{A}$, denote by τ_i the holding time after i-th jump of a CTMDP model, by $t_n := \sum_{i=0}^{n-1} \tau_i$ the n-th jump time with $t_0 := 0$, by x_n the state at t_n, and by $a_n := \pi(x_n, (H - t_n)^+)$ the action dictated by π at t_n, where $(H - t_n)^+$ denotes the remaining horizon. We write

$$X(u) := \sum_{n \geq 0} I_{t_n \leq u < t_{n+1}} x_n, \quad A(u) := \sum_{n \geq 0} I_{t_n \leq u < t_{n+1}} a_n, \quad u \geq 0,$$

as the corresponding continuous-time state and action processes respectively under π, where I_C denotes the indicator function on a set C. Define the value function under π:

$$V^\pi(x,t) := \mathbb{E}^\pi \left[\int_0^t r(X(u), A(u)) du \Big| X(0) = x \right], \quad (x,t) \in \mathcal{S} \times [0, H], \tag{2}$$

as well as the *optimal* value function (i.e. the maximum mean accumulated reward)

$$V^*(x,t) := \sup_{\pi \in \Pi} V^\pi(x,t), \quad (x,t) \in \mathcal{S} \times [0, H], \tag{3}$$

where Π denotes the set of all deterministic Markov policies. Because both the state and action spaces are finite, the existence of optimal policies is guaranteed under Assumption 1 (Mamer 1986).

Consider an agent who repeatedly interacts with an unknown finite-horizon CTMDP $\mathcal{M}$ over a total of K episodes, where the time length of each episode is H. For simplicity we assume the functional form of the reward function $(r(x,a)))_{s \in \mathcal{S}, a \in \mathcal{A}}$ is known, but the transition probabilities $(p(\cdot | x, a))_{x \in \mathcal{S}, a \in \mathcal{A}}$ and the rates of exponential holding times $(\lambda(x,a))_{x \in \mathcal{S}, a \in \mathcal{A}}$ are unknown. Extending our algorithm and analysis in this paper to the case of unknown random rewards is not difficult;

see, e.g., Zanette and Brunskill (2019) for a standard treatment. In each episode $k = 1, 2, \ldots, K$, an arbitrary fixed initial state $x_0^k = x_0$ is picked for the CTMDP[3]. An algorithm **algo** initializes and implements a policy π_1 for the first episode, and computes and executes policy π_k throughout episode k based on past data up to the end of episode $k-1$, $k = 2, \ldots, K$. The cumulative expected regret of **algo** over K episodes of interactions with $\mathcal{M}$ is defined by

$$\text{Regret}(\mathcal{M}, \textbf{algo}, K) := \mathbb{E}\left[\sum_{k=1}^{K} V^*(x_0^k, H) - \sum_{k=1}^{K} V^{\pi_k}(x_0^k, H)\right]. \tag{4}$$

The worst-case regret of **algo** is defined by $\sup_{\mathcal{M} \in \mathcal{C}} \text{Regret}(\mathcal{M}, \textbf{algo}, K)$, where $\mathcal{C}$ is the set of all CTMDPs with S states, A actions, horizon $[0, H]$, and the holding-time rates and rewards satisfying Assumption 1. Our objective is to obtain a bound on the worst-case regret that scales sublinearly in K while having polynomial dependance on the other parameters including S, A, H and $\lambda_{\max}$.

Remark 1. *In this paper, we assume the transition probability $p(\cdot|\cdot, \cdot)$, reward function $r(\cdot, \cdot)$ and the holding time rate $\lambda(\cdot, \cdot)$ to be all time-invariant, i.e., the model is time homogeneous. This assumption simplifies the parameter estimation, algorithm development and regret analysis. For episodic learning of discrete-time tabular MDPs, extending the results from the homogeneous setting to nonhomogeneous one is straightforward. For instance, one can still use the empirical transition probability to estimate the time-dependent transition probability $p_h(y|x, a)$ where $h = 1, 2, \ldots, H$ for a discrete-time tabular MDP with horizon length H steps. However, it is much more challenging to extend our current results to nonhomogeneous CTMDPs, precisely because the continuity of the time parameter. For illustrations, consider the time-dependent transition probability $p_t(y|x, a)$ with $t \in [0, H]$ for fixed x, a, y in a CTMDP. This is an infinite-dimensional quantity due to $t \in [0, H]$. In addition, there are only finite (random) number of jumps in each episode of the CTMDP under Assumption 1. This implies for almost every $t \in [0, H]$, one will not observe any transition occurred exactly at time t in the episodic learning of CTMDPs with K episodes. Therefore, it is impossible to estimate $p_t(y|x, a)$ reliably for all $t \in [0, H]$, unless one imposes strong assumptions such as special structures of the function $p_t(y|x, a)$. This further illustrates the fundamental difference between episodic learning of CTMDPs and DTMDPs, even for model-based learning. Hence, in this paper we restrict our CTMDP model to the homogeneous setting. Nevertheless, we believe considering episodic RL for nonhomogeneous CTMDP models is a very interesting topic worthy of future research.*

2.1 Model-based dynamic programming approach

Let us recall the dynamic programming approach for the model-based planning problem, i.e., solving the finite-horizon CTMDP (3) when all the model parameters are known.

It has been established that the optimal value function (3) satisfies an optimality condition via

---

[3]The results of the paper can also be extended to the case where the initial states are drawn from a fixed distribution over $\mathcal{S}$.

the following Hamilton–Jacobi–Bellman (HJB) equation

$$\frac{\partial v(x,t)}{\partial t} = \sup_{a \in \mathcal{A}} \left[r(x,a) - \lambda(x,a)v(x,t) + \lambda(x,a) \sum_{z \in \mathcal{S}} v(z,t)p(z|x,a) \right], \quad (x,t) \in \mathcal{S} \times (0,H], \quad (5)$$

$$v(x,0) = 0;$$

see, e.g., Theorem 1 in Miller (1968) or Proposition 4.1 in Guo et al. (2015). Note that this is a system of ordinary differential equations (ODEs) as the state space $\mathcal{S}$ is finitely discrete. The model-based approch is to first solve (5) offline and then obtain the optimal policy by maximizing the right hand side of (5). However, it is impossible to adapt this ODE approach to the learning setting because (5) is unspecified when its coefficients are unknown. Instead, we choose an alternative optimality condition that gives rise to a value iteration algorithm for computing the optimal value function V^*. This alternative condition is equivalent to the HJB equation in the model-based setting; yet its idea can be carried over to the RL setting.

To this end, we first introduce some notations. Denote by $\mathbb{M}$ the Banach space of all real-valued measurable functions $u : \mathcal{S} \times [0,H] \to [0,\infty)$ satisfying $||u||_\infty := \sup_{(x,t) \in \mathcal{S} \times [0,H]} u(x,t) < \infty$ and $u(x,0) \equiv 0$ for all $x \in \mathcal{S}$. Given a CTMDP $\mathcal{M}$ in (1), we define two associated operators $\mathcal{T}^a$ and G from $\mathbb{M}$ to $\mathbb{M}$ as follows: for $u \in \mathbb{M}$, $(x,t) \in \mathcal{S} \times [0,H]$ and $a \in \mathcal{A}$,

$$(\mathcal{T}^a u)(x,t) := r(x,a) \cdot \int_0^t e^{-\lambda(x,a)s}ds + \sum_y \int_0^t \lambda(x,a)e^{-\lambda(x,a)s}u(y,t-s)p(y|x,a)ds$$

$$= r(x,a) \cdot \mathbb{E}_{\tau \sim \exp(\lambda(x,a))}[t \wedge \tau] + \mathbb{E}_{\tau \sim \exp(\lambda(x,a)), Y \sim p(\cdot|x,a)}[u(Y,(t-\tau)^+)], \quad (6)$$

where $t \wedge \tau = \min\{t,\tau\}$ and $\exp(\lambda)$ denotes an exponential distribution with rate λ, and

$$(Gu)(x,t) := \sup_{a \in \mathcal{A}}(\mathcal{T}^a u)(x,t).$$

Then we have the following results; see Theorem 1 in Mamer (1986) and Theorems 3.1 and 3.2 in Huang and Guo (2011).

- (Optimality equation) The optimal value function V^* in (3) is the unique solution in $\mathbb{M}$ to the Bellman optimality equation $V^* = GV^*$.

- (Verification) The function $a^*(x,t) := \arg\max_{a \in \mathcal{A}}(\mathcal{T}^a V^*)(x,t)$, that maps a state–time pair to an action, is an optimal policy for (3).

- (Value iteration) Let $V_0^* := 0$ and $V_{n+1}^* := GV_n^*$. Then $V_{n+1}^* \geq V_n^*$ and $\lim_{n \to \infty} ||V_n^* - V^*||_\infty = 0$.

Intuitively, one can view V_n^* as the optimal value function of a modified control problem in which the objective is to maximize the expected total reward from the first n transitions or up to time H, whichever comes first (Mamer 1986). The results above will play an essential role in the design and analysis of our RL algorithm for episodic CTMDPs. In particular, the value iteration naturally introduces a sequence of numbers – those of the iterations – and makes it possible to apply induction arguments to the continuous-time setting; such induction arguments have been critical in regret analysis on episodic RL algorithms for discrete-time tabular MDPs.

2.2 Issues with upfront time discretization

Before we proceed, let us discuss an "obvious" approach for treating a CTMDP – one that first discretizes time and then applies existing RL algorithms to the resulting DTMDP. There are several issues with this approach as we discuss below.

Consider a uniform time discretization of the horizon $[0, H]$ with discretization step size $\Delta > 0$. For the simplicity of discussion, we assume H/Δ is an integer. Then the CTMDP model is approximated by a DTMDP:

$$\sup_\pi V_\Delta^\pi(x, H) := \mathbb{E}^\pi \left[\sum_{j=0}^{H/\Delta-1} r(X(j\Delta), A(j\Delta))|X(0) = x \right] \cdot \Delta, \tag{7}$$

where $r(x, a) \in [0, 1]$ for all x, a, and the value function $V_\Delta^\pi(x, H)$ approximates (2) for the CTMDP when Δ is small. Importantly, the number of decision steps in this DTMDP is $\frac{H}{\Delta}$, while in the original CTMDP the expected number of decision steps is only $O(H)$ due to Assumption 1. One can apply existing learning algorithms for DTMDP to this model, which however will lead to vacuous regret bounds as $\Delta \to 0$. To see this, consider for example the UCBVI–BF algorithm with Bernstein bonus (Azar et al. 2017), which has a regret bound $\tilde{O}(\bar{H}\sqrt{SAK} + \bar{H}^2 S^2 A + \bar{H}^{1.5}\sqrt{K})$ for a DTMDP with $\bar{H}$ decision steps and stationary transitions, where K is the number of episodes. This bound matches the regret lower bound for learning tabular DTMDP when $K \geq \bar{H}^2 S^3 A$ and $SA > H$ (Domingues et al. 2021). Applying UCBVI–BF to the episodic learning of DTMDP (7), we have

$$\text{Regret}^\Delta := \mathbb{E} \left[KV_\Delta^*(x_0, H) - \sum_{k=0}^{K-1} V_\Delta^{\pi_k}(x_0, H) \right]$$

$$= \Delta \cdot \tilde{O} \left(\frac{H}{\Delta}\sqrt{SAK} + \left(\frac{H}{\Delta} \right)^2 S^2 A + \left(\frac{H}{\Delta} \right)^{1.5} \sqrt{K} \right),$$

where V_Δ^* is the optimal value function for the DTMDP (7) and π_k is the executed policy in episode k under the UCBVI–BF algorithm. It is clear that the bound for Regret^Δ grows to ∞ when $\Delta \to 0$. Considering alternative learning algorithms such as EULER in Zanette and Brunskill (2019) for DTMDPs leads to a similar explosion of regret bound.

Another issue with the upfront discretization approach is that the error bound between the value function of the CTMDP and that of the DTMDP from time-discretization is in general unknown, i.e., it is hard to compute $V^\pi - V_\Delta^\pi$ with explicit constants. Hence, while one may expect that Regret^Δ converges to the regret of the CTMDP model when $\Delta \to 0$, one is generally unable to quantify the gap explicitly so as to give information in S, A, H and the bounds on the holding time rates $\lambda_{\max}$ and $\lambda_{\min}$.

Finally, from the implementation perspective, it is challenging to select the discretization step size Δ properly. Indeed, it is well known in the RL community that the performance of RL algorithms can be very sensitive with respect to discreteization step size; see, e.g., Tallec et al. (2019) in which it is empirically shown that standard $Q-$learning methods are not robust to changes in time discretization of continuous-time control problems.

3 The CT-UCBVI Algorithm

In this section, we introduce an RL algorithm, named the CT-UCBVI (continuous-time upper confidence bound value iteration) algorithm, for learning episodic tabular CTMDPs. It is a continuous-time analogue of the UCBVI algorithm developed in Azar et al. (2017) for learning discrete-time MDPs.

The CT-UCBVI algorithm updates the estimates of model parameters $(\lambda(\cdot, \cdot), p(\cdot|\cdot, \cdot))$ at the beginning of each episode. Let $(x_n^l, a_n^l, \tau_n^l)_n$ be the trajectories of the state, action and holding time in episode l of the CTMDP, and $t_n^l = \sum_{i=0}^{n-1} \tau_i^l$ where the discrete index n counts the number of jumps/transitions occurred in this episode; see Section 2. A state–action–state triplet (x, a, y) means that the process is in x, takes an action a and then moves to y. Similarly, a state–action pair (x, a) signifies that the process is in x and takes an action a.

At the start of episode k, we set the accumulated duration at (x, a) up to the end of episode $k - 1$ as $T^k(x, a)$, for $x \in \mathcal{S}$ and $a \in \mathcal{A}$. Mathematically, it is given by

$$T^k(x, a) = \sum_{n=0}^{\infty} \sum_{l=1}^{k-1} \tau_n^l \cdot 1_{\{(x_n^l, a_n^l) = (x, a)\}} \cdot 1_{t_n^l \leq H}. \tag{8}$$

Here, the indicator function on the event $\{t_n^l \leq H\}$ ensures that the final jumping time of the CTMDP in episode l does not exceed the end of the horizon, H. Note that the observation of the 'last' holding time in each episode may be truncated by H, i.e., if $t_n^l \leq H < t_{n+1}^l$, then the actual observed holding time is $\tau_n^l = H - t_n^l$ instead of $t_{n+1}^l - t_n^l$. In this case, the truncated τ_n^l is *not* a sample of an exponential distribution, noting we have slightly abused the notation for τ_n^l here.

We also set the observed accumulated visit counts to (x, a, y) up to the end of episode $k - 1$ as $N^k(x, a, y)$ for $x, y \in \mathcal{S}$ and $a \in \mathcal{A}$. Note that x_{n+1}^l is *not* observed if $t_n^l < H < t_{n+1}^l$.[4] Hence, for the purpose of estimating parameters of the CTMDP, we also introduce $N^{k,+}(x, a)$, which differs from from the cumulative visit counts to (x, a) up to the end of episode $k - 1$ in that does not include the visit count to (x, a) if the next state is not observed due to the finite horizon. Specifically, we let

$$N^k(x, a, y) = \sum_{n=0}^{\infty} \sum_{l=1}^{k-1} 1_{\{(x_n^l, a_n^l, x_{n+1}^l) = (x, a, y)\}} \cdot 1_{t_{n+1}^l \leq H}, \tag{9}$$

$$N^{k,+}(x, a) = \sum_{n=0}^{\infty} \sum_{l=1}^{k-1} 1_{\{(x_n^l, a_n^l) = (x, a)\}} \cdot 1_{t_{n+1}^l \leq H}. \tag{10}$$

At the beginning of episode k, we compute the empirical transition probabilities by

$$\hat{p}_k(y|x, a) := \frac{N^k(x, a, y)}{\max\{1, N^{k,+}(x, a)\}}, \tag{11}$$

as well as an estimator for the holding-time rate $\lambda(x, a)$ by

$$\hat{\lambda}_k(x, a) := \min\left\{ \frac{N^{k,+}(x, a)}{T^k(x, a)}, \lambda_{\max} \right\}. \tag{12}$$

---

[4]This seemingly innocent subtlety actually causes a substantial difficulty in our regret analysis below.

If (x, a) has not been sampled prior to episode k, we define $\hat{p}_k(y|x, a) = 0$ and $\hat{\lambda}_k(x, a) = 0$ for all $y \in \mathcal{S}$. [5] In particular, when $k = 1$, there is no data available so $\hat{p}_1(y|x, a) = 0$ and $\hat{\lambda}_1(x, a) = 0$. Note that we truncate the empirical holding-time rate $\frac{N^{k,+}(x,a)}{T^k(x,a)}$ by $\lambda_{\max}$ – while this is natural under our assumption $\lambda(x, a) \le \lambda_{\max}$, the truncation turns out also important in our regret analysis below.

The CT-UCBVI algorithm encourages exploration by awarding some 'bonus' for exploring certain state–action pairs in learning CTMDPs. The bonus, which is a function of (x, a), is constructed based on the level of uncertainty in learned transition probabilities (11) and holding time rates (12) to incentivize the agent to explore those pairs with high uncertainty. Mathematically, the bonus function b^k for each episode k is defined by

$$b^k(x, a, \delta) = \left(\frac{\lambda_{\max}}{1 - e^{-\lambda_{\max}H}} \vee 1 \right) \cdot \left(H^2 \sqrt{\frac{\lambda_{\max}L(\delta)}{\max\{T^k(x,a), L(\delta)/\lambda_{\max}\}}} + H \sqrt{\frac{2[S\ln 2 + \ln(SAHK^2/\delta)]}{\max\{1, N^{k,+}(x,a)\}}} \right),$$
(13)

where $(x, a) \in \mathcal{S} \times \mathcal{A}$, $\delta \in (0, 1)$ is an input parameter in our learning algorithm, and $L(\delta) := 4\ln(2SAK/\delta)$.

Let us provide some explanations on the expression of the bonus $b^k(x, a, \delta)$. In (13), the term $\sqrt{\frac{\lambda_{\max}L(\delta)}{\max\{T^k(x,a), L(\delta)/\lambda_{\max}\}}} := \beta_k(x, a, \delta)$ is the confidence bound for the holding time rate (see Lemma 1 in Section 6.1), and the term $\sqrt{\frac{2[S\ln 2 + \ln(SAHK^2/\delta)]}{\max\{1, N^{k,+}(x,a)\}}} := \alpha_k(x, a, \delta)$ is the confidence bound for the transition probabilities (see Lemma 3 in Section 6.1). By setting the bonus $b^k(x, a, \delta)$ as in (13), one can ensure optimism in regret analysis, i.e., the computed value function from our algorithm is a high probability upper bound of the true optimal value function (up to some error). See Lemma 5 in Section 6.1 for further details.

The learning algorithm CT-UCBVI, presented as Algorithm 1, proceeds as follows. In episode $k = 1, \ldots, K$, the algorithm uses the past data (trajectories of state, action and holding time) up to the end of episode $k - 1$ to estimate the parameters of the CTMDP, and to compute the reward bonus function b^k. Then it applies the modified value iteration procedure (Algorithm 2) to the estimated model with a combined reward function $r + b^k$, and executes the returned policy π^k from the value iteration in episode k. Finally, it observes the data collected in episode k and updates the data for model estimation in the next episode. The algorithm repeats the above procedure for K episodes.

Meanwhile, the subroutine, Algorithm 2, is adapted from the value iteration scheme developed in Huang and Guo (2011). The main modification needed here is that we truncate the value function by t in the algorithm. Note that because the reward $r(x, a) \in [0, 1]$, the value $V^{\pi}(x, t)$ of any policy π under the true CTMDP model is bounded by t. Such a truncation trick is standard in the discrete-time RL literature (see, e.g., Azar et al. 2017, Jin et al. 2018) which facilitates the regret analysis.

---

[5]Strictly speaking, $\hat{p}_k(\cdot|x, a)$ is not a probability vector on $\mathcal{S}$ in this case. However, this will not affect our analysis because we set $\hat{\lambda}_k(x, a) = 0$ when (x, a) has not been sampled before, see (15).

Algorithm 1 The CT-UCBVI Algorithm

Input: Parameters $\delta, \epsilon_k \in (0, 1)$ for $k \geq 1$, $\lambda_{\max}, \mathcal{S}, \mathcal{A}, H$ and reward function r.

1: Initialization: set initial state x_0;
2: **for** episode $k = 1, 2, \ldots, K$ **do**
3:     Past data $\mathcal{D} = \{(x_n^l, a_n^l, \tau_n^l)_{n \geq 0} : l < k\}$;
4:     For all $(x, a, y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:
5:     Compute the empirical transition probabilities $\hat{p}_k(y|x, a)$ by (11);
6:     Compute the estimated holding-time rate $\hat{\lambda}_k(x, a)$ by (12);
7:     Compute the reward bonus $b^k(x, a, \delta)$ by (13);
8:     Apply modified value iteration (Algorithm 2) to the CTMDP model $(\mathcal{S}, \mathcal{A}, r(\cdot, \cdot) + b^k(\cdot, \cdot, \delta), \hat{\lambda}_k(\cdot, \cdot), \hat{p}_k(\cdot|\cdot, \cdot), H, x_0)$ with accuracy parameter ϵ_k and return value function $\hat{V}_{n_k}^k$;
9:     Compute policy π^k:

$$\pi^k(x, t) = \arg\max_{a \in \mathcal{A}} \{\mathcal{T}^{a,k} \hat{V}_{n_k}^k(x, t)\}, \quad \text{for all } (x, t) \in \mathcal{S} \times [0, H], \tag{14}$$

    where

$$\mathcal{T}^{a,k} \hat{V}_{n_k}^k(x, t) = [r(x, a) + b^k(x, a, \delta)] \cdot \int_0^t e^{-\hat{\lambda}_k(x,a)s} ds$$

$$+ \sum_y \int_0^t \hat{\lambda}_k(x, a) e^{-\hat{\lambda}_k(x,a)s} \hat{V}_{n_k}^k(y, t - s) \hat{p}_k(y|x, a) ds; \tag{15}$$

10:     Execute π^k in episode k;
11:     Observe data $(x_n^k, a_n^k, \tau_n^k)_{n \geq 0}$ up to time H;
12: **end for**

Algorithm 2 Modified Value Iteration for CTMDP

Input: A CTMDP model $(\mathcal{S}, \mathcal{A}, r(\cdot, \cdot), \lambda(\cdot, \cdot), p(\cdot|\cdot, \cdot), H, x_0)$ and accuracy paremeter ϵ.

1: Initialization: Set $n = 0$ and $V_n(x, t) = 0$ for all $(x, t) \in \mathcal{S} \times [0, H]$;
2: Repeat
3:     $n = n + 1$;
4:     Compute

$$V_{n+1}(x, t) = \min\{t, \max_{a \in \mathcal{A}} \mathcal{T}^a V_n(x, t)\}, \quad \text{for all } (x, t) \in \mathcal{S} \times [0, H]; \tag{16}$$

5: Until $||V_n - V_{n-1}||_\infty < \epsilon$;
6: Return V_n.

Remark 2. *In Step 8 of the CT-UCBVI algorithm, we apply the modified value iteration to solve the CTMDP model $(\mathcal{S}, \mathcal{A}, r(\cdot, \cdot) + b^k(\cdot, \cdot, \delta), \hat{\lambda}_k(\cdot, \cdot), \hat{p}_k(\cdot|\cdot, \cdot), H, x_0)$ approximately. An alternative method might be to time-discretize the HJB equation (5) and solve it numerically to obtain an approximated value function. The problem with this approach for regret analysis, however, is that the error resulting from time–discretization is unknown.*

Remark 3. *In the CT-UCBVI algorithm, we assume that the type of integral in (15) can be explicitly computed for every $t \in [0, H]$. Then the policy $\pi^k(x, t)$ can be computed by (14) for all $(x, t) \in \mathcal{S} \times [0, H]$, because the state and action spaces are both finite. In our actual algorithm implementations in the simulation experiments, we discretize the integral and first compute $\pi^k(x, t_i)$ for $x \in \mathcal{S}$ where $0 = t_0 < t_1 < \ldots < t_N = H$. Given a state x, to compute $\pi^k(x, t)$ for all $t \in [0, H]$, we then use piecewise-constant interpolation. The approximation error can be made arbitrarily small with a small grid size.*

4 Main Results

Recall that $\mathcal{C}$ denote the set of all CTMDPs with S states, A actions, horizon $[0, H]$, and the holding-time rates and rewards satisfying Assumption 1. The main result of the paper is the following theorem, whose proof is delayed to Section 6.

Theorem 1. *The CT-UCBVI algorithm achieves the following worst-case regret bound:*

$$
\begin{aligned}
&\sup_{\mathcal{M} \in \mathcal{C}} Regret(\mathcal{M}, CT\text{-}UCBVI, K) \\
&\leq 3 \left(\left(\frac{\lambda_{\max}}{1 - e^{-\lambda_{\max} H}} \vee 1 \right) H + 1 \right) \cdot \sqrt{SAK} \cdot \left(H + \frac{1}{\lambda_{\min}} \right) \cdot (\lambda_{\max} H + 1) \frac{\ln K}{\ln(\ln(K) + 1)} \\
&\quad \cdot \left(\sqrt{2S + 6 \ln(2SAKH)} + 4\lambda_{\max} H \sqrt{\ln(2SAKH)} \right) \\
&\quad + \sum_{k=1}^{K} \epsilon_k \cdot e^{\lambda_{\max} H} + 1.
\end{aligned}
\tag{17}
$$

where ϵ_k, $k = 1, 2, \cdots, K$, are the accuracy parameters given in Algorithm 1.

Theorem 1 immediately implies the following result.

Corollary 1. *If $\lambda_{\max}, H > 1$ and $\epsilon_k = 1/k^{\alpha} \cdot e^{-\lambda_{\max} H}$ for $\alpha \geq 1/2$, then*

$$
\sup_{\mathcal{M} \in \mathcal{C}} Regret(\mathcal{M}, CT\text{-}UCBVI, K) = \tilde{O} \left(\lambda_{\max}^2 H^2 \left(H + \frac{1}{\lambda_{\min}} \right) \sqrt{SAK} \left[\sqrt{S} + \lambda_{\max} H \right] \right),
\tag{18}
$$

where the notation $\tilde{O}$ ignores the logarithmic factors in H, S, A and K.

Theorem 1 is complemented by Theorem 2 below, which gives a regret lower bound. We first state the following assumption.

Assumption 2. *The number of states and actions satisfy $S \geq 6, A \geq 2$, and there exists an integer d such that $S = 2 + \frac{A^d - 1}{A - 1}$.*

Assumption 2 is adapted from Domingues et al. (2021) (see Assumption 1 therein), where a rigorous proof of the minimax regret lower bound for episodic RL in discrete-time tabular MDPs is provided. [6] Note that $S = 2 + \frac{A^d - 1}{A - 1} \equiv 2 + \sum_{i=0}^{d-1} A^i$. Hence, one can use $S - 2$ states to construct a *full A-ary tree* of depth $d - 1$. In this rooted tree, each node has exactly A children and the total number of nodes is given by $\sum_{i=0}^{d-1} A^i = S - 2$. This tree construction allows one to design difficult CTMDP instances that are needed for proving the regret lower bound. Assumption 2 is imposed here for the purpose of a cleaner analysis, and it can be relaxed following the discussion in Appendix D of Domingues et al. (2021).

The proof of the following result is given in Section 6.

Theorem 2. *Under Assumption 2, for any algorithm **algo** there exists an $\mathcal{M} \in \mathcal{C}$ such that*

$$Regret(\mathcal{M}, \textbf{\textit{algo}}, K) \geq \frac{1}{12\sqrt{2}} \cdot \mathbb{E}[(H - \gamma_d)^+]\sqrt{SAK}, \quad \text{for all } K \geq SA/2, \tag{19}$$

where γ_d is a random variable following an Erlang distribution with rate parameter $\lambda_{\max}$ and shape parameter d. Moreover, if $d < (1 - c_2)\lambda_{\max}H$ for some universal constant $c_2 \in (0, 1)$, then

$$Regret(\mathcal{M}, \textbf{\textit{algo}}, K) \geq \frac{c_2}{12\sqrt{2}} \cdot H\sqrt{SAK}. \tag{20}$$

The regret upper bound in Theorem 1 and the lower bound in Theorem 2 suggest that our proposed CT-UCBVI algorithm achieves a regret rate with the optimal dependence on the number of episodes K as well as on the number of actions A, up to logarithmic factors. While the bounds on K are the most important as they inform the convergence rates of learning algorithms, it remains a significant open question whether one can improve the dependence of these bounds on S, H and $\lambda_{\max}$ to narrow down the gap between the upper and lower bounds. For episodic RL in discrete-time tabular MDPs with stationary transitions, the UCBVI–BF algorithm (with Bernstein's bonuses) in Azar et al. (2017) has a regret bound of $\tilde{O}(H\sqrt{SAK} + H^2 S^2 A + H\sqrt{HK})$. When $K \geq H^2 S^3 A$ and $SA \geq H$, the leading term in their regret bound is $\tilde{O}(H\sqrt{SAK})$, which matches the established lower bound of $\Omega(H\sqrt{SAK})$ for discrete-time episodic MDPs, up to logarithmic factors (Domingues et al. 2021). Our upper bound in Theorem 1 scales with S linearly. Compared with the leading term $\tilde{O}(H\sqrt{SAK})$ in Azar et al. (2017), the extra $\sqrt{S}$ factor comes from the fact that we apply concentration inequalities to the transition probability vector which is S-dimensional; see Lemma 3. Azar et al. (2017) instead maintain confidence intervals on the optimal value function directly; see also Dann et al. (2019), Zanette and Brunskill (2019) for similar approaches. The primary difficulty in extending the method of Azar et al. (2017) to our continuous-time setting is that the number of decision steps in each episode is random and different, and unbounded in general. In addition, due to the inherent random holding times, our regret upper bound has a worse dependance on H compared with the discrete-time counterparts. Azar et al. (2017) consider Bernstein-type "exploration bonuses" that use the empirical variance of the estimated value function at the next state to achieve tight (linear) dependance on H (in the leading term of their regret bound). It is an open problem how to extend their idea and technique to our setting in order to improve the dependancy of the upper bound on H.

---

[6]Assumption 1 of Domingues et al. (2021) requires $S = 3 + \frac{A^d - 1}{A - 1}$ because non-homogeneous transitions are considered there, as well as $H \geq 3d$ which is not needed in our analysis.

Remark 4. *If we assume the constant $\lambda_{\min} > 0$ in Assumption 1 is also known to the agent, then we may improve the H dependance in the regret upper bound. Specifically, we can truncate the empirical holding time rate (12) so that the new estimator $\hat{\lambda}_k(x, a) \geq \lambda_{\min}$ for all (x, a). Then in the proof of Lemma 6, the integral $\int_0^{(H-t_m^k)^+} e^{-\hat{\lambda}_k(x_0^k, a_0^k)s} ds$ (c.f. (40)) can be upper bounded by $\frac{1}{\lambda_{\min}}(1 - e^{-\lambda_{\min}(H-t_m^k)^+})$, instead of the quantity $(H - t_m^k)^+$. One can then verify that when $\lambda_{\max}, H > 1$, the regret bound in (18) can be improved to*

$$\tilde{O}\left(\frac{\lambda_{\max}^2 H^2}{\lambda_{\min}}\sqrt{SAK}\left[\sqrt{S} + \lambda_{\max}H\right]\right).$$

5 Simulation Results

In this section, we demonstrate the performance of the CT-UCBVI algorithm by simuations. As there are no existing benchmarks for episodic RL in CTMDPs to the best of our knowledge, we will take a continuous-time machine operation and repair problem taken from Puterman (2014).

We first discuss the numerical computation of the expected regret in (4) given a CTMDP. First, we need to compute $V^*(x_0, H)$, the optimal value function of the CTMDP when the model parameters are known. This can be done by applying the value iteration procedure in Algorithm 2, except that there is no need to consider the truncation by t in (16) in this case because it holds automatically that $V^*(x_0, t) \leq t$ for all $t \in [0, H]$. Second, we need to numerically compute $V^{\pi_k}(x_0, H)$, which is the cumulative reward collected up to time H in the known environment under the policy π^k for each episode $k = 1, 2, \ldots, K$. This is a problem of policy evaluation for the CTMDP. Lemma 3.1 in Huang and Guo (2011) yileds that

$$V^{\pi_k}(x, t) = \mathcal{T}^{\pi^k} V^{\pi_k}(x, t), \quad \text{for all } (x, t) \in \mathcal{S} \times [0, H],$$

where, with a slight abuse of notation (compare with (6)), the operator $\mathcal{T}^{\pi^k}$ is defined by

$$\mathcal{T}^{\pi^k} u(x, t) = r(x, \pi_k(x, t)) \cdot \int_0^t e^{-\lambda(x, \pi_k(x, t))s} ds$$
$$+ \sum_y \int_0^t \lambda(x, \pi_k(x, t)) e^{-\lambda(x, \pi_k(x, t))s} u(y, t - s) p(y | x, \pi_k(x, t)) ds.$$

Hence we can apply the following iterative scheme to compute $V^{\pi_k}(x, t)$ for $(x, t) \in \mathcal{S} \times [0, H]$:

- Set $u_0(x, t) = 0$ for all $(x, t) \in \mathcal{S} \times [0, H]$;

- Compute

$$u_{n+1}(x, t) = \mathcal{T}^{\pi^k} u_n(x, t), \quad \text{for } (x, t) \in \mathcal{S} \times [0, H],$$

until $||u_{n+1} - u_n||_\infty < \epsilon$ for some pre-specified accuracy parameter ϵ;

- Return u_n as an approximation of V^{π_k}.

Once we know how to compute V^* and V^{π_k}, we can then compute the expected regret in (4) by averaging the regret over multiple independent runs.

We take Problem 11.2 of Puterman (2014) for our simulation experiment. A machine is either in state 0 representing it is operating, or in state 1 signifying it is under repair. In each state, the decision maker can choose an action from $\mathcal{A} := \{\text{slow, fast}\} \equiv \{a_1, a_2\}$. Specifically, if the machine is in state 0 and a_1 is chosen, then the machine remains in the operating mode for an exponentially distributed amount of time with the failure rate $\lambda(0, a_1) = 3$ and the reward rate $r(0, a_1) = 5$; if a_2 is chosen, then the failure rate is $\lambda(0, a_2) = 5$ and the reward rate $r(0, a_2) = 8$. Hence, when the machine operates at the fast rate, it yields higher reward but breaks down more frequently than when it operates at the slow rate. On the other hand, when the machine is in state 1 and action a_1 is chosen, then the repair time distribution is exponential with the recovery rate $\lambda(1, a_1) = 2$ and the reward rate $r(1, a_1) = -4$. Similarly, $\lambda(1, a_2) = 7$ and $r(1, a_2) = -12$. Here the rewards are negative representing costs. All the failure and repair times are assumed to be mutually independent. In addition, the transition probabilities for this problem are given by $p(1|0, a_i) = p(0|1, a_i) = 1$ for $i = 1, 2$, i.e., the machine alternates between the operating mode and the repair mode. The length of the horizon for this CTMDP is set to be $H = 1$, and the objective of the CTMDP is to maximize the expected cumulative reward over the horizon $[0, H]$.

The learning problem under consideration assumes that the reward rate is known, but the failure and repair rates as well as the transition probabilities are unknown to the agent. We consider the episodic RL setting where every episode restarts with a fixed initial state $x_0 = 0$. We implement the proposed CT-UCBVI algorithm with parameters $\delta = 0.05, \epsilon_k = 1/\sqrt{k}, \lambda_{\max} = 7, \mathcal{S} = \{0, 1\}, \mathcal{A} = \{a_1, a_2\}$ and $H = 1$. To fit Assumption 1, we rescale the reward rates to $[0, 1]$ by a linear transformation so that $r(0, a_1) = 17/20, r(0, a_2) = 1, r(1, a_1) = 8/20$ and $r(1, a_2) = 0$.

Figure 1 illustrates the performance of the CT-UCBVI algorithm on this example. The expected regret is computed by averaging the regret values over 30 independent runs of the algorithm. We do not show the standard deviation estimated from these 30 runs because it is very small and hardly visible in the plot. For comparison, we also plot the regret upper bound in the right hand side of (17) as a function of the number of episodes K. Because the bound in (17) is a worst-case bound whose magnitude is much greater than that of the expected regret of the algorithm on this specific instance, we plot the two curves on a log-log scale. As we can observe from Figure 1, the growth rate of the expected regret of the CT-UCBVI algorithm on this specific example is similar to that of the worst-case bound up to 10 million episodes. However, the exact rate in terms of K requires a separate instance-dependent regret analysis of the algorithm and is left for future research.

6 Proofs

In this section we present proofs of the main results, Theorems 1 and 2.

6.1 Preliminary lemmas

In this subsection, we state and prove a few lemmas needed for the main proofs.

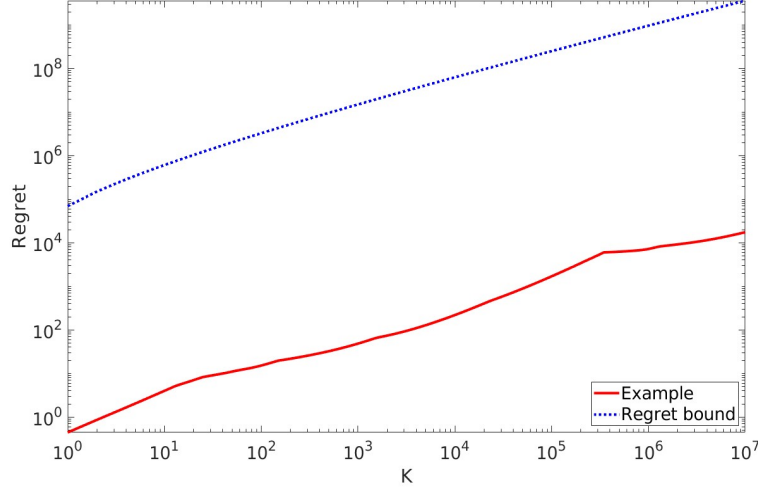The first lemma provides a high probability bound on the estimator $\hat{\lambda}_k(x, a)$ in (12) for the

Figure 1: Performance of the CT-UCBVI algorithm plotted on a log-log scale. The line "Example" represents the average regret of the CT-UCBVI algorithm on the machine replacement/repair example. The dash line "Regret bound" plots the theoretical regret upper bound in (17) as a function of the number of episodes K.

holding-time rate. Recall $L(\delta) = 4\ln(2SAK/\delta)$, and define

$$
\beta_k(x,a,\delta) := \sqrt{\frac{\lambda_{\max}L(\delta)}{\max\{T^k(x,a), L(\delta)/\lambda_{\max}\}}} = \begin{cases} \lambda_{\max} & \text{if} \quad T^k(x,a) < \frac{L(\delta)}{\lambda_{\max}}, \\ \sqrt{\frac{\lambda_{\max}L(\delta)}{T^k(x,a)}} & \text{otherwise.} \end{cases} \tag{21}
$$

Then we have the following result.

Lemma 1. *For any $\delta \in (0,1)$, we have*

$$
\mathbb{P}\left(|\lambda(x,a) - \hat{\lambda}_k(x,a)| \le \beta_k(x,a,\delta), \quad \text{for all } x,a,k\right) \ge 1 - \delta. \tag{22}
$$

Note that the samples of holding time at a given state–action pair may not be i.i.d. due to the truncation at time H. Hence, we are unable to directly apply concentration inequalities for i.i.d. exponential distributions to derive the confidence bound (22). The key idea to get around is to 'piece together' the samples of exponential holding time at a pair (x,a) as realizations of inter-arrival times of a Poisson process with rate $\lambda(x,a)$ modelling the number of visits to the pair over time. Even if the time is eventually truncated, the number of visits still follows a Poisson distribution. This enables us to apply the tail bounds for Poisson distribution, which is presented below for reader's convenience.

Lemma 2 (Tail bounds for Poisson distribution Canonne 2017)**.** *Let the random variable X follow a Poisson distribution with mean λ. Then we have*

$$
\mathbb{P}(|X - \lambda| > x) \le e^{-\frac{x^2}{2(\lambda+x)}}, \quad x > 0.
$$

17

Proof of Lemma 1. Fix x, a, k, δ. If $T^k(x, a) = 0$, i.e., the state–action pair (x, a) has not been sampled before episode k, then clearly $\hat{\lambda}_k(x, a) = 0$ and $\beta_k(x, a, \delta) = \lambda_{\max}$. Since $\lambda(x, a) \le \lambda_{\max}$ for all (x, a), it follows trivially that $|\lambda(x, a) - \hat{\lambda}_k(x, a)| \le \beta_k(x, a, \delta)$.

So we focus on the non-trivial case when $T^k(x, a) > 0$. For each episode $i < k$, we piece together the holding times at (x, a) and denote by $v_i(x, a)$ the total amount of time staying at (x, a) in episode i. Denote by $n_i(x, a)$ the total number of visits to (x, a) in episode i, excluding the visit count to (x, a) if the next state is not observed due to the finite horizon. Given $v_i(x, a)$, the variable $n_i(x, a)$ follows a Poisson distribution with mean $\lambda(x, a)v_i(x, a)$. Since the holding times at (x, a) across different episodes are independent, we obtain that given $(v_i(x, a) : i < k)$, the counts $N^{k,+}(x, a) = \sum_{i<k} n_i(x, a)$, follows a Poisson distribution with mean $\lambda(x, a)T^k(x, a)$ where $T^k(x, a) = \sum_{i<k} v_i(x, a)$.

Given $T^k(s, a) = T > 0$, $N^{k,+}(x, a)$ follows a Poisson distribution with mean $\lambda(x, a)T$. We now estimate the following conditional probability, noting (12):

$$
\begin{aligned}
&\mathbb{P}(|\hat{\lambda}_k(x, a) - \lambda(x, a)| > \beta_k(x, a, \delta)|T^k(s, a) = T) \\
&= \mathbb{P}(|N^{k,+}(x, a)/T^k(x, a) - \lambda(x, a)| > \beta_k(x, a, \delta), N^{k,+}(x, a)/T^k(x, a) \le \lambda_{\max}|T^k(s, a) = T) \\
&\quad + \mathbb{P}(|\lambda_{\max} - \lambda(x, a)| > \beta_k(x, a, \delta), N^{k,+}(x, a)/T^k(x, a) \ge \lambda_{\max}|T^k(s, a) = T) \\
&\le \mathbb{P}(|N^{k,+}(x, a)/T^k(x, a) - \lambda(x, a)| > \beta_k(x, a, \delta)|T^k(s, a) = T) \\
&\quad + \mathbb{P}(N^{k,+}(x, a)/T^k(x, a) - \lambda(x, a) \ge \lambda_{\max} - \lambda(x, a), \lambda_{\max} - \lambda(x, a) > \beta_k(x, a, \delta)|T^k(s, a) = T) \\
&\le 2\mathbb{P}(|N^{k,+}(x, a)/T^k(x, a) - \lambda(x, a)| > \beta_k(x, a, \delta)|T^k(s, a) = T). \tag{23}
\end{aligned}
$$

Recall that $\beta_k(x, a, \delta) = \sqrt{\frac{\lambda_{\max}L(\delta)}{T^k(x,a)}} < \lambda_{\max}$ when $T^k(x, a) > \frac{L(\delta)}{\lambda_{\max}}$. Applying Lemma 2, we obtain for $T > \frac{L(\delta)}{\lambda_{\max}}$,

$$
\begin{aligned}
&\mathbb{P}(|N^{k,+}(x, a)/T^k(x, a) - \lambda(x, a)| > \beta_k(x, a, \delta)|T^k(s, a) = T) \\
&= \mathbb{P}(|N^{k,+}(x, a) - \lambda(x, a)T| > \beta_k(x, a, \delta)T|T^k(s, a) = T) \\
&\le \exp\left(-\frac{(\beta_k(x, a, \delta)T)^2}{2(\lambda(x, a)T + \beta_k(x, a, \delta)T)}\right) \\
&\le \exp\left(-\frac{(\frac{\lambda_{\max}L(\delta)}{T})T}{2(\lambda(x, a) + \lambda_{\max})}\right) \\
&\le e^{-L(\delta)/4} = \frac{\delta}{2SAK},
\end{aligned}
$$

where we have used the facts that $\lambda(x, a) \le \lambda_{\max}$ and $\beta_k(x, a, \delta) \le \lambda_{\max}$ for all (x, a) in the last two inequalities. Together with (23), the above inequality yields that for $T > \frac{L(\delta)}{\lambda_{\max}}$,

$$
\mathbb{P}(|\hat{\lambda}_k(x, a) - \lambda(x, a)| > \beta_k(x, a, \delta)|T^k(s, a) = T) \le \frac{\delta}{SAK}. \tag{24}
$$

This inequality actually also holds for $T \le \frac{L(\delta)}{\lambda_{\max}}$. To see this, first note that in this case we have $\beta_k(x, a, \delta) = \lambda_{\max}$ by the definition (21). Hence, it follows from (12) that, with probability one, $|\hat{\lambda}_k(x, a) - \lambda(x, a)| \le \lambda_{\max}$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$.

18

Define the following two events

$$E = \left\{ \left| \hat{\lambda}_k(x,a) - \lambda(x,a) \right| \leq \beta_k(x,a,\beta), \quad \text{for all } x, a, k \right\},$$
$$E_{k,x,a} = \left\{ \left| \hat{\lambda}_k(x,a) - \lambda(x,a) \right| \leq \beta_k(x,a,\beta) \right\}.$$

For fixed k, x, a, δ, (24) implies

$$\mathbb{P}(E_{k,x,a}^c) = \mathbb{E}[\mathbb{P}(E_{k,x,a}^c | T^k(x,a))] \leq \frac{\delta}{SAK}.$$

Applying a union bound, we deduce that $\mathbb{P}(E^c) \leq \sum_{k,s,a} P(E_{k,s,a}^c) \leq SAK \cdot \frac{\delta}{SAK} = \delta$. The proof is complete. $\square$

The next lemma, which is a known result, gives L_1 concentration bounds for the empirical transition probabilities. Write

$$\alpha_k(x,a,\delta) := \sqrt{\frac{2[S \ln 2 + \ln(SAHK^2/\delta)]}{\max\{1, N^{k,+}(x,a)\}}}. \tag{25}$$

Lemma 3 (Lemma 17 in Jaksch et al. 2010). *For any $\delta \in (0,1)$, we have*

$$\mathbb{P}\left(\|p(\cdot|x,a) - \hat{p}_k(\cdot|x,a)\|_1 \leq \alpha_k(x,a,\delta), \quad \text{for all } x, a, k \right) \geq 1 - \delta. \tag{26}$$

Define

$$\mathcal{G} = \left\{ |\lambda(x,a) - \hat{\lambda}_k(x,a)| \leq \beta_k(x,a,\delta) \text{ and } \|p(\cdot|x,a) - \hat{p}_k(\cdot|x,a)\|_1 \leq \alpha_k(x,a,\delta), \text{ for all } x, a, k \right\}, \tag{27}$$

which is the event that all the estimated parameters lie in the confidence bounds. Lemmas 1 and 3 yield that $\mathbb{P}(\mathcal{G}^c) \leq 2\delta$. With the concentration bounds of the estimated parameters, we next bound the error in applying the operator $\mathcal{T}^a$ defined in (6) due to model estimation errors.

Lemma 4. *Fix $t \leq H$ and a function $f(\cdot, \cdot) : \mathcal{S} \times [0, H] \to [0, H]$ satisfying $0 \leq f(x,t) \leq t$ for all x, t and $f(\cdot, 0) = 0$. Then, conditional on the occurrence of the event $\mathcal{G}$, we have*

$$\left| \mathbb{E}_{\tau \sim \exp(\hat{\lambda}_k(x,a)), Y \sim \hat{p}_k(\cdot|x,a)}[f(Y, (t-\tau)^+)] - \mathbb{E}_{\tau \sim \exp(\lambda(x,a)), Y \sim p(\cdot|x,a)}[f(Y, (t-\tau)^+)] \right|$$

$$= \left| \sum_{y \in \mathcal{S}} \int_0^t \hat{\lambda}_k(x,a) e^{-\hat{\lambda}_k(x,a)s} f(y, t-s) \hat{p}_k(y|x,a) ds - \sum_{y \in \mathcal{S}} \int_0^t \lambda(x,a) e^{-\lambda(x,a)s} f(y, t-s) p(y|x,a) ds \right|$$

$$\leq \frac{t^2}{2} \beta_k(x,a,\delta) + t\alpha_k(x,a,\delta), \tag{28}$$

for all $(x,a) \in \mathcal{S} \times \mathcal{A}$ and $k = 1, \ldots, K$, where β_k and α_k are given in (21) and (25) respectively.

Proof of Lemma 4. We estimate

$$\left| \sum_y \int_0^t \hat{\lambda}_k(x,a) e^{-\hat{\lambda}_k(x,a)s} f(y, t-s) \hat{p}_k(y|x,a) ds - \sum_y \int_0^t \lambda(x,a) e^{-\lambda(x,a)s} f(y, t-s) p(y|x,a) ds \right|$$

$$\leq \left| \sum_y \int_0^t [\hat{\lambda}_k(x,a) e^{-\hat{\lambda}_k(x,a)s} - \lambda(x,a) e^{-\lambda(x,a)s}] f(y, t-s) \hat{p}_k(y|x,a) ds \right|$$

$$+ \left| \sum_y \int_0^t \lambda(x,a) e^{-\lambda(x,a)s} f(y, t-s) [\hat{p}_k(y|x,a) - p(y|x,a)] ds \right|$$

$$\leq \int_0^t \left| \hat{\lambda}_k(x,a) e^{-\hat{\lambda}_k(x,a)s} - \lambda(x,a) e^{-\lambda(x,a)s} \right| \cdot (t-s) ds$$

$$+ \int_0^t \lambda(x,a) e^{-\lambda(x,a)s} \left| \sum_y f(y, t-s) [\hat{p}_k(y|x,a) - p(y|x,a)] \right| ds, \tag{29}$$

where in the last inequality we have used the properties that $0 \leq f(x,t) \leq t$ for all x, t, and that $\sum_y \hat{p}_k(y|x,a) = 1$.

We next bound the right hand side of (29). For the first term, note that for a fixed $s \in [0, t]$, the function $g(\lambda) := \lambda e^{-\lambda s}$ has the first-order derivative $g'(\lambda) = (1 - \lambda s) e^{-\lambda s}$, which is easily seen to satisfy $|g'(\lambda)| \leq 1$ for $\lambda \geq 0$. Hence g is Lipschitz continuous with a Lipschitz constant one. As a result,

$$\int_0^t \left| \hat{\lambda}_k(x,a) e^{-\hat{\lambda}_k(x,a)s} - \lambda(x,a) e^{-\lambda(x,a)s} \right| \cdot (t-s) ds \tag{30}$$

$$\leq \int_0^t (t-s) \cdot |\hat{\lambda}_k(x,a) - \lambda(x,a)| ds$$

$$= \frac{t^2}{2} \cdot |\hat{\lambda}_k(x,a) - \lambda(x,a)| \leq \frac{t^2}{2} \cdot \beta_k(x,a,\delta),$$

conditional on the occurrence of $\mathcal{G}$.

It remains to bound the second term on the right hand side of (29). Because $0 \leq f(x,t) \leq t$ for all x, t, it follows from Hölder's inequality that

$$\left| \sum_y f(y, t-s) [\hat{p}_k(y|x,a) - p(y|x,a)] \right| \leq ||f(\cdot, t-s)||_\infty \cdot ||\hat{p}_k(\cdot|x,a) - p(\cdot|x,a)||_1 \leq (t-s) \cdot \alpha_k(x,a,\delta), \tag{31}$$

conditional on $\mathcal{G}$. Combining (29), (30) and (31) yields (28). $\square$

With Lemma 4, we can show that the computed value function $\hat{V}_{n_k}^k$ from Step 8 of the CT-UCBVI algorithm is a high-probability upper bound of the optimal value function V^* up to some error. This is stated in the next result.

Lemma 5. *Conditional on the occurrence of the event $\mathcal{G}$, we have*

$$V^*(x,t) \leq \hat{V}_{n_k}^k(x,t) + \epsilon_k \cdot e^{\lambda_{\max} H}, \quad \text{for all } (x,t) \in \mathcal{S} \times [0, H], \; k = 1, \ldots, K, \tag{32}$$

where ϵ_k, $k = 1, \ldots, K$, are the accuracy parameters given in Algorithm 1.

Proof of Lemma 5. From the value iteration, we have

$$
\begin{aligned}
V_{n+1}^*(x,t) &= \sup_{a \in \mathcal{A}} (\mathcal{T}^a V_n^*)(x,t) \\
&= \sup_{a \in \mathcal{A}} \left\{ r(x,a) + \mathbb{E}_{\tau \sim \exp(\lambda(x,a)), Y \sim p(\cdot|x,a)} [V_n^*(Y, (t-\tau)^+)] \right\},
\end{aligned}
$$

and $\lim_{n \to \infty} V_n^* = V^*$. Meanwhile, in episode k, the CT-UCBVI algorithm applies the modified value iteration procedure to the CTMDP $(\mathcal{S}, \mathcal{A}, r(\cdot, \cdot) + b^k(\cdot, \cdot, \delta), \hat{\lambda}_k(\cdot, \cdot), \hat{p}_k(\cdot|\cdot, \cdot), H, x_0)$ with the accuracy parameter ϵ_k, where

$$
\hat{V}_{n+1}^k(x,t) = \min\{t, \sup_{a \in \mathcal{A}} (\mathcal{T}^{a,k} \hat{V}_n^k)(x,t)\}, \tag{33}
$$

with the operator $\mathcal{T}^{a,k}$ given in (15). Because $\mathcal{T}^{a,k}$ is a monotone nondecreasing mapping (see Remark 3.1 in Huang and Guo 2011), we obtain by induction that

$$
H \geq \hat{V}_{n+1}^k(x,t) \geq \hat{V}_n^k(x,t). \tag{34}
$$

Hence, the sequence $\hat{V}_n^k(x,t)$ converges as $n \to \infty$ and we denote its limit by $\hat{V}_\infty^k(x,t) := \lim_{n \to \infty} \hat{V}_n^k(x,t)$ for $(x,t) \in \mathcal{S} \times [0,H]$.

We first prove that $V^*(x,t) \leq \hat{V}_\infty^k(x,t)$ for all $(x,t) \in \mathcal{S} \times [0,H]$. It suffices to show $V_n^*(x,t) \leq \hat{V}_n^k(x,t)$ for all n, which we now prove by induction on n. At $n = 0$, we have $V_0^*(x,t) = \hat{V}_0^k(x,t) = 0$ for all $(x,t) \in \mathcal{S} \times [0,H]$. Assuming $V_n^*(x,t) \leq \hat{V}_n^k(x,t)$ for all x,t, we next prove that $V_{n+1}^*(x,t) \leq \hat{V}_{n+1}^k(x,t)$ for all x,t.

Fixing an action a, we compute

$$
\begin{aligned}
&(\mathcal{T}^{a,k} \hat{V}_n^k)(x,t) - (\mathcal{T}^a V_n^*)(x,t) \\
&= [b^k(x,a,\delta) + r(x,a)] \cdot \int_0^t e^{-\hat{\lambda}_k(x,a)s} ds - r(x,a) \cdot \int_0^t e^{-\lambda(x,a)s} ds \\
&\quad + \mathbb{E}_{\tau \sim \exp(\hat{\lambda}_k(x,a)), Y \sim \hat{p}_k(\cdot|x,a)} [\hat{V}_n^k(Y, (t-\tau)^+)] - \mathbb{E}_{\tau \sim \exp(\lambda(x,a)), Y \sim p(\cdot|x,a)} [V_n^*(Y, (t-\tau)^+)] \\
&\geq b^k(x,a,\delta) \cdot \int_0^t e^{-\lambda_{\max} s} ds + r(x,a) \left[\int_0^t e^{-\hat{\lambda}_k(x,a)s} ds - \int_0^t e^{-\lambda(x,a)s} ds \right] \\
&\quad + \mathbb{E}_{\tau \sim \exp(\hat{\lambda}_k(x,a)), Y \sim \hat{p}_k(\cdot|x,a)} [V_n^*(Y, (t-\tau)^+)] - \mathbb{E}_{\tau \sim \exp(\lambda(x,a)), Y \sim p(\cdot|x,a)} [V_n^*(Y, (t-\tau)^+)], \tag{35}
\end{aligned}
$$

where the last inequality is due to the induction hypothesis and $\hat{\lambda}_k(x,a) \leq \lambda_{\max}$. Because the reward function is bounded in $[0,1]$, we can use a similar argument as in (40) to deduce that, conditional on $\mathcal{G}$,

$$
\left| r(x,a) \left[\int_0^t e^{-\hat{\lambda}_k(x,a)s} ds - \int_0^t e^{-\lambda(x,a)s} ds \right] \right| \leq \frac{t^2}{2} \beta_k(x,a,\delta).
$$

Moreover, $0 \leq V_n^*(x,t) \leq V^*(x,t) \leq t$ for all (x,t) and n. Lemma 4 then implies

$$
\begin{aligned}
&\left| \mathbb{E}_{\tau \sim \exp(\hat{\lambda}_k(x,a)), Y \sim \hat{p}_k(\cdot|x,a)} [V_n^*(Y, (t-\tau)^+)] - \mathbb{E}_{\tau \sim \exp(\lambda(x,a)), Y \sim p(\cdot|x,a)} [V_n^*(Y, (t-\tau)^+)] \right| \\
&\leq \frac{t^2}{2} \beta_k(x,a,\delta) + t \alpha_k(x,a,\delta).
\end{aligned}
$$

21

It then follows from (35) that

$$(\mathcal{T}^{a,k}\hat{V}_n^k)(x,t) - (\mathcal{T}^a V_n^*)(x,t) \geq b^k(x,a,\delta) \cdot \int_0^t e^{-\lambda_{\max}s}ds - t^2\beta_k(x,a,\delta) - t\alpha_k(x,a,\delta).$$

Therefore, $(\mathcal{T}^{a,k}\hat{V}_n^k)(x,t) - (\mathcal{T}^a V_n^*)(x,t) \geq 0$ for all $(x,t) \in \mathcal{S} \times [0,H]$ if the following holds:

$$b^k(x,a,\delta) \geq \sup_{t\in[0,H]} \frac{t^2\beta_k(x,a,\delta) + t\alpha_k(x,a,\delta)}{\int_0^t e^{-\lambda_{\max}s}ds}. \tag{36}$$

One can easily verify that the function to be maximized on the right hand side of (36) is increasing in $t \geq 0$, and hence the maximum is attained at $t = H$, i.e.,

$$\sup_{t\in[0,H]} \frac{t^2\beta_k(x,a,\delta) + t\alpha_k(x,a,\delta)}{\int_0^t e^{-\lambda_{\max}s}ds} = \frac{H^2\beta_k(x,a,\delta) + H\alpha_k(x,a,\delta)}{\int_0^H e^{-\lambda_{\max}s}ds} = \frac{\lambda_{\max}}{1 - e^{-\lambda_{\max}H}} \left[H^2\beta_k(x,a,\delta) + H\alpha_k(x,a,\delta)\right].$$

Hence (36) indeed holds by the definition of $b^k(x,a,\delta)$ in (13). This proves $V_{n+1}^*(x,t) \leq \hat{V}_{n+1}^k(x,t)$ for all x,t, and thereby the desired inequality $V^*(x,t) \leq \hat{V}_\infty^k(x,t)$ for all x,t.

We are now ready to prove (32). We first show that in each episode k, when the modified value iteration stops in n_k iterations with $||\hat{V}_{n_k}^k - \hat{V}_{n_k-1}^k||_\infty \leq \epsilon_k$, it returns $\hat{V}_{n_k}^k$ with

$$||\hat{V}_\infty^k - \hat{V}_{n_k}^k||_\infty \leq \epsilon_k \cdot e^{\lambda_{\max}H}. \tag{37}$$

Indeed, the definition of $\mathcal{T}^{a,k}$ in (15) yields that for any functions $u,v \in \mathbb{M}$,

$$\left|(\mathcal{T}^{a,k}u)(x,t) - (\mathcal{T}^{a,k}v)(x,t)\right| = \left|\sum_y \int_0^t \hat{\lambda}_k(x,a)e^{-\hat{\lambda}_k(x,a)s} \cdot (u-v)(y,t-s) \cdot \hat{p}_k(y|x,a)ds\right|$$

$$\leq ||u-v||_\infty \cdot \int_0^t \hat{\lambda}_k(x,a)e^{-\hat{\lambda}_k(x,a)s}ds,$$

where we have used the fact that $\sum_y \hat{p}_k(y|x,a) = 1$. Consequently,

$$||\mathcal{T}^{a,k}u - \mathcal{T}^{a,k}v||_\infty \leq ||u-v||_\infty \cdot \sup_{x\in\mathcal{S}} \int_0^H \hat{\lambda}_k(x,a)e^{-\hat{\lambda}_k(x,a)s}ds$$

$$\leq ||u-v||_\infty \cdot (1 - e^{-\lambda_{\max}H}),$$

noting that $\hat{\lambda}_k(x,a) \leq \lambda_{\max}$ for all x,a. This implies that the operator $\mathcal{T}^{a,k}$ is a contraction mapping with a contraction coefficient $\rho := 1 - e^{-\lambda_{\max}H} \in (0,1)$. It follows that the mapping from u to $\sup_{a\in\mathcal{A}} \mathcal{T}^{a,k}u$ is also a contraction with the same coefficient ρ; see, e.g., Theorem 2 in Denardo (1967). Because the function $\min\{H,x\}$ is Lipschitz in $x \geq 0$ with the Lipschitz constant one, we infer from (33) that the mapping from $\hat{V}_n^k$ to $\hat{V}_{n+1}^k$ in the modified value iteration is a contraction with the coefficient ρ. A standard argument, see, e.g., the proof of Theorem 6.3.1 in Puterman (2014), yields that (37) holds. As we have proved $V^*(x,t) \leq \hat{V}_\infty^k(x,t)$, it follows that $V^*(x,t) \leq \hat{V}_{n_k}^k(x,t) + \epsilon_k \cdot e^{\lambda_{\max}H}$ for all x,t. The proof is complete. $\qquad\square$

In contrast to the discrete-time RL setting (see, e.g., Lemma 18 in Azar et al. 2017), we have an additional error term $\epsilon_k \cdot e^{\lambda_{\max}H}$ appearing in (32). This is because the modified value iteration

22

in Algorithm 2 for solving finite-horizon CTMDPs converges asymptotically with a linear rate $\rho := 1 - e^{-\lambda_{\max}H}$ when $n \to \infty$. When this iteration is stopped at a finite number of iterations n_k, it naturally leads to some approximation error. In the above proof, by showing that the operator (15) is a contraction mapping with the coefficient ρ, we control this error using similar arguments as in the analysis of value iteration for discrete-time infinite-horizon discounted MDPs (Puterman 2014, Chapter 6.3).

In the next and final lemma of this subsection, we bound the per episode regret. For each episode k, denote by $\mathcal{H}_{k-1}$ the history up to the end of episode $k-1$, i.e. $\mathcal{H}_{k-1} = \{(x_n^l, a_n^l, \tau_n^l)_n : l = 1, \ldots, k-1\}$. Note that the policy π^k specified in (14) depends on $\mathcal{H}_{k-1}$. Also recall the definitions of α_k, β_k and b^k in (25), (21) and (13) respectively.

Lemma 6. *Conditional on the occurrence of $\mathcal{G}$, for each episode k, we have*

$$V^*(x_0^k, H) - V^{\pi_k}(x_0^k, H)$$
$$\leq \sum_{m=0}^{\infty} \mathbb{E}\left([b^k(x_m^k, a_m^k, \delta) + \alpha_k(x_m^k, a_m^k, \delta) + (H - t_m^k)^+ \cdot \beta_k(x_m^k, a_m^k, \delta)] \cdot (H - t_m^k)^+ \big| \mathcal{H}_{k-1} \right)$$
$$+ \epsilon_k \cdot e^{\lambda_{\max}H},$$

where $(x_i^k, a_i^k, \tau_i^k)_{i \geq 0}$ consists of the trajectories of the state, action and holding time corresponding to the policy π^k, $t_m^k = \sum_{i=0}^{m-1} \tau_i^k$, and the expectation is taken over the randomness in the holding times $(\tau_i^k)_{i \geq 0}$ and state transitions $(x_i^k)_{i \geq 1}$.

Proof of Lemma 6. Lemma 5 yields $V^*(x, H) \leq \hat{V}_{n_k}^k(x, H) + \epsilon_k \cdot e^{\lambda_{\max}H}$. So we only need to bound the quantity $\hat{V}_{n_k}^k(x_0^k, H) - V^{\pi_k}(x_0^k, H)$.

Recall that for $i \leq n_k - 1$, we have the modified value iteration:

$$\hat{V}_{i+1}^k(x, t) = \min\{t, \sup_{a \in \mathcal{A}}(\mathcal{T}^{a,k}\hat{V}_i^k)(x, t)\} \quad \text{and} \quad \pi^k(x, t) = \arg\max_{a \in \mathcal{A}}\{(\mathcal{T}^{a,k}\hat{V}_{n_k}^k)(x, t)\}, \quad \text{for all } x, t.$$

By (34) and the monotonicity of the operator $\mathcal{T}^{a,k}$, we have

$$\hat{V}_{n_k}^k(x_0^k, H) \leq \sup_{a \in \mathcal{A}}(\mathcal{T}^{a,k}\hat{V}_{n_k-1}^k)(x_0^k, H) \leq \sup_{a \in \mathcal{A}}(\mathcal{T}^{a,k}\hat{V}_{n_k}^k)(x_0^k, H)$$
$$= \left[r(x_0^k, a_0^k) + b^k(x_0^k, a_0^k, \delta) \right] \cdot \int_0^H e^{-\hat{\lambda}_k(x_0^k, a_0^k)s} ds$$
$$+ \mathbb{E}_{\tau_0^k \sim \exp(\hat{\lambda}_k(x_0^k, a_0^k)), x_1^k \sim \hat{p}_k(\cdot|x_0^k, a_0^k)}[\hat{V}_{n_k}^k(x_1^k, (H - \tau_0^k)^+)], \tag{38}$$

where $a_0^k := \pi^k(x_0^k, H)$ with H being the remaining horizon at time 0. In addition, given $\mathcal{H}_{k-1}$, the policy π_k in (14) is Markovian and we have the following relation (see Lemma 3.1 in Huang and Guo 2011):

$$V^{\pi_k}(x_0^k, H) = r(x_0^k, a_0^k) \cdot \int_0^H e^{-\lambda(x_0^k, a_0^k)s} ds + \mathbb{E}_{\tau_0^k \sim \exp(\lambda(x_0^k, a_0^k)), x_1^k \sim p(\cdot|x_0^k, a_0^k)}[V^{\pi^k}(x_1^k, (H - \tau_0^k)^+)]. \tag{39}$$

In view of (38) and (39), we can directly compute that

$$\left[r(x_0^k, a_0^k) + b^k(x_0^k, a_0^k, \delta) \right] \cdot \int_0^H e^{-\hat{\lambda}_k(x_0^k, a_0^k)s} ds - r(x_0^k, a_0^k) \cdot \int_0^H e^{-\lambda(x_0^k, a_0^k)s} ds$$

$$= r(x_0^k, a_0^k) \cdot \left| \int_0^H e^{-\hat{\lambda}_k(x_0^k, a_0^k)s} ds - \int_0^H e^{-\lambda(x_0^k, a_0^k)s} ds \right| + b^k(x_0^k, a_0^k, \delta) \cdot \int_0^H e^{-\hat{\lambda}_k(x_0^k, a_0^k)s} ds$$

$$\leq r(x_0^k, a_0^k) \cdot \frac{H^2}{2} \left| \hat{\lambda}_k(x_0^k, a_0^k) - \lambda(x, a) \right| + b^k(x_0^k, a_0^k, \delta) \cdot H$$

$$\leq \frac{H^2}{2} \beta_k(x_0^k, a_0^k, \delta) + b^k(x_0^k, a_0^k, \delta) \cdot H, \tag{40}$$

where in the second inequality we use the fact that the function $\int_0^H e^{-\lambda s} ds$ is Lipschitz continuous in $\lambda \in [0, \lambda_{\max}]$ with a Lipschitz constant $H^2/2$, and in the third inequality we use the assumption that the event $\mathcal{G}$ occurs and the fact that rewards are bounded by one. Then we deduce from (38) and (39) that, given $\mathcal{H}_{k-1}$,

$$\hat{V}_{n_k}^k(x_0^k, H) - V^{\pi_k}(x_0^k, H)$$

$$\leq \frac{H^2}{2} \beta_k(x_0^k, a_0^k, \delta) + b^k(x_0^k, a_0^k, \delta) \cdot H$$

$$\quad + \mathbb{E}_{\tau_0^k \sim \exp(\hat{\lambda}_k(x_0^k, a_0^k)), x_1^k \sim \hat{p}_k(\cdot|x_0^k, a_0^k)} [\hat{V}_{n_k}^k(x_1^k, (H - \tau_0^k)^+)]$$

$$\quad - \mathbb{E}_{\tau_0^k \sim \exp(\lambda(x_0^k, a_0^k)), x_1^k \sim p(\cdot|x_0^k, a_0^k)} [V^{\pi^k}(x_1^k, (H - \tau_0^k)^+)]$$

$$= \frac{H^2}{2} \beta_k(x_0^k, a_0^k, \delta) + b^k(x_0^k, a_0^k, \delta) \cdot H$$

$$\quad + \mathbb{E}_{\tau_0^k \sim \exp(\hat{\lambda}_k(x_0^k, a_0^k)), x_1^k \sim \hat{p}_k(\cdot|x_0^k, a_0^k)} [\hat{V}_{n_k}^k(x_1^k, (H - \tau_0^k)^+)]$$

$$\quad - \mathbb{E}_{\tau_0^k \sim \exp(\lambda(x_0^k, a_0^k)), x_1^k \sim p(\cdot|x_0^k, a_0^k)} [\hat{V}_{n_k}^k(x_1^k, (H - \tau_0^k)^+)]$$

$$\quad + \mathbb{E}_{\tau_0^k \sim \exp(\lambda(x_0^k, a_0^k)), x_1^k \sim p(\cdot|x_0^k, a_0^k)} [\hat{V}_{n_k}^k(x_1^k, (H - \tau_0^k)^+) - V^{\pi_k}(x_1^k, (H - \tau_0^k)^+)]$$

$$\leq H^2 \beta_k(x_0^k, a_0^k, \delta) + H \cdot b^k(x_0^k, a_0^k, \delta) + H \cdot \alpha_k(x_0^k, a_0^k, \delta)$$

$$\quad + \mathbb{E}_{\tau_0^k \sim \exp(\lambda(x_0^k, a_0^k)), x_1^k \sim p(\cdot|x_0^k, a_0^k)} [\hat{V}_{n_k}^k(x_1^k, (H - t_1^k)^+) - V^{\pi_k}(x_1^k, (H - t_1^k)^+)],$$

where the last inequality follows from Lemma 4 and the fact that $t_1^k = \tau_0^k$. Using a similar argument, we can bound $\hat{V}_{n_k}^k(x_1^k, (H - t_1^k)^+) - V^{\pi_k}(x_1^k, (H - t_1^k)^+)$:

$$\hat{V}_{n_k}^k(x_1^k, (H - t_1^k)^+) - V^{\pi_k}(x_1^k, (H - t_1^k)^+)$$

$$\leq ((H - t_1^k)^+)^2 \beta_k(x_1^k, a_1^k, \delta) + (H - t_1^k)^+ \cdot b^k(x_1^k, a_1^k, \delta) + (H - t_1^k)^+ \cdot \alpha_k(x_1^k, a_1^k, \delta)$$

$$\quad + \mathbb{E}_{\tau_1^k \sim \exp(\lambda(x_1^k, a_1^k)), x_2^k \sim p(\cdot|x_1^k, a_1^k)} [\hat{V}_{n_k}^k(x_2^k, (H - t_2^k)^+) - V^{\pi_k}(x_2^k, (H - t_2^k)^+)],$$

where $t_2^k = t_1^k + \tau_1^k = \tau_0^k + \tau_1^k$.

Indeed, we can apply the above procedure recursively to obtain, noting that $\hat{V}_{n_k}^k(x, 0) =$

24

$V^{\pi_k}(x, 0) = 0$ for all $x \in \mathcal{S}$,

$$\hat{V}^k_{n_k}(x^k_0, H) - V^{\pi_k}(x^k_0, H) \leq \sum_{m=0}^{\infty} \mathbb{E}\left[b^k(x^k_m, a^k_m, \delta) \cdot (H - t^k_m)^+ \big| \mathcal{H}_{k-1}\right]$$
$$+ \sum_{m=0}^{\infty} \mathbb{E}\left[\beta_k(x^k_m, a^k_m, \delta) \cdot [(H - t^k_m)^+]^2 \big| \mathcal{H}_{k-1}\right]$$
$$+ \sum_{m=0}^{\infty} \mathbb{E}\left[\alpha_k(x^k_m, a^k_m, \delta) \cdot (H - t^k_m)^+ \big| \mathcal{H}_{k-1}\right],$$

where the expectation is taken over the randomness in the holding times $(\tau^k_i)_{i \geq 0}$ and state transitions $(x^k_i)_{i \geq 1}$ under π_k, conditional on $\mathcal{H}_{k-1}$. The proof is complete. $\square$

6.2 Proof of Theorem 1

In this section, we prove Theorem 1, with the objective being to bound the expected total regret over K episodes. The main difficulty in the analysis is to bound the regret from estimation errors of transition probabilities and holding time rates for the CTMDP. In particular, the standard pigeon-hole argument can not be directly applied to our setting. To illustrate this point as well as both the subtlety and complexity in bounding the regret from estimation errors of transition probabilities of CTMDPs, we consider one sample path where $t^k_1 > H$ for all $k = 1, \ldots, K$ (recall that t^k_1 is the first transition time in episode k, and it follows an exponential distribution). In such a case, we can infer from (10) that $N^{k,+}(x, a) = 0$ for all (x, a) and k, which implies that for this path,

$$\sum_{k=1}^{K} \frac{1}{\sqrt{\max\{1, N^{k,+}(x^k_0, a^k_0)\}}} = K. \tag{41}$$

Hence, the standard pathwise pigeon-hole argument used for obtaining *sublinear* regret in episodic DTMDPs (see, e.g., Azar et al. (2017)) no longer works in our setting. We overcome this difficulty by devising a new probabilistic argument to bound the regret from estimation errors of transition probabilities (which needs to be weighted by the remaining time horizon $(H - t^k_m)^+$). The result is presented as Proposition 1.

Proposition 1.

$$\mathbb{E}\left[\sum_{k=1}^{K} \sum_{m=0}^{\infty} \frac{(H - t^k_m)^+}{\sqrt{\max\{1, N^{k,+}(x^k_m, a^k_m)\}}}\right] \leq 3\left(H + \frac{1}{\lambda_{\min}}\right) \cdot \sqrt{SAK} \cdot \frac{(\lambda_{\max} H + 1)\ln K}{\ln(\ln(K) + 1)}.$$

To bound the regret from estimation error in the holding time rate, which involves the continuous varible $T^k(x, a)$ (the cumulative amount of time spent at (x, a)), we use an idea similar to that in the proof of Lemma 1. That is, we convert bounding the function of $T^k(x, a)$ to bounding the function of $N^{k,+}(x, a)$ by exploiting properties of Poisson distributions. The result is given in Proposition 2.

Proposition 2.

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{m=0}^{\infty}\frac{(H-t_m^k)^+}{\sqrt{\max\{T^k(x_m^k,a_m^k),L(\delta)/\lambda_{\max}\}}}\right]$$

$$\leq \sqrt{2\lambda_{\max}}\cdot 3\left(H+\frac{1}{\lambda_{\min}}\right)\cdot\sqrt{SAK}\cdot\frac{(\lambda_{\max}H+1)\ln K}{\ln(\ln(K)+1)}.$$

Proofs of Propositions 1 and 2 are deferred to the end of this section.

Proof of Theorem 1.   Write the total regret as

$$\text{Regret}(\mathcal{M},\text{CT-UCBVI},K)$$

$$=\mathbb{E}\left[\sum_{k=1}^{K}V^*(x_0^k,H)-\sum_{k=1}^{K}V^{\pi_k}(x_0^k,H)\right]$$

$$=\mathbb{E}\left[\mathbf{1}_{\mathcal{G}}\cdot\sum_{k=1}^{K}(V^*(x_0^k,H)-V^{\pi_k}(x_0^k,H))\right]+\mathbb{E}\left[\mathbf{1}_{\mathcal{G}^c}\cdot\sum_{k=1}^{K}(V^*(x_0^k,H)-V^{\pi_k}(x_0^k,H))\right]$$

$$\leq\mathbb{E}\left[\mathbf{1}_{\mathcal{G}}\cdot\sum_{k=1}^{K}(V^*(x_0^k,H)-V^{\pi_k}(x_0^k,H))\right]+2\delta KH, \tag{42}$$

where the inequality holds because $\mathbb{P}(\mathcal{G}^c)\leq 2\delta$ and $0\leq V^{\pi^k}(x_0^k,H)\leq V^*(x_0^k,H)\leq H$ (by the assumption that the reward rates are bounded between zero and one). In the following we choose $\delta=1/(2KH)$ so that $2\delta KH=1$ in (42) and $L(\delta)=4\ln(2SAK/\delta)>1$.

For notational simplicity, let $C:=\frac{\lambda_{\max}}{1-e^{-\lambda_{\max}H}}\vee 1$. It follows from Lemma 6 and the definitions of α_k,β_k and b^k in (25), (21) and (13) that, conditional on $\mathcal{G}$,

$$V^*(x_0^k,H)-V^{\pi_k}(x_0^k,H)$$

$$\leq\epsilon_k\cdot e^{\lambda_{\max}H}+(CH+1)\cdot\sum_{m=0}^{\infty}\mathbb{E}\left[\sqrt{\frac{2[S\ln 2+\ln(SAHK^2/\delta)]}{\max\{1,N^{k,+}(x_m^k,a_m^k)\}}}\cdot(H-t_m^k)^+\Big|\mathcal{H}_{k-1}\right]$$

$$+(CH^2+H)\cdot\sum_{m=0}^{\infty}\mathbb{E}\left[\sqrt{\frac{\lambda_{\max}L(\delta)}{\max\{T^k(x_m^k,a_m^k),L(\delta)/\lambda_{\max}\}}}\cdot(H-t_m^k)^+\Big|\mathcal{H}_{k-1}\right]$$

where $T^k(x,a)$ and $N^{k,+}(x,a)$ are given in (8) and (10) respectively. This implies

$$\mathbb{E}\left[\mathbf{1}_{\mathcal{G}}\cdot\sum_{k=1}^{K}(V^*(x_0^k,H)-V^{\pi_k}(x_0^k,H))\right]$$

$$\leq\sum_{k=1}^{K}\epsilon_k\cdot e^{\lambda_{\max}H}+(CH+1)\cdot\sum_{k=1}^{K}\sum_{m=0}^{\infty}\mathbb{E}\left[\sqrt{\frac{2[S\ln 2+\ln(SAHK^2/\delta)]}{\max\{1,N^{k,+}(x_m^k,a_m^k)\}}}\cdot(H-t_m^k)^+\right]$$

$$+(CH^2+H)\cdot\sum_{k=1}^{K}\sum_{m=0}^{\infty}\mathbb{E}\left[\sqrt{\frac{\lambda_{\max}L(\delta)}{\max\{T^k(x_m^k,a_m^k),L(\delta)/\lambda_{\max}\}}}\cdot(H-t_m^k)^+\right]. \tag{43}$$

26

Applying Propositions 1 and 2, we can readily infer from (43) that

$$\text{Regret}(\mathcal{M}, \text{CT-UCBVI}, K)$$
$$\leq 3\,(CH+1) \cdot \sqrt{SAK} \cdot \left(H + \frac{1}{\lambda_{\min}}\right) \cdot (\lambda_{\max} H + 1)\frac{\ln K}{\ln(\ln(K)+1)} \cdot$$
$$\cdot \left(\sqrt{2S + 6\ln(2SAKH)} + 4\lambda_{\max} H \sqrt{\ln(2SAKH)}\right)$$
$$+ \sum_{k=1}^{K} \epsilon_k \cdot e^{\lambda_{\max} H} + 1.$$

The proof is complete. $\qquad\square$

6.2.1   Proofs of Propositions 1 and 2

Proof of Proposition 1.   Noting $(H - t_m^k)^+ = (H - t_m^k)1_{t_m^k < H}$ and $\lim_{x \to 0+} \frac{x}{1 - e^{-\lambda_{\min} x}} = 1/\lambda_{\min} \in (0, \infty)$, we have

$$\frac{(H - t_m^k)^+}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}} = \frac{(1 - e^{-\lambda_{\min}(H - t_m^k)^+}) \cdot 1_{t_m^k < H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}} \cdot \frac{(H - t_m^k)^+}{1 - e^{-\lambda_{\min}(H - t_m^k)^+}}. \tag{44}$$

Using the fact that $e^{\lambda_{\min} x} \geq 1 + \lambda_{\min} x$ for any $x \geq 0$, one can easily verify that $x/(1 - e^{-\lambda_{\min} x}) \leq x + \frac{1}{\lambda_{\min}}$ for any $x \geq 0$. It follows that

$$\frac{(H - t_m^k)^+}{1 - e^{-\lambda_{\min}(H - t_m^k)^+}} \leq (H - t_m^k)^+ + \frac{1}{\lambda_{\min}} \leq H + \frac{1}{\lambda_{\min}}.$$

This implies

$$\frac{(H - t_m^k)^+}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}} \leq \frac{(1 - e^{-\lambda_{\min}(H - t_m^k)^+}) \cdot 1_{t_m^k < H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}} \cdot \left(H + \frac{1}{\lambda_{\min}}\right). \tag{45}$$

Because the holding time $t_{m+1}^k - t_m^k$ has an exponential distribution with rate parameter $\lambda(x_m^k, a_m^k)$, we can compute

$$\mathbb{E}\left[\frac{1_{t_{m+1}^k \leq H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}} \,\Big|\, (x_m^k, a_m^k, t_m^k), N^{k,+}(x_m^k, a_m^k)\right]$$
$$= \mathbb{P}(t_{m+1}^k - t_m^k \leq (H - t_m^k)^+ | x_m^k, a_m^k, t_m^k) \cdot \frac{1_{t_m^k < H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}}$$
$$= (1 - e^{-\lambda(x_m^k, a_m^k)(H - t_m^k)^+}) \cdot \frac{1_{t_m^k < H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}}$$
$$\geq \frac{(1 - e^{-\lambda_{\min}(H - t_m^k)^+}) \cdot 1_{t_m^k < H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}},$$

where the last inequality is due to the fact that $\lambda(x,a) \geq \lambda_{\min}$ for all (x,a). It then follows from (45) that

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{m=0}^{\infty}\frac{(H-t_m^k)^+}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}}\right] \leq \left(H + \frac{1}{\lambda_{\min}}\right) \cdot \mathbb{E}\left[\sum_{m=0}^{\infty}\sum_{k=1}^{K}\frac{1_{t_{m+1}^k \leq H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}}\right].$$
(46)

We next bound the right-hand side of (46). For a state–action pair $(x,a) \in \mathcal{S} \times \mathcal{A}$, denote by $Y_m^l(x,a) := 1_{\{(x_m^l, a_m^l)=(x,a)\}} \cdot 1_{t_{m+1}^l \leq H}$, and let

$$N_m^{k,+}(x,a) := \sum_{l=1}^{k-1} Y_m^l(x,a) = \sum_{l=1}^{k-1} 1_{\{(x_m^l, a_m^l)=(x,a)\}} \cdot 1_{t_{m+1}^l \leq H}.$$

It is clear that $0 \leq Y_m^k(x,a) \leq 1 \leq \max\{1, N_m^{k,+}(x,a)\}$ and $N_m^{k,+}(x,a) \leq N^{k,+}(x,a)$. We can then obtain

$$\sum_{k=1}^{K}\frac{1_{t_{m+1}^k \leq H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}} \leq \sum_{k=1}^{K}\frac{1_{t_{m+1}^k \leq H}}{\sqrt{\max\{1, N_m^{k,+}(x_m^k, a_m^k)\}}}$$

$$\leq \sum_{(x,a)}\sum_{k=1}^{K}\frac{Y_m^k(x,a)}{\sqrt{\max\{1, N_m^{k,+}(x,a)\}}}$$

$$\leq 3\sum_{(x,a)}\sqrt{N_m^{(K+1),+}(x,a)}$$

$$\leq 3\sqrt{SA \cdot \sum_{(x,a)} N_m^{(K+1),+}(x,a)}$$

$$= 3\sqrt{SA \cdot \sum_{l=1}^{K} 1_{t_{m+1}^l \leq H}},$$

where the third inequality is due to Lemma 19 in Jaksch et al. (2010), and the fourth inequality follows from Cauchy–Schwarz inequality. Thus, we have

$$\mathbb{E}\left[\sum_{m=0}^{\infty}\sum_{k=1}^{K}\frac{1_{t_{m+1}^k \leq H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}}\right] \leq 3\sqrt{SA} \cdot \mathbb{E}\left[\sum_{m=0}^{\infty}\sqrt{\sum_{l=1}^{K} 1_{t_{m+1}^l \leq H}}\right].$$
(47)

Given the policy π^k, denote by $J^k(H) := \sup\{m : t_m^k \leq H\}$, and set $J_M^K(H) := \max\{J^k(H) : k = 1, 2, \ldots, K\}$. It is clear that

$$\sum_{m=0}^{\infty}\sqrt{\sum_{l=1}^{K} 1_{t_{m+1}^l \leq H}} = \sum_{m=0}^{J_M^K(H)-1}\sqrt{\sum_{l=1}^{K} 1_{t_{m+1}^l \leq H}} \leq \sqrt{K} \cdot J_M^K(H).$$
(48)

In the equation above, we have allowed the agent to make (the same) $(J_M^K(H) + 1)$ number of decisions for each episode, and hence it is possible that some of the decision steps are taken after

the time H. Such decision steps are *fictitious*, introduced only for the purpose of a cleaner analysis. We then obtain from (47) that

$$\mathbb{E}\left[\sum_{m=0}^{\infty}\sum_{k=1}^{K}\frac{1_{t_{m+1}^k \leq H}}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}}\right] \leq 3\sqrt{SAK} \cdot \mathbb{E}(J_M^K(H)). \tag{49}$$

It remains to bound $\mathbb{E}(J_M^K(H))$. Noting $\lambda(x, a) \leq \lambda_{\max}$ for all (x, a), we get that $J^k(H) \leq Z^k(H)$ almost surely for each $k = 1, \ldots, K$, where $\{Z^k(\cdot) : k = 1, \ldots, K\}$ are K independent Poisson processes with a common rate $\lambda_{\max}$. It then follows from the definition of $J_M^K(H)$ that

$$\mathbb{E}[J_M^K(H)] \leq \mathbb{E}\left[\max_{k=1,\ldots,K}\{Z^k(H)\}\right] \leq \frac{(\lambda_{\max}H + 1)\ln K}{\ln(\ln(K) + 1)}, \tag{50}$$

where the second inequality is a standard bound on the expected maximum value of several independent Poisson random variables with a common mean $\lambda_{\max}H$; see e.g. Exercise 2.18 in Boucheron et al. (2013). Therefore, we can infer from (49) and (46) that

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{m=0}^{\infty}\frac{(H - t_m^k)^+}{\sqrt{\max\{1, N^{k,+}(x_m^k, a_m^k)\}}}\right] \leq 3\left(H + \frac{1}{\lambda_{\min}}\right) \cdot \sqrt{SAK} \cdot \frac{(\lambda_{\max}H + 1)\ln K}{\ln(\ln(K) + 1)}.$$

The proof is complete. $\qquad\square$

Proof of Proposition 2. Fix $(x, a) \in \mathcal{S} \times \mathcal{A}$ and k. As already shown in the proof of Lemma 1, given $T^k(x, a)$, $N^{k,+}(x, a)$ follows a Poisson distribution with mean $\lambda(x, a)T^k(x, a)$. Hence we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\sqrt{\max\{N^{k,+}(x, a), 1\}}}\middle| T^k(x, a)\right] &\geq \frac{1}{\sqrt{\mathbb{E}[\max\{N^{k,+}(x, a), 1\}|T^k(x, a)]}} \\
&= \frac{1}{\sqrt{e^{-\lambda(x,a)T^k(x,a)} + \lambda(x, a)T^k(x, a)}} \\
&\geq \frac{1}{\sqrt{2\max\{\lambda(x, a)T^k(x, a), 1\}}} \\
&\geq \frac{1}{\sqrt{2\lambda_{\max}}} \cdot \frac{1}{\sqrt{\max\{T^k(x, a), 1/\lambda_{\max}\}}},
\end{aligned}$$

where the first inequality is due to Jensen's inequality, the second inequality follows from the fact that $e^{-t} + t \leq 2\max\{t, 1\}$ for $t \geq 0$, and the last inequality is due to $\lambda(x, a) \leq \lambda_{\max}$. It follows that

$$\begin{aligned}
\mathbb{E}&\left[\frac{(H - t_m^k)^+}{\sqrt{\max\{N^{k,+}(x_m^k, a_m^k), 1\}}}\right] \\
&= \mathbb{E}\left(\mathbb{E}\left[\frac{(H - t_m^k)^+}{\sqrt{\max\{N^{k,+}(x_m^k, a_m^k), 1\}}}\middle| (x_m^k, a_m^k, t_m^k), T^k(x_m^k, a_m^k)\right]\right) \\
&\geq \frac{1}{\sqrt{2\lambda_{\max}}} \cdot \mathbb{E}\left(\frac{(H - t_m^k)^+}{\sqrt{\max\{T^k(x_m^k, a_m^k), 1/\lambda_{\max}\}}}\right). \tag{51}
\end{aligned}$$

Consequently,

$$\sum_{k=1}^{K} \sum_{m=0}^{\infty} \mathbb{E} \left[\frac{(H - t_m^k)^+}{\sqrt{\max\{T^k(x_m^k, a_m^k), L(\delta)/\lambda_{\max}\}}} \right]$$

$$\leq \sum_{k=1}^{K} \sum_{m=0}^{\infty} \mathbb{E} \left[\frac{(H - t_m^k)^+}{\sqrt{\max\{T^k(x_m^k, a_m^k), 1/\lambda_{\max}\}}} \right]$$

$$\leq \sum_{k=1}^{K} \sum_{m=0}^{\infty} \sqrt{2\lambda_{\max}} \cdot \mathbb{E} \left[\frac{(H - t_m^k)^+}{\sqrt{\max\{N^{k,+}(x_m^k, a_m^k), 1\}}} \right]$$

$$\leq \sqrt{2\lambda_{\max}} \cdot 3 \left(H + \frac{1}{\lambda_{\min}} \right) \cdot \sqrt{SAK} \cdot \frac{(\lambda_{\max} H + 1) \ln K}{\ln(\ln(K) + 1)},$$

where the first inequality follows from $L(\delta) > 1$, the second inequality is due to (51), and the last inequality is due to Proposition 1. The proof of Proposition 2 is complete. $\square$

6.3 Proof of Theorem 2

We adapt the proof of Theorem 9 in Domingues et al. (2021) for episodic RL in discrete-time MDPs. The main thrust of their proof is to design a class of hard MDP instances that are based on the construction of MDPs in (Lattimore and Szepesvári 2020, Chapter 38.7) and Jaksch et al. (2010). We make necessary modifications of such MDP instances to fit into the continuous-time setting and prove our regret lower bound.

Proof of Theorem 2. Because the proof is similar to Domingues et al. (2021), we only provide the key steps.

- *Step 1. Constructing CTMDP instances.*

  Given S and A, by Assumption 2, we can use $S-2$ states to construct a full A-ary tree of depth $d-1$. In this rooted tree, each node has exactly A children and the total number of nodes is given by $\sum_{i=0}^{d-1} A^i = S - 2$. The remaining two states are denoted by s_g ('good' state) and s_b ('bad' state), both of which are absorbing states. The transitions are deterministic within the tree: the initial state of the CTMDPs is the root of the tree, and given a node/state, action a leads to the a-th child. Let L be the number of leaves, and the collection of the leaves are denoted by $\mathcal{L} = \{s_1, \ldots, s_L\}$. For each $s_i \in \mathcal{L}$, any action will lead to a transition to either the good state or the bad state according to

  $$p(s_g|s_i, a) = \frac{1}{2} + \epsilon(s_i, a), \quad p(s_b|s_i, a) = \frac{1}{2} - \epsilon(s_i, a),$$

  where the function ϵ will be specified later. The reward function $r(s, a) = 1$ if $s = s_g$ and it is zero otherwise. The holding time for any state–action pair is exponentially distributed with rate $\lambda_{\max}$, except when the state is $s = s_g$ or $s = s_b$. Let $\mathcal{M}_0$ be the CTMDP with $\epsilon(s, a) = 0$ for all $(s, a) \in \mathcal{L} \times \mathcal{A}$. For the elements (state-action pairs) in the set $\mathcal{J} := \mathcal{L} \times \mathcal{A}$ we order them from 1 to $|\mathcal{J}|$. Let $\mathcal{M}_j$, $j \geq 1$, be the CTMDP with $\epsilon(s, a) = \Delta$ for the j-th state–action pair in $\mathcal{J}$ and $\epsilon(s, a) = 0$ otherwise, where Δ is some parameter to be tuned later.

- *Step 2. Computing the expected regret of an **algo** in $\mathcal{M}_j$.*

  Given a learning algorithm **algo**, for notational simplicity, denote by R_j its expected regret for the CTMDP $\mathcal{M}_j$. The optimal cumulative expected reward over $[0, H]$ in the CTMDP $\mathcal{M}_j$ for $j \geq 1$ is given by

  $$\mathbb{E}[(H - \gamma_d)^+] \cdot (1/2 + \Delta),$$

  where γ_d is the sum of d i.i.d exponential random variables and it follows the Erlang distribution with shape parameter d and rate parameter $\lambda_{\max}$. This is because the optimal policy for $\mathcal{M}_j$ is simply to traverse the tree to reach the leaf state (corresponding to the state in the j-th state–action pair in $\mathcal{J}$ for which ϵ equals to Δ) in $d-1$ transitions, and then take the corresponding action to reach the good state s_g in one more transition (with probability $1/2 + \Delta$) to earn the reward. Denote by $\mathbb{P}_j$ and $\mathbb{E}_j$, $j \geq 1$, the probability measure and expectation in the CTMDP $\mathcal{M}_j$ by following the **algo**, and by $\mathbb{P}_0$ and $\mathbb{E}_0$ the corresponding operators in $\mathcal{M}_0$. Then the cumulative expected reward of the **algo** in $\mathcal{M}_j$ in episode k is given by

  $$\mathbb{E}[(H - \gamma_d)^+] \cdot \mathbb{E}_j[1_{x_d^k = s_g}],$$

  where x_d^k is the state after d transitions in episode k. It follows that the expected regret is given by

  $$
  \begin{aligned}
  R_j &= K\mathbb{E}[(H - \gamma_d)^+] \cdot (1/2 + \Delta) - \mathbb{E}[(H - \gamma_d)^+] \cdot \sum_{k=1}^{K} \mathbb{E}_j[1_{x_d^k = s_g}] \\
  &= K\mathbb{E}[(H - \gamma_d)^+] \cdot (1/2 + \Delta) - \mathbb{E}[(H - \gamma_d)^+] \cdot (K/2 + \Delta\mathbb{E}_j\mathbf{T}_j) \\
  &= K\mathbb{E}[(H - \gamma_d)^+]\Delta \cdot (1 - \frac{1}{K}\mathbb{E}_j\mathbf{T}_j),
  \end{aligned}
  $$

  where $\mathbf{T}_j \leq K$ denotes the total number of visits to the state–action pair $j \in \mathcal{J}$ during all the K episodes while the **algo** is applied to $\mathcal{M}_j$. Hence, noting that $|\mathcal{J}| = LA$, we conclude that the maximum regret of the **algo** over all possible $\mathcal{M}_j$ for $j \in \mathcal{J}$ is lower bounded by

  $$\max_{j \in \mathcal{J}} R_j \geq \frac{1}{LA} \sum_{j \in \mathcal{J}} R_j = K\mathbb{E}[(H - \gamma_d)^+]\Delta \cdot \left(1 - \frac{1}{KLA} \sum_{j \in \mathcal{J}} \mathbb{E}_j\mathbf{T}_j\right). \tag{52}$$

- *Step 3. Upper bounding $\sum_{j \in \mathcal{J}} \mathbb{E}_j\mathbf{T}_j$.*

  Fix $j \in \mathcal{J}$. Since $\mathbf{T}_j/K \in [0, 1]$, one can show that $kl(\frac{1}{K}\mathbb{E}_0[\mathbf{T}_j], \frac{1}{K}\mathbb{E}_j[\mathbf{T}_j]) \leq \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_j)$, where KL denotes the Kullback–Leibler divergence between two probability measures and $kl(p, q)$ denotes the KL divergence between two Bernoulli distributions with success probabilities p and q respectively (Garivier et al. 2019). It then follows from Pinsker's inequality that

  $$\frac{1}{K}\mathbb{E}_j[\mathbf{T}_j] \leq \frac{1}{K}\mathbb{E}_0[\mathbf{T}_j] + \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_j)}.$$

Because $\mathcal{M}_0$ and $\mathcal{M}_j$ only differ when the state–action pair j is visited and the difference is only in the transition probabilities to the good or bad states, one can prove that

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_j) = \mathbb{E}_0[\mathbf{T}_j] \cdot kl(1/2, 1/2 + \Delta) \leq \mathbb{E}_0[\mathbf{T}_j] \cdot 4\Delta^2,$$

when $\Delta \leq 1/4$. Hence,

$$\frac{1}{K}\mathbb{E}_j[\mathbf{T}_j] \leq \frac{1}{K}\mathbb{E}_0[\mathbf{T}_j] + \sqrt{2}\Delta\sqrt{\mathbb{E}_0[\mathbf{T}_j]}.$$

Note that $\sum_{j \in \mathcal{J}} \mathbf{T}_j \leq K$. It follows from the Cauchy–Schwarz inequality that

$$\frac{1}{K}\sum_{j \in \mathcal{J}}\mathbb{E}_j[\mathbf{T}_j] \leq 1 + \sqrt{2}\Delta\sum_{j \in \mathcal{J}}\sqrt{\mathbb{E}_0[\mathbf{T}_j]} \leq 1 + \sqrt{2}\Delta\sqrt{LAK}. \tag{53}$$

- *Step 4. Optimizing Δ and finishing the proof.*

  We infer from (52) and (53) that

  $$\max_{j \in \mathcal{J}} R_j \geq K\mathbb{E}[(H - \gamma_d)^+]\Delta \cdot \left(1 - \frac{1}{LA}(1 + \sqrt{2}\Delta\sqrt{LAK})\right).$$

  This lower bound is maximized by taking $\Delta = \frac{1}{2\sqrt{2}}(1 - \frac{1}{LA})\sqrt{\frac{LA}{K}} \leq 1/4$ (by $K \geq SA/2$). Hence

  $$\max_{j \in \mathcal{J}} R_j \geq \frac{1}{4\sqrt{2}}\mathbb{E}[(H - \gamma_d)^+] \cdot \left(1 - \frac{1}{LA}\right)\sqrt{LAK}.$$

  Because $A \geq 2$ and $S \geq 6$, we have $L = \frac{1}{A} + (1 - \frac{1}{A})(S - 2) \geq S/4$. So we obtain

  $$\max_{j \in \mathcal{J}} R_j \geq \frac{1}{12\sqrt{2}}\mathbb{E}[(H - \gamma_d)^+] \cdot \sqrt{SAK}.$$

  Therefore, there exists a CTMDP $\mathcal{M}_j$ for some $j \in \mathcal{J}$ such that (19) holds. Finally, (20) follows from (19) by Jensen's inequality. The proof is complete. $\qquad\square$

7 Conclusions

In this paper we study RL for tabular CTMDPs with unknown parameters in the finite-horizon, episodic setting. We develop a learning algorithm and establish a worst-case regret upper bound. Meanwhile, we prove a regret lower bound, showing that the square-root regret rate achieved by our proposed algorithm actually has the optimal dependance on the numbers of episodes and actions. Numerical experiments are conducted to illustrate the performance of our learning algorithm.

Our work serves as a first step towards a better understanding of efficient episodic RL for CTMDPs. It is an open question how to improve our worst-case (or instance-independent) regret bounds and narrow the gap between the upper and lower bounds. There are also many other open problems, including, to name but a few, instance-dependent regret bounds, learning in CTMDPs with large or infinite state spaces, and episodic RL for semi-Markov decision processes that allow general holding time distributions. We leave them for future research.

References

Agarwal, A., N. Jiang, S. M. Kakade, and W. Sun (2021). Reinforcement learning: Theory and algorithms. *https://rltheorybook.github.io/rltheorybook_AJKS.pdf*.

Azar, M. G., I. Osband, and R. Munos (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR.

Basei, M., X. Guo, A. Hu, and Y. Zhang (2021). Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Available at SSRN 3848428*.

Bäuerle, N. and U. Rieder (2011). *Markov decision processes with applications to finance*. Springer Science & Business Media.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bradtke, S. J. and M. O. Duff (1995). Reinforcement learning methods for continuous-time markov decision. *Advances in Neural Information Processing Systems 7 7*, 393.

Canonne, C. L. (2017). A short note on poisson tail bounds. *http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf*.

Dann, C., L. Li, W. Wei, and E. Brunskill (2019). Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR.

Das, T. K., A. Gosavi, S. Mahadevan, and N. Marchalleck (1999). Solving semi-markov decision problems using average reward reinforcement learning. *Management Science 45*(4), 560–574.

Denardo, E. V. (1967). Contraction mappings in the theory underlying dynamic programming. *SIAM Review 9*(2), 165–177.

Domingues, O. D., P. Ménard, E. Kaufmann, and M. Valko (2021). Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR.

Fruit, R. and A. Lazaric (2017). Exploration-exploitation in MDPs with options. In *Artificial Intelligence and Statistics*, pp. 576–584. PMLR.

Gallego, G. and G. Van Ryzin (1994). Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science 40*(8), 999–1020.

Gao, X. and X. Y. Zhou (2022). Logarithmic regret bounds for continuous-time average-reward markov decision processes. *arXiv preprint arXiv:2205.11168*.

Garivier, A., P. Ménard, and G. Stoltz (2019). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research 44*(2), 377–399.

Guo, X. and O. Hernández-Lerma (2009). *Continuous time markov decision processes: theory and applications*. Springer-Verlag.

Guo, X., A. Hu, and Y. Zhang (2021). Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls. *arXiv preprint arXiv:2104.09311*.

Guo, X., X. Huang, and Y. Huang (2015). Finite-horizon optimality for continuous-time markov decision processes with unbounded transition rates. *Advances in Applied Probability 47*(4), 1064–1087.

Huang, Y. and X. Guo (2011). Finite horizon semi-markov decision processes with application to maintenance systems. *European Journal of Operational Research 212*(1), 131–140.

Jaksch, T., R. Ortner, and P. Auer (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research 11*(Apr), 1563–1600.

Jia, Y. and X. Y. Zhou (2022a). Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research 23*, (154)1–55.

Jia, Y. and X. Y. Zhou (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research 23*, (275)1–50.

Jia, Y. and X. Y. Zhou (2023). q-Learning in continuous time. *Journal of Machine Learning Research 24*, (161)1–61.

Jin, C., Z. Allen-Zhu, S. Bubeck, and M. I. Jordan (2018). Is Q-learning provably efficient? *Advances in Neural Information Processing Systems 31*.

Lattimore, T. and C. Szepesvári (2020). *Bandit algorithms*. Cambridge University Press.

Lippman, S. A. (1976). Countable-state, continuous-time dynamic programming with structure. *Operations Research 24*(3), 477–490.

Mamer, J. W. (1986). Successive approximations for finite horizon, semi-markov decision processes with application to asset liquidation. *Operations Research 34*(4), 638–644.

Miller, B. L. (1968). Finite state continuous time markov decision processes with a finite planning horizon. *SIAM Journal on Control 6*(2), 266–280.

Osband, I. and B. Van Roy (2017). Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pp. 2701–2710. PMLR.

Piunovskiy, A. and Y. Zhang (2020). Continuous-time markov decision processes. *Probability Theory and Stochastic Modelling*.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

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

Szpruch, L., T. Treetanthiploet, and Y. Zhang (2021). Exploration-exploitation trade-off for continuous-time episodic reinforcement learning with linear-convex models. *arXiv preprint arXiv:2112.10264*.

Tallec, C., L. Blier, and Y. Ollivier (2019). Making deep q-learning methods robust to time discretization. In *International Conference on Machine Learning*, pp. 6096–6104. PMLR.

Wang, H., T. Zariphopoulou, and X. Y. Zhou (2020). Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research 21*, (198)1–34.

Zanette, A. and E. Brunskill (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR.