Logarithmic regret bounds for continuous-time average-reward Markov decision processes

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

May 24, 2022

Abstract

We consider reinforcement learning for continuous-time Markov decision processes (MDPs) in the infinite-horizon, average-reward setting. In contrast to discrete-time MDPs, a continuous-time process moves to a state and stays there for a random holding time after an action is taken. With unknown transition probabilities and rates of exponential holding times, we derive instance-dependent regret lower bounds that are logarithmic in the time horizon. Moreover, we design a learning algorithm and establish a finite-time regret bound that achieves the logarithmic growth rate. Our analysis builds upon upper confidence reinforcement learning, a delicate estimation of the mean holding times, and stochastic comparison of point processes.

1   Introduction

Reinforcement learning (RL) is the problem of an agent learning how to map states to actions in order to maximize the reward over time in an unknown environment. It has received significant attention in the past decades, and the key challenge is in balancing the trade-off between exploration and exploitation (Sutton and Barto 2018). The common model for RL is a Markov Decision Process (MDP), which provides a mathematical framework for modeling sequential decision making problems under uncertainty. Most of the current studies on RL focus on developing algorithms and analysis for *discrete-time* MDPs. In contrast, less attention has been paid to *continuous-time* MDPs. However, there are many real-world applications where one *needs* to consider continuous-time MDPs. Examples include autonomous driving, control of queueing systems, high frequency trading and control of infectious diseases; see, e.g., Guo and Hernández-Lerma (2009), Chapter 11 of Puterman (2014) and the references therein. In this paper, we study learning in continuous-time MDPs (CTMDPs), aiming to understand fundamental performance limits of learning algorithms as well as to develop learning algorithms with theoretical guarantees.

In RL, there are typically three different settings/criteria: (1) infinite-horizon discounted reward setting, (2) infinite-horizon average-reward (ergodic) setting, and (3) finite-horizon episodic setting. In this paper, we focus on the infinite-horizon average-reward setting, where the agent aims to learn from data (states, actions and rewards) a policy that optimizes the long-run average reward. This setting has been heavily studied for learning in discrete-time MDPs; see e.g. Ortner and Auer (2007), Bartlett and Tewari (2009), Jaksch et al. (2010), Filippi et al. (2010), Fruit et al. (2018), Wei et al. (2020). More importantly, for many CTMDPs such as the optimal control of queueing

---

[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`.

systems and their applications, it is useful to study performance measures such as the average number of customers in the system and optimize the long-run average reward/cost as the systems often run for a long time; see e.g. Dai and Gluzman (2022).

In this paper, we focus on theoretical analysis and algorithm design for learning in tabular CTMDPs with finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$. Upon arriving at a state s, the agent takes an action a and receives some reward. The system remains in state s for a random holding time which is *exponentially* distributed with some unknown rate. Then it transits to another state with some unknown probability, and this series of events is repeated. The performance of the agent's algorithm is often measured by the total *regret* after some time period, [3] which is defined as the difference between the rewards collected by the algorithm during learning and the rewards of an optimal policy should the model parameters be completely known. If a learning algorithm achieves *sublinear* regret (in terms of the time horizon), then the average reward of the algorithm converges to the optimal average reward per unit time, while the convergence speed is determined by the specific growth rate of the regret.

While computational RL methods have been developed for CTMDPs as well as the more general Semi-Markov Decision Processes (SMDPs) in the earlier literature (Bradtke and Duff 1995, Das et al. 1999), available theoretical results on regret bounds for continuous-time RL are very limited. Several recent papers (Basei et al. 2021, Guo et al. 2021, Szpruch et al. 2021) study continuous-time RL for linear–quadratic/convex models and propose algorithms with sublinear regret upper bounds in the finite-horizon episodic setting. On the other hand, research on continuous-time RL for the infinite-horizon average reward setting is scarce. Fruit and Lazaric (2017) study learning in continuous-time 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 their algorithm achieves $O(\sqrt{n})$ regret after n decision steps (ignoring logarithmic factors and hiding dependency on other constants) in the infinite-horizon average reward setting. They also establish a minimax lower bound, in the sense that for any algorithm there exists an SMDP such that the expected regret is of $O(\sqrt{n})$.

These $\sqrt{n}-$type regret bounds are *worst-case* performance bounds without exploiting any special structure of the underlying CTMDPs; hence they tend to be overly pessimistic about the complexity of the learning problem (Simchowitz and Jamieson 2019). This motivates us to seek tighter regret bounds that are instance-dependent. The goal of the present paper is two fold. First, we derive instance-dependent regret lower bounds that are logarithmic in the time horizon T satisfied by *any* algorithm for learning average-reward CTMDPs. This provides the corresponding fundamental performance limits of learning algorithms. Second, we design learning algorithms that indeed achieve the logarithmic rate.

While instance-dependent logarithmic regret bounds have been studied for learning discrete-time MDPs in the average-reward setting (Burnetas and Katehakis 1997, Tewari and Bartlett 2007, Jaksch et al. 2010, Ok et al. 2018, Fruit et al. 2019), there are no such results for the continuous-time average-reward RL to our best knowledge. In this paper we fill this gap. Specifically, we first establish instance-dependent logarithmic regret lower bound for learning average-reward CTMDPs. Methodologically, our approach is related to Burnetas and Katehakis (1997). Since the set of

---

[3]Another popular performance measure is the sample complexity, which is the amount of data required to learn a near-optimal policy; see e.g. Brunskill and Li (2014), Jin and Sidford (2021), Wang (2017).

decision epochs of CTMDPs are still discrete, we first prove a regret lower bound that is logarithmic in the number of decision steps. The major difference and difficulty, compared with Burnetas and Katehakis (1997), is that now we need to take into account the random holding time in a state after an action is executed. The key idea to overcome this difficulty is, assuming the mean holding times are uniformly bounded, to bound the number of decision steps on an arbitrary time interval $[0, T]$ by invoking stochastic comparisons of point processes (Whitt 1981). This allows us to further obtain a logarithmic regret lower bound in the time horizon T.

For regret upper bound and algorithm design, we propose the CT-UCRL (continuous-time upper confidence reinforcement learning) algorithm for learning CTMDP and establish a finite-time instance-dependent upper bound for its regret. Our algorithm is a variant of the UCRL2 algorithm in Jaksch et al. (2010) for learning discrete-time MDPs although, again, we need to deal with the additional holding times inherent in CTMDPs. Our alogrithm is also related to the SMDP-UCRL algorithm for learning SMDPs in Fruit and Lazaric (2017). The main difference in the algorithm design is that we estimate the mean holding times with more refined estimators with tighter confidence bounds than the empirical mean in Fruit and Lazaric (2017). Note that the holding times follow exponential distributions, which are *unbounded and have heavier tails* than Gaussian distributions. As a consequence, the empirical mean holding times have exponential upper tails, leading to difficulties in the proof of logarithmic regret guarantees of learning algorithms, as is known in the multi-armed-bandit literature (Bubeck et al. 2013). In terms of theoretical results, Fruit and Lazaric (2017) consider subexponetial rewards and holding times and show that their algorithm achieves a regret upper bound of $O(\sqrt{n})$ after n decision steps (ignoring dependance on other constants). In contrast, we show that the CT-UCRL algorithm has a regret bound of $O(\frac{\log n}{g})$ where g is an instance-dependent gap parameter. To prove this gap-dependent logarithmic regret bound, we extend the approach of Jaksch et al. (2010) to the CTMDP setting, where the main difficulty is to analyze the random holding times. Moreover, we show that the expected regret of the CT-UCRL algorithm is also logarithmic in the time horizon T, again by using stochastic comparisons of point processes.

The remainder of the paper is organized as follows. In Section 2, we formulate the problem of learning in CTMDPs. In Section 3 we present the result on the asymptotic regret lower bound, while in Section 4 we devise the CT-UCRL algorithm and establish a finite-time instance-dependent upper bound for its regret. Finally Section 5 concludes. All the proofs are presented in the Appendix.

2   Formulation of Learning in CTMDPs

We consider a CTMDP with a finite state space $\mathcal{S}$ and a finite action space $\mathcal{A}$. Given a (deterministic, stationary, and Markov) *policy* π, which is a map from $\mathcal{S}$ to $\mathcal{A}$, the process moves as follows (Puterman 2014). At time 0, the system is at state $s_0 \in \mathcal{S}$ and the agent chooses an action $a_0 = \pi(s_0) \in \mathcal{A}$ with a reward $r(s_0, a_0) \in [0, 1]$ received as a consequence of the action. The system remains in state s_0 for a random holding time period τ_0 that follows an exponential distribution with parameter $\lambda(s_0, a_0)$. Then it jumps to state $s_1 \in \mathcal{S}$ with transition probability $p(s_1|s_0, a_0)$ at which another action $a_1 = \pi(s_1)$ is made. This series of events is repeated indefinitely. Given a policy π which, unless otherwise specified, is a map from $\mathcal{S}$ to $\mathcal{A}$ throughout this paper, denote by $\omega_n \equiv (s_0, a_0, \tau_0, s_1, a_1, \tau_1 \cdots, s_{n-1}, a_{n-1}, \tau_{n-1}, s_n)$ the history of the CTMDP, starting from s_0, up

to the nth decision epoch. Here, a major difference from discrete-time MDPs is that the history now contains the holding times $\tau_0, \tau_1, \cdots, \tau_{n-1}$. Denote by $N(t)$ the total number of actions made up to (and including) time $t > 0$.

We consider learning in an *unknown* CTMDP where the rates of exponential holding times $(\lambda(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$ and the transition probabilities $(p(\cdot|s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$ are unknown. For simplicity we assume the functional form of the reward function $(r(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$ is known.[4] A *learning algorithm*, denoted by $\mathcal{G}$, at each decision epoch $n \in \mathbb{N}$ generates a mapping π_n that maps the state s_n to an action $a_n = \pi_n(s_n)$. This mapping π_n itself could be determined based on the whole history w_n through for instance estimations of unknown model parameters. So a learning algorithm learns a policy *recursively* over decision epochs. The goal is to minimize the expected regret to be defined shortly.

Denote by $\mathcal{M} = (p(\cdot|s,a), \lambda(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$ the unknown parameters of the CTMDP. For notational simplicity, if a CTMDP is described by the quadruplet $(\mathcal{S}, \mathcal{A}, \mathcal{M}, r)$ where $r = (r(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$, it is referred to as the CTMDP $\mathcal{M}$.[5]

We make two assumptions on the CTMDP $\mathcal{M}$ throughout the paper. The first one is on the unknown transition probability matrix, denoted by $P = (p(j|s,a))_{s,j\in\mathcal{S},a\in\mathcal{A}}$. We are first given a family of sets $\mathcal{S}^+(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. These sets are assumed to be known to the agent and they are independent of $\mathcal{M}$. Now define

$$\Theta(s,a) = \left\{ q \in \mathbb{R}^{|\mathcal{S}|} : \sum_{y\in\mathcal{S}} q(y) = 1, q(y) > 0, \ \forall \ y \in \mathcal{S}^+(s,a) \text{ and } q(y) = 0, \ \forall \ y \notin \mathcal{S}^+(s,a) \right\}.$$

Following Burnetas and Katehakis (1997) (see Assumption (A) therein), we make the following assumption on the transition probability matrix P:

Assumption 1. *The transition probability vectors $p(\cdot|s,a) \in \Theta(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. In addition, the transition matrices $[p(j|s,\pi(s))]_{s,j\in\mathcal{S}}$ are irreducible for all deterministic, stationary and Markov policies π.*

Note that $p(\cdot|s,a) \in \Theta(s,a)$ implies that the support of the transition probability vector $p_s(a)$ is given by the known set $\mathcal{S}^+(s,a)$. For instance, with controlled birth–death processes, one may have $\mathcal{S}^+(s,a) = \{s+1, s-1\}$.

The second assumption is on the rates of the holding times.

Assumption 2. *We assume there are two known constants $\lambda_{\min}, \lambda_{\max} \in (0, \infty)$, such that*

$$\lambda_{\min} \le \lambda(s,a) \le \lambda_{\max}, \quad \text{for all } s \in \mathcal{S}, a \in \mathcal{A}.$$

---

[4]This assumption is not restrictive for the purpose of obtaining logarithmic regret bounds. For lower bounds, it can be relaxed following the discussion in Section 7 of Burnetas and Katehakis (1997) and Ok et al. (2018). For upper bounds, our proposed algorithm can be easily extended to incorporate the empirical estimates of the reward function following Jaksch et al. (2010), without changing the logarithmic growth rate of the regret.

[5]For simplicity, we only consider the 'lump sum reward' $r(s,a)$ in the CTMDP. More generally, one can also incorporate the 'running reward/cost' where the agent accrues a reward at rate $c(s', s, a)$ upon arriving at state s' after taking action a in the preceding state s. See Chapter 11 of Puterman (2014) for more details.

We denote by $\mathcal{H}$ the collection of CTMDPs that satisfy the above two assumptions.

Given a CTMDP $\mathcal{M}$, set

$$\rho^*(\mathcal{M}) = \max_\pi \left\{ \rho_{s_0}^\pi(\mathcal{M}) := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_{s_0}^\pi \sum_{n=0}^{N(T)-1} r(s_n, a_n) \right\}, \tag{1}$$

where the maximum is over all policies under the assumption of full knowledge of $\mathcal{M}$'s parameters. This quantity $\rho^*(\mathcal{M})$, or simply ρ^*, is the (ground truth) optimal long-run average reward. For a CTMDP satisfying Assumptions 1 and 2, there exist a function $h^* : \mathcal{S} \to \mathbb{R}$ and a constant ρ^* that is independent of the initial state s_0 such that the following Bellman optimality equation holds for the average reward CTMDP (1):

$$0 = \max_{a \in \mathcal{A}} \left\{ r(s,a) - \rho^*/\lambda(s,a) + \sum_{j \in \mathcal{S}} p(j|s,a) h^*(j) - h^*(s) \right\}, \quad s \in \mathcal{S}; \tag{2}$$

see, e.g., Chapter 11.5.3 of Puterman (2014). Here, ρ^* is the optimal average reward per unit time given in (1), and h^* is called the bias function or the relative value function (defined up to an additive constant). In addition, the maximizer of the right hand side of (2), which is a function of the state s, constitutes a stationary, deterministic Markov policy that is optimal for problem (1); see Chapter 11.5.3 of Puterman (2014) and Proposition 2 of Fruit and Lazaric (2017). We call this policy the optimal greedy policy. In the sequel, when we need to stress the dependence of ρ^* and $h^*(s)$ on the underlying CTMDP $\mathcal{M}$, we will write them as $\rho^*(\mathcal{M})$ and $h^*(s; \mathcal{M})$ respectively.

We measure the performance of a learning algorithm by the regret defined below.

Definition 1. *Given a learning algorithm $\mathcal{G}$ that generates action a_n at state s_n, its regret up to time T is defined by*

$$R_T^{\mathcal{G}}(s_0, \mathcal{M}) = T\rho^* - \sum_{n=0}^{N(T)-1} r(s_n, a_n), \tag{3}$$

where ρ^* is given by (1) and $\sum_{n=0}^{N(T)-1} r(s_n, a_n)$ is the accumulated reward collected under $\mathcal{G}$.

This notion of regret, proposed by Jaksch et al. (2010), is defined against the long-run average criterion ρ^*. Its relation to a more direct definition of regret is as follows. Consider the total expected reward under a policy π over horizon $T > 0$

$$V_T^\pi(s_0) = \mathbb{E}_{s_0}^\pi \sum_{n=0}^{N(T)-1} r(s_n, a_n)$$

and the corresponding optimal reward $V_T(s_0) = \sup_\pi V_T^\pi(s_0)$ when the true CTMDP $\mathcal{M}$ is known. Then the following would be a more natural definition of regret of the learning algorithm $\mathcal{G}$:

$$\hat{R}_T^{\mathcal{G}}(s_0, \mathcal{M}) = V_T(s_0) - \sum_{n=0}^{N(T)-1} r(s_n, a_n).$$

However, if there exists a constant C that is independent of T and s_0 such that

$$|V_T(s_0) - T\rho^*| \leq C, \tag{4}$$

then the gap between $\hat{R}_T^{\mathcal{G}}(s_0, \mathcal{M})$ and $R_T^{\mathcal{G}}(s_0, \mathcal{M})$ is $O(1)$ as $T \to \infty$. Hence, in this case if one can show $R_T^{\mathcal{G}}(s_0, \mathcal{M})$ is sublinear in T as T grows, then $\hat{R}_T^{\mathcal{G}}(s_0, \mathcal{M})$ is also sublinear in T with the same growth rate.

Lemma 1. *There exists a constant $C > 0$ independent of T and s_0 such that (4) holds for all s_0.*

This lemma therefore justifies the notion of regret defined by Definition 1. Its proof is given in Appendix A.

In the study of regret upper bounds, we will employ the well-known *uniformization* technique. Specifically, noting $\lambda(s, a) \leq \lambda_{\max}$ for all s, a under Assumption 2, we rewrite equation (2) as follows (Puterman 2014, Chapter 11.5.3)

$$\check{h}^*(s) = \max_{a \in \mathcal{A}} \left\{ \check{r}(s, a) - \check{\rho}^* + \sum_{j \in \mathcal{S}} \check{p}(j|s, a) \check{h}^*(j) \right\}, \tag{5}$$

where $\check{r}(s, a) = r(s, a)\lambda(s, a)/\lambda_{\max}$, $\check{\rho}^* = \rho^*/\lambda_{\max}$, $\check{h}^* = h^*$, $\check{p}(j|s, a) = p(j|s, a)\lambda(s, a)/\lambda_{\max}$ for $j \neq s$ and $\check{p}(j|s, a) = 1 - (1 - p(j|s, a))\lambda(s, a)/\lambda_{\max}$ for $j = s$. This is called the uniformization transformation. This transformation leads to a new but equivalent CTMDP where the decision times are determined by the arrival times of a homogeneous Poisson process with uniform rate $\lambda_{\max}$ that no longer depends on the state. Moreover, we recognize that (5) is the optimality equation for an average-reward *discrete-time* MDP with reward function $\check{r}$ and transition probabilities $(\check{p}(j|s, a))_{s,j,a}$. This implies that the original CTMDP model (1) is also equivalent to a (uniformized) discrete-time MDP model under the *infinite-horizon* average-reward criterion. This being said, we can *not* directly apply the existing regret bound results for discrete-time MDPs to obtain regret bounds for the CTMDPs we consider. This is because we are ultimately interested in how the total regret of a learning algorithm grows as a function of time, whereas on a *finite horizon* the continuous-time and uniformized discrete-time MDPs are *not* equivalent – indeed it is unclear how the gap between the two problems scale as a function of time. This represents the major challenge we need to overcome in our regret analysis.

3 Asymptotic Regret Lower Bound

In this section we establish (asymptotic) logarithmic regret lower bound for learning CTMDPs. Lower bounds are important as they provide fundamental performance limits of learning algorithms.

3.1 Preliminaries

We first introduce some definitions and quantities that will be used later.

A learning algorithm $\mathcal{G}$ is called *uniformly fast convergent* (UF) if

$$\mathbb{E}_{s_0}[R_T^{\mathcal{G}}(s_0, \mathcal{M})] = o(T^{\alpha}) \quad \text{as } T \to \infty, \tag{6}$$

for any $\alpha > 0, s_0 \in \mathcal{S}$ and any CTMDP $\mathcal{M} \in \mathcal{H}$. Denote by C_F the class of UF algorithms. *A priori* it is not clear whether or not the set C_F is empty; however, we will develop an algorithm in Section 4 which shows that UF algorithms indeed exist in average-reward CTMDPs. On the other hand, a UF algorithm is defined to perform uniformly well on all CTMDP instances, but may not perform arbitrarily well on a specific instance; so it is used to derive a lower bound (Theorem 1).

For a learning algorithm $\mathcal{G}$, define the expected regret up to $N-$th decision epoch S_N by

$$\mathcal{R}_N^{\mathcal{G}}(s_0, \mathcal{M}) := \mathbb{E}_{s_0}[R_{S_N}^{\mathcal{G}}(s_0, \mathcal{M})] = \mathbb{E}_{s_0}[\rho^* \cdot S_N - \sum_{n=0}^{N-1} r(s_n, a_n)], \tag{7}$$

where $R_T^{\mathcal{G}}(s_0, \mathcal{M})$ is defined in (3) for any $T > 0$.

Let $O(s; \mathcal{M})$ be the set of maximizers of the right hand side of the equation (2). Define the "suboptimality gap" of an action a at state s by

$$\phi^*(s, a; \mathcal{M}) = -r(s, a) + \rho^*/\lambda(s, a) - \sum_{j \in \mathcal{S}} p(j|s, a)h^*(j) + h^*(s). \tag{8}$$

Clearly, $\phi^*(s, a; \mathcal{M}) = 0$ if $a \in O(s; \mathcal{M})$, and $\phi^*(s, a; \mathcal{M}) > 0$ for any s and $a \notin O(s; \mathcal{M})$.

The following are natural extensions of some critical state–action pairs and related quantities from discrete-time MDPs in Burnetas and Katehakis (1997) to their continuous-time counterparts.

- For (s, a) such that $a \notin O(s; \mathcal{M})$, let $\overline{\mathcal{M}} \in \mathcal{H}$ be the modification of the original model $\mathcal{M}$ in the following manner: the transition probability distribution $p(\cdot|s, a)$ is modified to $\bar{p}(\cdot|s, a) \in \Theta(s, a)$, the rate of holding time $\lambda(s, a)$ is modified to $\bar{\lambda}(s, a) \in [\lambda_{\min}, \lambda_{\max}]$, and all the other parameters remain unchanged. When we need to stress the dependence of $\overline{\mathcal{M}}$ on $s, a, \mathcal{M}, \bar{p}(\cdot|s, a), \bar{\lambda}$, we will write it as $\overline{\mathcal{M}}(s, a; \mathcal{M}, \bar{p}(\cdot|s, a), \bar{\lambda}(s, a))$.

- For (s, a) with $a \notin O(s; \mathcal{M})$, denote by $\Delta\Theta(s, a; \mathcal{M})$ the set of parameter values $(\bar{p}(\cdot|s, a), \bar{\lambda}(s, a))$ that make action a at state s uniquely optimal under the modified model $\overline{\mathcal{M}}$, and by $B(\mathcal{M})$ the set of critical state-action pairs where $B(\mathcal{M}) = \{(s, a) : a \notin O(s; \mathcal{M}), \Delta\Theta(s, a; \mathcal{M}) \neq \emptyset\}$.

- Define the 'distance' between the two models $\mathcal{M}$ and $\overline{\mathcal{M}}$ at (s, a):

$$I([p(\cdot|s, a), \lambda(s, a)], [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)])$$
$$= KL(p(\cdot|s, a), \bar{p}(\cdot|s, a)) + KL(\tau(s, a), \bar{\tau}(s, a))$$
$$= \sum_{j \in \mathcal{S}} p(j|s, a) \log(p(j|s, a)/\bar{p}(j|s, a)) + \log(\lambda(s, a)/\bar{\lambda}(s, a)) + \bar{\lambda}(s, a)/\lambda(s, a) - 1,$$

where $\tau(s, a)$ and $\bar{\tau}(s, a)$ are the exponential distributions with parameters $\lambda(s, a)$ and $\bar{\lambda}(s, a)$ respectively, and $KL(\cdot, \cdot)$ is the Kullback–Leibler (KL) divergence between two probability distributions. Note that $KL(p(\cdot|s, a), \bar{p}(\cdot|s, a))$ is well defined since both $p(\cdot|s, a)$ and $\bar{p}(\cdot|s, a)$ are in the set $\Theta(s, a)$, and we use the fact that the KL divergence of two exponential distributions with parameters α and β is $\log(\alpha/\beta) + \beta/\alpha - 1$.

- Let

$$K(s, a; \mathcal{M}) = \inf\{I([p(\cdot|s, a), \lambda(s, a)], [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)]) :$$
$$[\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)] \in \Delta\Theta(x, a; \mathcal{M})\}. \tag{9}$$

For fixed (s, a), $K(s, a; \mathcal{M}) = 0$ if $a \in O(s; \mathcal{M})$. When $a \notin O(s; \mathcal{M})$, $0 < K(s, a; \mathcal{M}) < \infty$ if $(s, a) \in B(\mathcal{M})$ and $K(s, a; \mathcal{M}) = \infty$ if $(s, a) \notin B(\mathcal{M})$. As we will see later, $K(s, a; \mathcal{M})$ is a measure of the importance of the critical pair (s, a).

- Define

$$C(\mathcal{M}) = \sum_{(s,a) \in B(\mathcal{M})} \frac{\phi^*(s, a; \mathcal{M})}{K(s, a; \mathcal{M})}, \tag{10}$$

with the convention that the fraction is 0 if both the numerator and denominator are 0. This constant $C(\mathcal{M})$ provides an aggregate measure of importance of all the critical state-action pairs, which in turn characterizes the level of difficulty for learning in the CTMDP $\mathcal{M}$ as we will see in Theorem 1. Note that $C(\mathcal{M}) = 0$ only in the degenerate cases when all policies are optimal (i.e. $a \in O(x; \mathcal{M})$ for all a) and/or none of the non-optimal actions can be made optimal by changing only its transition probability vector and the rate of holding times (i.e. $\Delta\Theta(s, a; \mathcal{M}) = \emptyset$).

3.2   Main result

This section provides the main result on the asymptotic lower bound for the regret.

Theorem 1 (Logarithmic instance-dependent regret lower bound)**.** *For any learning algorithm $\mathcal{G}$ that is UF and any CTMDP $\mathcal{M} \in \mathcal{H}$, the expected regret up to $N-$th decision epoch satisfies*

$$\liminf_{N \to \infty} \frac{\mathcal{R}_N^{\mathcal{G}}(s_0, \mathcal{M})}{\log N} \geq C(\mathcal{M}), \tag{11}$$

where the instance-dependent constant $C(\mathcal{M})$ is given in (10). Moreover

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{s_0}[R_T^{\mathcal{G}}(s_0, \mathcal{M})]}{\log T} \geq C(\mathcal{M}). \tag{12}$$

A proof of Theorem 1 is deferred to Appendix B.

Remark 1. *In deriving the regret lower bound, the assumption that the holding time is exponential is used only at two places: Proof of (12) and Proof of Lemma 4 in Appendix B. In both places, the exponential distribution facilitates the comparison of counting processes and the conversion of the regret bound in terms of discrete decision epoch N to the regret bound of continuous time T. Specifically, together with Assumption 2 (which involves only bounds on mean holding times), the exponential assumption allows us to bound the counting process associated with the number of decision epochs under any learning algorithm by tractable Poisson processes. On the other hand, it is possible to extend (12) to the more general semi-Markov decision processes which extend CTMDPs*

8

by allowing general holding time distributions. In such an extension, one needs extra assumptions on the holding time distributions so that the aforementioned comparison of counting processes still works. For instance, one may need bounds on the failure rate of the residual life time distribution of the counting process associated with decision epochs under any learning algorithm (Whitt 1981). This in turn requires more information about the distribution of the holding times beyond the mere bounds on the mean. Here we do not pursue this extension in this paper.

To understand how the instance-dependent constant $C(\mathcal{M})$ in the regret lower bound scales with the sizes of state and action spaces along with the other model parameters, we present the following result, which provides an upper bound on $C(\mathcal{M})$. Write $S = |\mathcal{S}|$, $A = |\mathcal{A}|$, and define $H := \max_{s \in \mathcal{S}} h^*(s; \mathcal{M}) - \min_{s \in \mathcal{S}} h^*(s; \mathcal{M})$, where we recall that h^* is the bias function in the average reward optimality equation (2).

Proposition 1. *We have*

$$C(\mathcal{M}) \leq \frac{(H + 2\lambda_{\max})^2 SA}{\min_{(s,a) \in B(\mathcal{M})} \phi^*(s, a; \mathcal{M})} \cdot \frac{\lambda_{\max}}{\lambda_{\min}^3}.$$

A proof of this result is given in Appendix C.

4 The CT–UCRL Algorithm and Its Instance-Dependent Regret Upper Bound

In this section we present the CT-UCRL algorithm for learning in CTMDP and establish a finite-time instance-dependent upper bound for its regret.

First, we introduce the following definition from Jaksch et al. (2010) and Fruit and Lazaric (2017), which is a real-valued measure of the connectedness of an MDP.

Definition 2. *The diameter $D(\mathcal{M})$ of an MDP $\mathcal{M}$, either continuous time or discrete time, is defined by*

$$D(\mathcal{M}) = \max_{s,s' \in \mathcal{S}} \left\{ \min_{\pi:\mathcal{S} \to \mathcal{A}} \mathbb{E}^\pi [T(s')|s_0 = s] \right\} \tag{13}$$

where $T(s')$ is the first time when state s' is reached.

Remark 2. *It is known that the diameter of a discrete-time MDP with finite state space is finite if and only if the MDP is communicating (i.e. for any pair of states s, s', there exists a policy under which the probability of eventually reaching s' starting from s is positive); see Jaksch et al. (2010). In our case, the diameter $D(\mathcal{M})$ of the CTMDP $\mathcal{M}$ is also finite. This is because by Assumption 2 we can uniformize the CTMDP to obtain an equivalent discrete-time MDP $\mathcal{M}^{eq}$. It follows from the proof of Lemma 6 in Fruit and Lazaric (2017) that $D(\mathcal{M}) = D(\mathcal{M}^{eq})/\lambda_{\max}$. By Assumption 1, $\mathcal{M}^{eq}$ is communicating; hence $D(\mathcal{M}^{eq})$ is finite.*

In the following, we first put forward the CT-UCRL algorithm in Section 4.1, and then present its logarithmic regret upper bound in Section 4.2.

4.1 The CT-UCRL algorithm

In this section we present the CT-UCRL algorithm.

4.1.1 Refined estimator for mean holding time

In the CT-UCRL algorithm we need to estimate the unknown quantities, i.e., transition probabilities and mean holding times, hopefully with tight confidence bounds. For the former, one can simply use the empirical transition probabilities as in Jaksch et al. (2010). However, in order to obtain logarithmic regret bounds, the estimation of the mean holding times requires extra delicate analysis because the holding time follows an exponential distribution which has tails heavier than those of a Gaussian distribution. This represents one of the main differences and difficulties in treating the continuous-time case.

Specifically, for a given state-action pair (s, a), suppose $(X_i)_{i=1}^n$ are i.i.d exponential holding times with mean $1/\lambda(s, a)$. In the following we omit the dependancy of λ on (s, a) for notational simplicity. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ be the sample mean. Then it is known that the upper tail of the sample mean $\bar{X}$ is exponential instead of Gaussian, though the lower tail is Gaussian. Mathematically, we have for any $x > 0$,

$$\mathbb{P}\left(\lambda \bar{X} - 1 \geq \frac{x}{n} + \sqrt{\frac{2x}{n}}\right) \leq e^{-x}, \quad \text{and} \quad \mathbb{P}\left(\lambda \bar{X} - 1 \leq -\sqrt{\frac{2x}{n}}\right) \leq e^{-x}.$$

However, to establish the logarithmic regret bound of the CT-UCRL algorithm, it turns out in our analysis we need an estimator $\hat{X}$ (based on n i.i.d observations) for the mean holding time $1/\lambda$ with two-sided Gaussian tail bounds. Hence, we need to replace the empirical mean by other refined estimators. One simple choice (see e.g. Bubeck et al. 2013) is the truncated empirical mean $\hat{X}$ defined by

$$\hat{X} = \frac{1}{n} \sum_{i=1}^n X_i \cdot 1_{X_i \leq \sqrt{\frac{2i}{\lambda_{\min}^2 \log(1/\delta)}}}, \tag{14}$$

where the constant $2/\lambda_{\min}^2$ acts as an upper bound of the second moment $\mathbb{E}[X_i^2] = \frac{2}{\lambda^2}$ by Assumption 2. Lemma 1 of Bubeck et al. (2013) shows that for $\delta \in (0, 1)$, the following estimate holds with probability at least $1 - \delta$:

$$\left|\hat{X} - \frac{1}{\lambda}\right| \leq \frac{4}{\lambda_{\min}} \sqrt{\frac{2 \log(1/\delta)}{n}}. \tag{15}$$

Our proof of the logarithmic regret of the CT-CURL algorithm crucially relies on (15).

Remark 3. *Instead of the truncated empirical mean, one can also use other refined mean estimators such as the median of means and Catoni's M estimator. These estimators have similar performance guarantees as (15) with the truncated empirical mean; see Bubeck et al. (2013) for a detailed discussion about these estimators for the purpose of deriving logarithmic regret for multi-armed bandits with heavy tailed rewards. We take the truncated empirical mean estimator in this paper due to its simplicity and the low computational cost per update.*

4.1.2 The CT-UCRL algorithm

The CT-UCRL learning algorithm in CTMDPs is presented as Algorithm 1. Let us provide an overview of the algorithm and then elaborate some details.

The algorithm proceeds in episodes $k = 1, 2, 3, \ldots$ with variable lengths. At the start of episode k, the algorithm constructs estimators and confidence intervals for the unknown transition probabilities and mean holding times; see (16)–(17). These lead to a set of statistically plausible CTMDPs denoted by $\mathcal{C}_k$. The algorithm finds a CTMDP $\widetilde{M}_k \in \mathcal{C}_k$, referred to as the optimistic CTMDP, that maximizes the average reward among the plausible CTMDPs in certain sense (to be specified shortly) and computes the corresponding optimal greedy policy $\tilde{\pi}_k$; see (18). The details of this computation will be discussed later. The policy $\tilde{\pi}_k$ is executed until the end of episode k when the number of visits to some state-action pair doubles (Step 6 of Algorithm 1). More specifically, episode k ends when for some state action pair (s, a), the number of visits to (s, a) in episode k, $v_k(s, a)$, equals $\max\{1, N_k(s, a)\}$ which is the total number of visits to (s, a) up to and including episode $k - 1$. Then a new episode starts and the whole process repeats.

We now discuss (18), the problem of finding the CTMDP model in the confidence region with the nearly largest average reward. This problem can be solved by introducing an "extended" CTMDP that combines all the CTMDPs in the confidence region and finding the average-reward optimal policy on the extended CTMDP. The solution procedure, known as the extended value iteration, is discussed in details in Fruit and Lazaric (2017) for SMDPs, though their confidence region differs from ours. Here we outline the key ideas for readers' convenience.

Introduce the extended CTMDP $\widetilde{M}_k^+$ with the state space $\mathcal{S}$ and a (state-dependent) action space $\mathcal{A}_s^+$ defined as

$$\mathcal{A}_s^+ = \{(a, p(\cdot|s, a), \lambda(s, a)) : a \in \mathcal{A}, (p(\cdot|s, a), \lambda(s, a)) \in \mathcal{P}_k(s, a) \times C_k(s, a)\},$$

where $\mathcal{P}_k(s, a)$ and $C_k(s, a)$ are the confidence sets in (16) and (17) respectively for a given state-action pair (s, a). The reward function, transition probabilities and rates of holding times associated $\mathcal{A}_s^+$ are specified as follows. At a given state–action pair $(s, a_s^+) \equiv (s, a, p(\cdot|s, a), \lambda(s, a)) \in \mathcal{S} \times \mathcal{A}_s^+$, the reward is $r^+(s, a_s^+) = r(s, a)$, the transition probabilities are $p^+(\cdot|s, a_s^+) = p(\cdot|s, a)$, and the rates of holding times are $\lambda^+(s, a_s^+) = \lambda(s, a)$. Then, following a similar discussion as in Section 3.1.1 of Jaksch et al. (2010), we conclude that finding a CTMDP $\widetilde{M} \in \mathcal{C}_k$ and a policy $\tilde{\pi}$ on $\widetilde{M}$ such that $\rho_s^{\tilde{\pi}}(\widetilde{M}) = \max_{s, \pi, \mathcal{M}' \in \mathcal{C}_k} \rho_s^{\pi}(\mathcal{M}')$ for all initial state s corresponds to finding the average-reward optimal policy on $\widetilde{M}_k^+$. Specifically, for any given CTMDP $\mathcal{M}' \in \mathcal{C}_k$ and any policy $\pi' : \mathcal{S} \to \mathcal{A}$ on $\mathcal{M}'$, there is a policy π^+ on the extended CTMDP $\widetilde{M}_k^+$ such that the same transition probabilities, holding-time rates and rewards are induced by π' on $\mathcal{M}'$ and π^+ on $\widetilde{M}_k^+$. On the other hand, for each policy π^+ on $\widetilde{M}_k^+$, there is a CTMDP $\mathcal{M}' \in \mathcal{C}_k$ and a policy π' on it so that again the same transition probabilities, holding-time rates and rewards are induced.

It remains to discuss how to solve the extended CTMDP $\widetilde{M}_k^+$ under the average-reward criteria for each episode k. One can apply the uniformization procedure, convert the extended CTMDP to an equivalent discrete-time MDP model $\widetilde{M}_{k,eq}^+$, and apply a value iteration scheme to solve $\widetilde{M}_{k,eq}^+$ *approximately*. Specifically, denote the state values of iteration i by $u_i(s)$ for $s \in \mathcal{S}$. Then the value

Algorithm 1 The CT-UCRL Algorithm

Input: Confidence parameter $\delta \in (0,1)$, $\lambda_{\min}, \lambda_{\max} \in (0,\infty)$, $\mathcal{S}, \mathcal{A}$ and reward function r

1: Initialization: set $n = 1$ and observe initial state s_1.

2: **for** episode $k = 1, 2, 3, \ldots$ **do**

3:    **Initialize episode k:**

    (a)   Set the start decision epoch of episode k, $t_k := n$

    (b)   For all (s,a) initialize the state-action counter for episode k, $v_k(s,a) = 0$. Further set the state-action counts prior to episode k as $N_k(s,a) := \#\{i < t_k : s_i = s, a_i = s\}$.

    (c)   For $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, set the observed accumulated transition counts piror to episode k as $N_k(s,a,s') := \#\{i < t_k : s_i = s, a_i = s, s_{i+1} = s'\}$. Compute the empirical transition probabilities: $\hat{p}(s'|s,a) = \frac{N_k(s,a,s')}{\max\{1, N_k(s,a)\}}$. Also compute the *truncated* empirical mean $\frac{1}{\hat{\lambda}(s,a)}$ for the holding time by (following (14)):

$$\frac{1}{\hat{\lambda}(s,a)} = \frac{1}{N_k(s,a)} \sum_{j=1}^{t_k - 1} \tau_j \cdot 1_{s_j = s, a_j = a} \cdot 1_{\tau_j \leq \sqrt{\frac{2 T_j(s,a)}{\lambda_{\min}^2 \log(1/\delta)}}},$$

       where $T_j(s,a) := \#\{i \leq j : s_i = s, a_i = s\}$ denotes the state-action counter up to decision epoch j.

4:    **Compute policy $\tilde{\pi}_k$:**

    (a)   Let $\mathcal{C}_k$ be the set of all CTMDPs with states and actions as $\mathcal{M}$, and with transition probabilities $\tilde{p}(\cdot|s,a)$ and holding time rate parameter $\tilde{\lambda}(s,a)$ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\|\tilde{p}(\cdot|s,a) - \hat{p}(\cdot|s,a)\|_1 \leq \sqrt{\frac{14 S \log(2 A t_k/\delta)}{\max\{1, N_k(s,a)\}}} \tag{16}$$

$$\left| \frac{1}{\tilde{\lambda}(s,a)} - \frac{1}{\hat{\lambda}(s,a)} \right| \leq \frac{4}{\lambda_{\min}} \sqrt{\frac{14 \log(2 A S t_k/\delta)}{\max\{1, N_k(s,a)\}}} \quad \text{and} \quad \tilde{\lambda}(s,a) \in [\lambda_{\min}, \lambda_{\max}]. \tag{17}$$

    (b)   Find a CTMDP $\widetilde{M}_k \in \mathcal{C}_k$ and a policy $\tilde{\pi}_k$ reward such that

$$\rho_k^* := \min_{s \in \mathcal{S}} \rho_s^{\tilde{\pi}_k}(\widetilde{M}_k) \geq \max_{s \in \mathcal{S}} \max_{\pi} \max_{\mathcal{M}' \in \mathcal{C}_k} \rho_s^{\pi}(\mathcal{M}') - \frac{1}{\sqrt{t_k}}. \tag{18}$$

5:    **Execute policy $\tilde{\pi}_k$:**

6:    **while** $v_k(s_n, \tilde{\pi}_k(s_n)) < \max\{1, N_k(s_n, \tilde{\pi}_k(s_n))\}$ **do**

    (a)   Choose the action $a_n = \tilde{\pi}_k(s_n)$, observe the holding time τ_n and the next state s_{n+1}

    (b)   Update $v_k(s_n, a_n) = v_k(s_n, a_n) + 1$, and set $n = n + 1$. Here n is decision epoch.

7: **end for**

---

iteration on $\widetilde{M}_{k,eq}^+$ becomes the following

$$u_{i+1}(s)$$

$$= \max_{a_s^+ \in \mathcal{A}_s^+} \left\{ \check{r}(s, a_s^+) + \sum_{j \in \mathcal{S}} \check{p}(j|s, a_s^+) u_i(j) \right\}$$

$$= \max_{a \in \mathcal{A}} \left\{ \max_{\lambda(s,a) \in C_k(s,a)} \left\{ r(s,a) \frac{\lambda(s,a)}{\lambda_{\max}} + \left[\max_{p(\cdot|s,a) \in \mathcal{P}_k(s,a)} \sum_{j \in \mathcal{S}} p(j|s,a) u_i(j) - u_i(s) \right] \frac{\lambda(s,a)}{\lambda_{\max}} \right\} \right\}$$

$$+ u_i(s), \quad \forall s \in \mathcal{S}, \tag{19}$$

where from uniformation one has $\check{r}(s, a_s^+) = r^+(s, a_s^+) \lambda(s,a)/\lambda_{\max} = r(s,a)\lambda(s,a)/\lambda_{\max}$, $\check{p}(j|s, a_s^+) = p(j|s,a)\lambda(s,a)/\lambda_{\max}$ for $j \neq s$ and $\check{p}(s|s, a_s^+) = 1 - (1 - p(s|s,a))\lambda(s,a)/\lambda_{\max}$. If one stops the value iteration whenever

$$\max_{s \in \mathcal{S}} \{u_{i+1}(s) - u_i(s)\} - \min_{s \in \mathcal{S}} \{u_{i+1}(s) - u_i(s)\} < \varepsilon := \frac{1}{\sqrt{t_k}}, \tag{20}$$

then the greedy policy with respect to u_i is ε−optimal for the extended CTMDP $\widetilde{M}_k^+$ which yields the nearly optimistic policy $\tilde{\pi}_k$ in (18). In addition, the optimistic CTMDP $\widetilde{M}_k = (\widetilde{p}_k(\cdot|s,a), \widetilde{\lambda}_k(s,a))_{s \in \mathcal{S}, a \in \mathcal{A}}$ is obtained by solving the two inner optimizations in (19). The optimization over $p(\cdot|s,a)$ can be solved using the algorithm in Figure 2 of Jaksch et al. (2010). Given the solution to this problem and u_i, one can easily solve the optimization over $\lambda(s,a) \in \Lambda_k$ since the objective is linear in $\lambda(s,a)$ and the domain $C_k(s,a)$ is bounded with $\lambda(s,a) \in [\lambda_{\min}, \lambda_{\max}]$. Note that the stopping condition (20) is reached in a finite number of steps, say i, and one has $|u_{i+1} - u_i(s) - \rho_k^*| < \frac{1}{\sqrt{t_k}}$ where ρ_k^* is given in (18); see Lemmas 1 and 7 of Fruit and Lazaric (2017) for details. We also remark that by taking the minimum over $s \in \mathcal{S}$ on the left-hand side of (18), it ensures that the gap between $\rho_s^{\tilde{\pi}_k}(\widetilde{M}_k)$ and $\max_{s,\pi,\mathcal{M}' \in \mathcal{C}_k} \rho_s^\pi(\mathcal{M}')$ is at most $\frac{1}{\sqrt{t_k}}$ for all initial states s.

4.2 Regret upper bound

We now present the logarithmic upper bound of the CT-UCRL algorithm (Algorithm 1). Recall $\rho_s^\pi(\mathcal{M})$ and $\rho^*(\mathcal{M})$ defined in equation (1). Set

$$g = \rho^*(\mathcal{M}) - \max_{s \in \mathcal{S}} \max_{\pi:\mathcal{S} \to \mathcal{A}} \{\rho_s^\pi(\mathcal{M}) : \rho_s^\pi(\mathcal{M}) < \rho^*(\mathcal{M})\}, \tag{21}$$

which is the gap in the average reward between the best and second best policy in the CTMDP $\mathcal{M}$. Set $\delta = 1/N$ in Algorithm 1 where N is the number of decision epochs.

Theorem 2 (Logarithmic instance-dependent regret upper bound). *For any CTMDP $\mathcal{M} \in \mathcal{H}$, any initial state s_0 and any $N \geq 1$, the expected regret of the algorithm $\bar{\mathcal{G}} := CT\text{-}UCRL$ up to the N-th decision epoch satisfies*

$$\mathcal{R}_N^{\bar{\mathcal{G}}}(s_0, \mathcal{M}) \leq 4C \cdot \frac{\log(N)}{g} + C', \tag{22}$$

where C' is a constant independent of N and

$$C = 3\left(34^2\lambda_{\max}^2 D(\mathcal{M})^2 S^2 A + 2\cdot 73^2 \frac{\lambda_{\max}^2}{\lambda_{\min}^2}SA + \frac{24SA}{\lambda_{\min}^2}\right) \tag{23}$$

with $D(\mathcal{M}) < \infty$ being the diameter of the CTMDP $\mathcal{M}$. Moreover

$$\mathbb{E}_{s_0}[R_T^{\bar{\mathcal{G}}}(s_0, \mathcal{M})] \leq \frac{4C}{g}\cdot\log\left(\lambda_{\max}T+2\right)+C'+\frac{\lambda_{\max}}{\lambda_{\min}}. \tag{24}$$

A proof of this result is deferred to Appendix D, where the constant C' will be explicitly given; see (63).

Remark 4. *For a discrete-time communicating MDP with S states and A actions, the UCRL2 algorithm proposed in Jaksch et al. (2010) has a instance-dependent regret upper bound of $O(\frac{D^2 S^2 A \log T}{g^*})$ in the average-reward setting, where D is the diameter of the MDP and g^* is the gap parameter that can be defined similarly to (21) for discrete-time MDPs. Equation (24) in Theorem 2 suggests that the instance-dependent regret upper bound of the CT-UCRL algorithm for learning CTMDPs is given by*

$$O\left(\frac{\lambda_{\max}^2 D(\mathcal{M})^2 S^2 A + \left[\frac{\lambda_{\max}^2+1}{\lambda_{\min}^2}\right]SA}{g}\cdot\log T\right).$$

Compared with the results for discrete-time MDPs, there are several differences. First, there is an additional term, $\left[\frac{\lambda_{\max}^2+1}{\lambda_{\min}^2}\right]SA\log T$, which partly quantities the extra difficulty in learning CTMDPs due to the random exponential holding times. Second, the coefficient of $S^2 A$ now depends on both the diameter of the CTMDP $D(\mathcal{M})$ and $\lambda_{\max}$. Finally, the gap parameter g in (21) is associated with CTMDPs and hence depends on the rates of holding times in general.

5 Conclusion

In this paper, we study RL for continuous-time average-reward Markov decision processes with unknown parameters. We establish instance-dependent logarithmic regret lower bounds that represent fundamental performance limits of learning algorithms. Meanwhile we devise the CT-UCRL algorithm and prove a non-asymptotic logarithmic regret upper bound. The analysis for deriving these results are substantially different from and more difficult than its discrete-time counterpart due to the presence of the random holding time at a state.

The study of regret analysis for RL in continuous time is still rare and far between. There are many open questions. A significant one is to extend the current setting to one with continuous space, e.g., where the dynamic is governed by a diffusion process. RL for diffusion processes has been studied recently; see e.g. Wang et al. (2020), Jia and Zhou (2022a,b), but these papers focus on the problems of generating trial-and-error policies, police evaluation and policy improvement, while that of the regret bounds remains completely untouched. Other research directions include non-asymptotic instance-dependent regret lower bound and finite-horizon episodic setting.

References

Bartlett, P. and A. Tewari (2009). Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*, pp. 35–42. AUAI Press.

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*.

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.

Brunskill, E. and L. Li (2014). Pac-inspired option discovery in lifelong reinforcement learning. In *International conference on machine learning*, pp. 316–324. PMLR.

Bubeck, S., N. Cesa-Bianchi, and G. Lugosi (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory 59*(11), 7711–7717.

Burnetas, A. N. and M. N. Katehakis (1997). Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research 22*(1), 222–255.

Dai, J. G. and M. Gluzman (2022). Queueing network controls via deep reinforcement learning. *Stochastic Systems 12*(1), 30–67.

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.

Filippi, S., O. Cappé, and A. Garivier (2010). Optimism in reinforcement learning and kullback-leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 115–122. IEEE.

Fruit, R. and A. Lazaric (2017). Exploration-exploitation in MDPs with options. In *Artificial Intelligence and Statistics*, pp. 576–584. PMLR, `http://proceedings.mlr.press/v54/fruit17a/fruit17a-supp.pdf`.

Fruit, R., A. Lazaric, and M. Pirotta (2019). Regret minimization in infinite-horizon finite markov decision processes. Tutorial at ALT'19.

Fruit, R., M. Pirotta, A. Lazaric, and R. Ortner (2018). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1578–1586. PMLR.

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*.

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*.

Jia, Y. and X. Y. Zhou (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *arXiv preprint arXiv:2111.11232*.

Jin, Y. and A. Sidford (2021). Towards tight bounds on the sample complexity of average-reward mdps. *arXiv preprint arXiv:2106.07046*.

Ok, J., A. Proutiere, and D. Tranos (2018). Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems 31*.

Ortner, P. and R. Auer (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems 19*, 49.

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

Simchowitz, M. and K. G. Jamieson (2019). Non-asymptotic gap-dependent regret bounds for tabular MDPs. *Advances in Neural Information Processing Systems 32*, 1153–1162.

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*.

Tewari, A. and P. Bartlett (2007). Optimistic linear programming gives logarithmic regret for irreducible MDPs. *Advances in Neural Information Processing Systems 20*.

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*, 1–34.

Wang, M. (2017). Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*.

Wei, C.-Y., M. J. Jahromi, H. Luo, H. Sharma, and R. Jain (2020). Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pp. 10170–10180. PMLR.

Whitt, W. (1981). Comparing counting processes and queues. *Advances in Applied Probability 13*(1), 207–220.

A   Proof of Lemma 1

Proof. Denote by $v_t(s)$ the ground truth supremum value of the cumulative expected reward collected from time t to T when the state at time t is $s \in \mathcal{S}$. Note that $v_0(s) = V_T(s)$ for all s. The HJB (Hamilton-Jacobi-Bellman) equation satisfied by this finite-horizon optimal value function is

$$\frac{dv_t(s)}{dt} + \max_{a \in \mathcal{A}} \left\{ r(s,a)\lambda(s,a) - \lambda(s,a)v_t(s) + \lambda(s,a) \sum_{z \in \mathcal{S}} p(z|s,a)v_t(z) \right\} = 0, \ (s,t) \in \mathcal{S} \times [0,H),$$
(25)

$$v_T(s) = 0, \quad s \in \mathcal{S},$$

where $\frac{dv_t(s)}{dt}$ denotes the derivative of $v_t(s)$ with respect to time t. On the other hand, it follows from the Bellman optimality equation (2) for the average-reward CTMDP (1) that

$$0 = -\rho^* + \max_{a \in \mathcal{A}} \left\{ r(s,a)\lambda(s,a) + \lambda(s,a) \sum_{z \in \mathcal{S}} p(z|s,a)h^*(z) - \lambda(s,a)h^*(s) \right\}, \quad s \in \mathcal{S}. \tag{26}$$

Since the bias function h^* is defined up to an addtive constant, without loss of generality we assume $\min_{s \in \mathcal{S}} h^*(s) = 0$. Define

$$g_t(s) = (T-t)\rho^* + h^*(s), \ (s,t) \in \mathcal{S} \times [0,T]. \tag{27}$$

It then follows from (26) and (27) that

$$\frac{dg_t(s)}{dt} + \max_{a \in \mathcal{A}} \left\{ r(s,a)\lambda(s,a) - \lambda(s,a)g_t(s) + \lambda(s,a) \sum_{z \in \mathcal{S}} p(z|s,a)g_t(z) \right\} = 0, \ (s,t) \in \mathcal{S} \times [0,H),$$
(28)

$$g_T(s) = h^*(s), \quad s \in \mathcal{S}.$$

The differential equation (28) is identical to (25) except for the terminal conditions where $g_T(s) = h^*(s)$. So $g_t(s)$ is actually the ground truth supremum value of the cumulative expected reward collected from t to T with a non-negative terminal reward $h^*(x_T) \geq 0$ when the state at time t is s. It then follows that

$$v_t(s) \leq g_t(s) = (T-t)\rho^* + h^*(s), \ (s,t) \in \mathcal{S} \times [0,T].$$

In particular, we have

$$V_T(s) = v_0(s) \leq T\rho^* + h^*(s) \leq T\rho^* + \sup_{s \in \mathcal{S}} h^*(s), \quad \text{for all } s \in \mathcal{S}.$$

Noting that $\sup_{s \in \mathcal{S}} h^*(s) < \infty$ because $\mathcal{S}$ is finite, we obtain for all s,

$$V_T(s) - T\rho^* \leq \sup_{s \in \mathcal{S}} h^*(s). \tag{29}$$

By considering $-h^*(s)$ in (27), we can use a similar argument to obtain for all s,

$$V_T(s) - T\rho^* \geq -\sup_{s \in \mathcal{S}} h^*(s). \tag{30}$$

The proof is complete by letting $C = \sup_{s \in \mathcal{S}} h^*(s)$. □

B Proof of Theorem 1

In this section we prove Theorem 1. Given a learning algorithm $\mathcal{G}$, denote by $T_N(s,a)$ the number of occurrences of the state-action pair (s,a) up to the $N-$th decision epoch, i.e. $T_N(s,a) = \sum_{n=0}^{N-1} 1_{s_n=s,a_n=a}$. We need several technical results for the proof of Theorem 1. For the ease of notation, unless specified otherwise, we use $\mathbb{P}_{s_0}$ and $\mathbb{E}_{s_0}$ to denote the probability and expectation under the model $\mathcal{M}$, algorithm $\mathcal{G}$ and initial state s_0 respectively.

Proposition 2. *For any algorithm $\mathcal{G}$, the expected regret up to the $N-$th decision epoch satisfies,*

$$\mathcal{R}_N^{\mathcal{G}}(s_0, \mathcal{M}) = \sum_{x \in \mathcal{S}} \sum_{a \notin O(s, \mathcal{M})} \mathbb{E}_{s_0}[T_N(s,a)] \cdot \phi^*(s,a; \mathcal{M}) + O(1), \quad as\ N \to \infty, \tag{31}$$

where $\phi^*(s,a; \mathcal{M})$ is given in (8).

Proof. The proof is an adaption of the proof of Proposition 1 in Burnetas and Katehakis (1997) to the CTMDPs. Write

$$D_N^{\mathcal{G}}(s_0, \mathcal{M}) = \mathcal{R}_N^{\mathcal{G}}(s_0, \mathcal{M}) + h^*(s_0; \mathcal{M}) = \mathbb{E}_{s_0}[\rho^* \cdot S_N - \sum_{n=0}^{N-1} r(s_n, a_n)] + h^*(s_0; \mathcal{M}), \tag{32}$$

where h is the bias function in the average reward optimality equation (2) and $S_N = \sum_{i=0}^{N-2} \tau_i$ with the convention that $\sum_{i=0}^{-1} \tau_i = 0$. Then we can compute

$$D_N^{\mathcal{G}}(s_0, \mathcal{M}) = \mathbb{E}_{s_0}\left(\mathbb{E}_{s_0}[\rho^* \cdot S_N - \sum_{n=0}^{N-1} r(s_n, a_n)|(a_0, s_1)]\right) + h^*(s_0; \mathcal{M})$$

$$= \mathbb{E}_{s_0}\left(\frac{\rho^*}{\lambda(s_0, a_0)} - r(s_0, a_0) + \mathbb{E}_{s_0}[\rho^* \cdot \sum_{i=1}^{N-2} \tau_i - \sum_{n=1}^{N-1} r(s_n, a_n)|(a_0, s_1)]\right) + h(s_0; \mathcal{M})$$

$$= \mathbb{E}_{s_0}[\phi^*(s_0, a_0; \mathcal{M})] + \mathbb{E}_{s_0}[D_{N-1}^{\mathcal{G}}(s_1, \mathcal{M})],$$

where τ_i is the holding time at a state after the $(i+1)-$th decision epoch, and the last equality follows by adding and subtracting the same quantity $\mathbb{E}_{s_0}(\mathbb{E}_{(a_0,s_1)}[h(s_1; \mathcal{M})])$ from the second last line of the above equation. It then follows from the above recursion that

$$D_N^{\mathcal{G}}(s_0, \mathcal{M}) = \sum_{n=0}^{N-1} \mathbb{E}_{s_0}[\phi^*(s_n, a_n; \mathcal{M})] = \sum_{s \in \mathcal{S}} \sum_{a \notin O(s, \mathcal{M})} \mathbb{E}_{s_0}[T_N(s,a)]\phi^*(s,a; \mathcal{M}),$$

where we use the fact that $\phi^*(s,a; \mathcal{M}) = 0$ for $a \in O(s; \mathcal{M})$ in the last equality. Because the state space is finite, we know $\sup_{s_0 \in \mathcal{S}} |h^*(s_0; \mathcal{M})|$ is a finite constant which is independent of N. Hence we obtain (31) and the proof is complete. □

Proposition 3. *For any UF algorithm $\mathcal{G}$ and critical state-action pair $(s,a) \in B(\mathcal{M})$,*

$$\liminf_{N \to \infty} \mathbb{E}_{s_0}[T_N(s,a)]/\log N \geq \frac{1}{K(s,a; \mathcal{M})},$$

where $K(s,a; \mathcal{M})$ is given in (9).

The proof of Proposition 3 is long and it requires several intermediate lemmas. The essential idea relies on a change-of-measure argument similar to that in Burnetas and Katehakis (1997), although considerable difficulty arises in applying the argument to the continuous-time case due to the inherent random holding time. In the CTMDP setting, the key observation is the following. The likelihood ratio of the two models $\mathcal{M}$ and $\overline{\mathcal{M}}$ given the history ω_N up to N-th decision epoch is given by

$$L(\mathcal{M}, \overline{\mathcal{M}}; \omega_N) = \mathbb{P}_{s_0}^{\mathcal{M}}(s_0, a_0, \tau_0, s_1, \ldots, \tau_{N-1}, s_N) / \mathbb{P}_{s_0}^{\overline{\mathcal{M}}}(s_0, a_0, \tau_0, s_1, \ldots, \tau_{N-1}, s_N). \tag{33}$$

We can compute

$$\mathbb{P}_{s_0}^{\mathcal{M}}(s_0, a_0, \tau_0, s_1, \ldots, \tau_{N-1}, s_N)$$
$$= \prod_{k=0}^{N-1} \mathbb{P}_{s_0}^{\mathcal{M}}(a_k | s_0, a_0, \tau_0, s_1, \ldots, \tau_{k-1}, s_k) \cdot \mathbb{P}_{s_0}^{\mathcal{M}}(\tau_k | s_k, a_k) \mathbb{P}_{s_0}^{\mathcal{M}}(s_{k+1} | s_k, a_k).$$

Since the algorithm $\mathcal{G}$ selects actions only using the history, we have

$$L(\mathcal{M}, \overline{\mathcal{M}}; \omega_N) = \prod_{k=0}^{N-1} \frac{\mathbb{P}_{s_0}^{\mathcal{M}}(\tau_k | s_k, a_k) \cdot \mathbb{P}_{s_0}^{\mathcal{M}}(s_{k+1} | s_k, a_k)}{\mathbb{P}_{s_0}^{\overline{\mathcal{M}}}(\tau_k | s_k, a_k) \cdot \mathbb{P}_{s_0}^{\overline{\mathcal{M}}}(s_{k+1} | s_k, a_k)}. \tag{34}$$

Let the discrete random variable $Z_j(s, a) \in \mathcal{S}$ denote the state visited immediately after the $j-$th occurrence of (s, a), and $t_j(s, a)$ the holding time at state s after the $j-$th occurrence of (s, a). For $k \geq 0$, $q, \bar{q} \in \Theta(s, a)$ and $\lambda, \bar{\lambda} \in [\lambda_{\min}, \lambda_{\max}]$, define

$$\Lambda_k([q, \lambda], [\bar{q}, \bar{\lambda}]) = \prod_{j=1}^{k} \frac{q_{Z_j(s,a)}}{\bar{q}_{Z_j(s,a)}} \cdot \frac{\lambda e^{-\lambda t_j(s,a)}}{\bar{\lambda} e^{-\bar{\lambda} t_j(s,a)}}, \tag{35}$$

representing the likelihood ratio of the two models $\mathcal{M}$ and $\overline{\mathcal{M}}$ corresponding to the history of transitions out of state s under action a. Finally, denote by $\mathbf{P}_{(p(\cdot|s,a),\lambda(s,a))}$ the probability measure generated by $(Z_j(s, a), t_j(s, a))_{j \geq 1}$, with $Z_j(s, a)$ following distribution $p(\cdot|s, a)$ and $t_j(s, a)$ following an exponential distribution with rate $\lambda(s, a)$.

Lemma 2. *We have*

(i) *If $\overline{\mathcal{M}} = \overline{\mathcal{M}}(s, a; \mathcal{M}, \bar{p}(\cdot|s, a), \bar{\lambda}(s, a))$, then*

$$L(\mathcal{M}, \overline{\mathcal{M}}; \omega_N) = \Lambda_{T_N(s,a)}([p(\cdot|s, a), \lambda(s, a)], [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)]), \tag{36}$$

where $T_N(s, a)$ is the number of occurrences of the state action pair (s, a) up to the $N-$th decision epoch.

(ii) *Let (b_n) be an increasing sequence of positive constants with with $b_n \to \infty$ as $n \to \infty$. Then,*

$$\mathbf{P}_{(p(\cdot|s,a),\lambda(s,a))} \left(\max_{k \leq \lfloor b_n \rfloor} \log \Lambda_k / \lfloor b_n \rfloor > (1 + c) \cdot I([p(\cdot|s, a), \lambda(s, a)], [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)]) \right) = o(1),$$

as $n \to \infty$, for any $c > 0$.

Proof. (i) By the construction of $\overline{\mathcal{M}}$, we have $\mathbb{P}^{\mathcal{M}}_{s_0}(s_{k+1}|s_k, a_k) = \mathbb{P}^{\overline{\mathcal{M}}}_{s_0}(s_{k+1}|s_k, a_k)$ unless $(s_k, a_k) = (s, a)$. Similarly, $\mathbb{P}^{\mathcal{M}}_{s_0}(\tau_k|s_k, a_k) = \mathbb{P}^{\overline{\mathcal{M}}}_{s_0}(\tau_k|s_k, a_k)$ unless $(s_k, a_k) = (s, a)$. Hence, all the terms except those corresponding to $(s_k, a_k) = (s, a)$ are cancelled in (34), leading to the desired result (36).

(ii) Since $Z_j(s, a), j = 1, 2, \ldots$ are i.i.d., and so are $t_j(s, a), j = 1, 2, \ldots$, the strong law of large numbers yields that as $k \to \infty$,

$$\log \Lambda_k([p(\cdot|s, a), \lambda(s, a)], [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)])/k$$
$$\to I([p(\cdot|s, a), \lambda(s, a)], [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)]), \quad \text{a.s. } \mathbf{P}_{(p(\cdot|s,a), \lambda(s,a))}.$$

Consequently, $\max_{k \leq \lfloor b_n \rfloor} \log \Lambda_k / \lfloor b_n \rfloor \to I([p(\cdot|s, a), \lambda(s, a)], [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)])$ a.s. for any increasing sequence of (b_n) with $b_n \to \infty$. The result then follows. $\qquad\square$

For a learning algorithm $\mathcal{G}$, denote by $T_N(s)$ the total number of visits of the embedded jump chain (s_n) of the CTMDP to the state s up to the N−th decision epoch, i.e., $T_N(s) = \sum_{n=0}^{N-1} 1_{s_n = s}$. Under Assumption 1, we have the following result.

Lemma 3 (Proposition 2 of Burnetas and Katehakis (1997))**.** *There exists some $\beta > 0$ such that for any algorithm $\mathcal{G}$ and $s \in \mathcal{S}$,*

$$\mathbb{P}_{s_0}(T_N(s) < \beta N) = o(1), \quad \text{as } N \to \infty.$$

Lemma 4. *If $\overline{\mathcal{M}} = \overline{\mathcal{M}}(s, a; \mathcal{M}, \bar{p}(\cdot|s, a), \bar{\lambda}(s, a))$, i.e., action a is uniquely optimal at state s under the model $\overline{\mathcal{M}}$, then for any UF algorithm $\mathcal{G}$,*

$$\mathbb{E}^{\overline{\mathcal{M}}}_{s_0}[T_N(s) - T_N(s, a)] = \mathbb{E}^{\overline{\mathcal{M}}}_{s_0}\Big[\sum_{a' \neq a} T_N(s, a')\Big] = o(N^\alpha), \quad \text{for any } \alpha > 0. \tag{37}$$

Proof. Since $\mathcal{G}$ is UF and $\overline{\mathcal{M}} \in \mathcal{H}$, by definition we have $\mathbb{E}_{s_0}[R^{\mathcal{G}}_T(s_0, \overline{\mathcal{M}})] = o(T^\alpha)$ as $T \to \infty$ for any $\alpha > 0$. Then, for any $\epsilon > 0$, there exists some T_0 such that $\mathbb{E}_{s_0}[R^{\mathcal{G}}_T(s_0, \overline{\mathcal{M}})] \leq \epsilon T^\alpha$ for $T \geq T_0$. It follows that $\mathbb{E}_{s_0}[R^{\mathcal{G}}_T(s_0, \overline{\mathcal{M}})] \leq \epsilon T^\alpha + T_0 \rho^*$ for any $T \geq 0$. By virtue of (7), we get

$$\mathcal{R}^{\mathcal{G}}_N(s_0, \mathcal{M}) := \mathbb{E}_{s_0}[R^{\mathcal{G}}_{S_N}(s_0, \mathcal{M})] = \mathbb{E}_{s_0}\left(\mathbb{E}_{s_0}[R^{\mathcal{G}}_{S_N}(s_0, \mathcal{M})|S_N]\right) \leq \epsilon \cdot \mathbb{E}_{s_0}[S_N^\alpha] + T_0 \rho^*. \tag{38}$$

Denote by (τ_i) the sequence of holding times under any algorithm $\mathcal{G}$ applied to the CTMDP $\overline{\mathcal{M}} \in \mathcal{H}$. Since $S_N = \sum_{i=0}^{N-2} \tau_i$, by Assumption 2, it is possible to construct a probability space and a sequence of i.i.d. exponential random variables $(\hat{\tau}_i)$ with rate $\lambda_{\min}$ such that the law of $(S_N : n \geq 1)$ is the same as the law of $(\tilde{S}_N : N \geq 1)$ and $\tilde{S}_N \leq \hat{S}_N = \sum_{i=0}^{N-2} \hat{\tau}_i$ for all N and all sample paths; see, e.g., Section 2 of Whitt (1981). It follows that $\mathbb{E}_{s_0}[S_N^\alpha] = \mathbb{E}_{s_0}[\tilde{S}_N^\alpha] \leq \mathbb{E}[\hat{S}_N^\alpha]$. Without loss of generality we consider $\alpha \in (0, 1)$. From the strong law of large numbers we have $\hat{S}_N^\alpha / N^\alpha \to (1/\lambda_{\min})^\alpha$ a.s. when $N \to \infty$. Moreover, $\mathbb{E}[\hat{S}_N^\alpha]/N^\alpha \leq \mathbb{E}[\hat{S}_N]/N = 1/\lambda_{\min} \cdot (N-1)/N \leq 1/\lambda_{\min} < \infty$ for each N. By the dominated convergence theorem we then obtain

$$\lim_{N \to \infty} \frac{\mathbb{E}[\hat{S}_N^\alpha]}{N^\alpha} = (1/\lambda_{\min})^\alpha.$$

As a consequence,

$$\limsup_{N\to\infty} \frac{\mathbb{E}_{s_0}[S_N^\alpha]}{N^\alpha} \le (1/\lambda_{\min})^\alpha.$$

It then follows from (38) that

$$\limsup_{N\to\infty} \frac{\mathcal{R}_N^{\mathcal{G}}(s_0, \overline{\mathcal{M}})}{N^\alpha} \le \epsilon \cdot (1/\lambda_{\min})^\alpha.$$

Since $\epsilon > 0$ is arbitrary, we obtain $\mathcal{R}_N^{\mathcal{G}}(s_0, \overline{\mathcal{M}}) = o(N^\alpha)$.

From Proposition 2, we know that for any learning algorithm $\mathcal{G}$, the expected regret up to $N-$th decision epoch satisfies

$$\mathcal{R}_N^{\mathcal{G}}(s_0, \overline{\mathcal{M}}) = \sum_{x\in\mathcal{S}} \sum_{a'\ne a} \mathbb{E}_{s_0}^{\overline{\mathcal{M}}}[T_N(s, a')]\phi^*(x, a'; \overline{\mathcal{M}}) + O(1),$$

where $\{a\} = O(x; \overline{\mathcal{M}})$. Since $\mathcal{R}_N^{\mathcal{G}}(s_0, \overline{\mathcal{M}}) = o(N^\alpha)$, and $\phi^*(x, a'; \overline{\mathcal{M}})$ are all positive and uniformly bounded away from zero for $a' \ne a$, we arrive at

$$\mathbb{E}_{s_0}^{\overline{\mathcal{M}}}[T_N(s) - T_N(s, a)] = \mathbb{E}_{s_0}^{\overline{\mathcal{M}}}[\sum_{a'\ne a} T_N(s, a')] = o(N^\alpha),$$

completing the proof. $\qquad\square$

Now we are ready to prove Proposition 3.

Proof. Given all the previous lemmas, the proof is similar to that of Theorem 2 in Burnetas and Katehakis (1997); so we only outline the key steps here. It suffices to show for any $\epsilon > 0$,

$$\lim_{N\to\infty} \mathbb{P}_{s_0}\left(T_N(s, a) < \frac{(1-\epsilon)\log N}{K(s, a; \mathcal{M})}\right) = 0.$$

By Lemma 3, it further remains to show

$$\mathbb{P}_{s_0}\left(T_N(s, a) < \frac{(1-\epsilon)\log N}{K(s, a; \mathcal{M})}, \quad T_N(s) \ge \beta N\right) = o(1), \quad \text{as } N \to \infty.$$

Let $\mathcal{G}$ be a UF algorithm, $(s, a) \in B(\mathcal{M})$ and $\delta = \epsilon/(2-\epsilon) > 0$. By the definition of $K(s, a; \mathcal{M})$ in (9), one can readily verify that

$$(1-\epsilon)/K(s, a; \mathcal{M}) \le (1-\delta)/I([p(\cdot|s, a), \lambda(s, a)], [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)]).$$

Write $b_N = (1-\delta)\log N / I([p(\cdot|s, a), \lambda(s, a)], [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)])$, ω_N the history up to decision epoch N, and $E_N^\delta = \{\omega_N : T_N(s) \ge \beta N, \quad T_N(s, a) < b_N\}$. Then it suffices to prove

$$\mathbb{P}_{s_0}(E_N^\delta) = o(1), \quad \text{as } N \to \infty. \tag{39}$$

Since $\mathcal{G}$ is UF, and action a is now uniquely optimal under $\overline{\mathcal{M}} = \overline{\mathcal{M}}(s, a; \mathcal{M}, \bar{p}(\cdot|s, a), \bar{\lambda}(s, a))$, it follows from Lemma 4 that

$$\mathbb{E}_{s_0}^{\overline{\mathcal{M}}}[T_N(s) - T_N(s, a)] = \mathbb{E}_{s_0}^{\overline{\mathcal{M}}}[\sum_{a' \neq a} T_N(s, a')] = o(N^\alpha).$$

Thus the Markov inequality implies

$$\mathbb{P}_{s_0}^{\mathcal{G}, \overline{\mathcal{M}}}(E_N^\delta) \leq \mathbb{P}_{s_0}^{\mathcal{G}, \overline{\mathcal{M}}}(T_N(s) - T_N(s, a) \geq \beta N - b_N) = \frac{o(N^{\delta/2})}{\beta N - b_N} = o(N^{\delta/2 - 1}). \qquad (40)$$

Let $B_N^\delta = \{\omega_N : \log \Lambda_{T_N(s,a)}([p(\cdot|s,a), \lambda(s,a)], [\bar{p}(\cdot|s,a), \bar{\lambda}(s,a)]) \leq (1 - \delta/2) \log N\}$. To show (39), we first show $\mathbb{P}_{s_0}(E_N^\delta \cap B_N^\delta) = o(1)$. Using the likelihood ratio (33), equation (36) and the definition of B_N^δ, we have

$$\mathbb{P}_{s_0}(E_N^\delta \cap B_N^\delta) \leq e^{(1-\delta/2)\log N} \cdot \mathbb{P}_{s_0}^{\mathcal{G}, \overline{\mathcal{M}}}(E_N^\delta \cap B_N^\delta) \leq N^{1-\delta/2} \cdot \mathbb{P}_{s_0}^{\mathcal{G}, \overline{\mathcal{M}}}(E_N^\delta) = o(1),$$

where the last equality is due to (40).

We next show $\mathbb{P}_{s_0}(E_N^\delta \cap \bar{B}_N^\delta) = o(1)$. Note that

$$\mathbb{P}_{s_0}(E_N^\delta \cap \bar{B}_N^\delta)$$
$$\leq \mathbf{P}_{(p(\cdot|s,a), \lambda(s,a))} \left(\max_{k \leq \lfloor b_N \rfloor} \log \Lambda_k / \lfloor b_N \rfloor > \frac{1 - \delta/2}{1 - \delta} \cdot I([p(\cdot|s,a), \lambda(s,a)], [\bar{p}(\cdot|s,a), \bar{\lambda}(s,a)]) \right).$$

From part (ii) of Lemma 2 we obtain that $\mathbb{P}_{s_0}(E_N^\delta \cap \bar{B}_N^\delta) = o(1)$. Hence we have proved (39). The proof is therefore complete. $\qquad \square$

Proof of Theorem 1. Equation (11) directly follows from Propositions 2 and 3 along with the definition of $C(\mathcal{M})$ in (10).

We next prove (12). Recall that

$$R_T^{\mathcal{G}}(s_0, \mathcal{M}) = \rho^* \cdot T - \sum_{n=0}^{N(T)-1} r(s_n, a_n),$$

where $N(T)$ is the number of decision epochs by time T under the learning algorithm $\mathcal{G}$. Since the reward is bounded by one, we have

$$|R_T^{\mathcal{G}}(s_0, \mathcal{M}) - R_{S_{N(T)+1}}^{\mathcal{G}}(s_0, \mathcal{M})| \leq \rho^* \cdot [S_{N(T)+1} - S_{N(T)}] + 1,$$

where we recall S_n denotes the $n-$th decision epoch under a given algorithm $\mathcal{G}$. It then follows from Assumption 2 that

$$\mathbb{E}_{s_0}|R_T^{\mathcal{G}}(s_0, \mathcal{M}) - R_{S_{N(T)+1}}^{\mathcal{G}}(s_0, \mathcal{M})| \leq \frac{\rho^*}{\lambda_{\min}} + 1. \qquad (41)$$

Dividing both sides by $\log T$ and sending T to infinity, we have

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{s_0}[R_T^{\mathcal{G}}(s_0, \mathcal{M})]}{\log T} \geq \liminf_{T \to \infty} \frac{\mathbb{E}_{s_0}[R_{S_{N(T)+1}}^{\mathcal{G}}(s_0, \mathcal{M})]}{\log T}.$$

Hence, to prove (12) it suffices to show

$$\liminf_{T \to \infty} \mathbb{E}_{s_0}[R^{\mathcal{G}}_{S_{N(T)+1}}(s_0, \mathcal{M})]/\log T \geq C(\mathcal{M}).$$

From (11), we know that for any $\epsilon > 0$, there exists $N_0 > 0$ such that $\mathcal{R}^{\mathcal{G}}_N(s_0, \mathcal{M})/\log N \geq (C(\mathcal{M}) - \epsilon)$ when $N \geq N_0$. Hence for all $N \geq 1$,

$$\mathcal{R}^{\mathcal{G}}_N(s_0, \mathcal{M}) = \mathbb{E}_{s_0}[R^{\mathcal{G}}_{S_N}(s_0, \mathcal{M})] \geq (C(\mathcal{M}) - \epsilon) \log N - C_0,$$

where $C_0 := \frac{\rho^* N_0}{\lambda_{\min}}$ is a constant that bounds the expected regret in the first N_0 decision steps under any policy π. It follows that

$$\begin{aligned}
\mathbb{E}_{s_0}[R^{\mathcal{G}}_{S_{N(T)+1}}(s_0, \mathcal{M})] &= \mathbb{E}_{s_0}\left(\mathbb{E}_{s_0}[R^{\mathcal{G}}_{S_{N(T)+1}}(s_0, \mathcal{M})|N(T)+1]\right) \\
&\geq (C(\mathcal{M}) - \epsilon)\mathbb{E}_{s_0}[\log(N(T)+1)] - C_0. \quad (42)
\end{aligned}$$

Note that due to Assumption 2, we have under any algorithm $\mathcal{G}$, the counting process $\{N(t) - 1 : t \geq 0\}$ is sandwiched pathwise between two Poisson processes $A_1(\cdot)$ and $A_2(\cdot)$ with rates $\lambda_{\min}$ and $\lambda_{\max}$ respectively, i.e. $A_1(t) \leq N(t) - 1 \leq A_2(t)$ for all $t \geq 0$ and all sample paths, see, e.g., Section 2 of Whitt (1981). Note that we consider $N(t) - 1$ because there is a decision made at time 0 so that $N(0) - 1 = 0$. For a Poisson process $A(\cdot)$ with any rate $\mu > 0$, we can infer from the strong law of large numbers that

$$\lim_{T \to \infty} \mathbb{E}\left[\log(A(T) + 2) - \log(T)\right] = \lim_{T \to \infty} \mathbb{E}\left[\log \frac{A(T) + 2}{T}\right] = \log(\mu),$$

where in the last equality follows from the generalized dominated convergence theorem given that $\log \frac{A(T)+2}{T} \leq \frac{A(T)+2}{T}$, and $\frac{A(T)+2}{T} \to \mu$ almost surely and in expectation as $T \to \infty$. Hence we have $\lim_{T \to \infty} \mathbb{E}_{s_0}[\log(A(T) + 2)]/\log T = 1$. Consequently

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{s_0}[\log(N(T) + 1)]}{\log T} = 1,$$

which, together with (42), yields

$$\liminf_{T \to \infty} \mathbb{E}_{s_0}[R^{\mathcal{G}}_{S_{N(T)+1}}(s_0, \mathcal{M})]/\log T \geq C(\mathcal{M}) - \epsilon.$$

The result then follows as ϵ is arbitrarily small. $\qquad \square$

C   Proof of Proposition 1

Proof. First, one can easily check that an equivalent formulation of $C(\mathcal{M})$ is as follows:

$$C(\mathcal{M}) = \inf_{\eta \geq 0} \sum_{(s,a) \in B(\mathcal{M})} \eta(s, a)\phi^*(s, a; \mathcal{M}) \quad (43)$$

$$\text{subject to} \quad \eta(s, a)[KL(p(\cdot|s, a), \bar{p}(\cdot|s, a)) + KL(\lambda(s, a), \bar{\lambda}(s, a))] \geq 1,$$

$$\text{for } [\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)] \in \Delta\Theta(s, a; \mathcal{M}) \text{ and all } (s, a) \in B(\mathcal{M}).$$

23

Indeed the optimal solution to the optimization problem (43) is given by $\eta^*(s, a) = 1/K(s, a; \mathcal{M})$ for $(s, a) \in B(\mathcal{M})$ by (9). To upper bound the objective value $C(\mathcal{M})$ in (43), we construct a smaller feasible set for the above optimization problem. For this purpose, we need a characterization of the set $\Delta\Theta(s, a; \mathcal{M})$. Recall the average reward optimality equation is given by

$$h^*(s; \mathcal{M}) = \max_{a \in \mathcal{A}} \left\{ r(s, a) - \rho^*(\mathcal{M})/\lambda(s, a) + \sum_{j \in \mathcal{S}} p(j|s, a)h^*(j; \mathcal{M}) \right\}, \quad s \in \mathcal{S}.$$

For $q \in \Theta(s, a), \theta \in [\lambda_{\min}, \lambda_{\max}]$ and $\rho \in [0, 1]$, we denote

$$\mathcal{L}(s, a; q, \theta, h, \rho) = r(s, a) + q \cdot h - \rho/\theta,$$

where $q \cdot h := \sum_{j \in \mathcal{S}} q(j)h(j)$. Then the optimality equation becomes

$$h^*(s; \mathcal{M}) = \max_{a \in \mathcal{A}} \left\{ \mathcal{L}(s, a; p(\cdot|s, a), \lambda(s, a), h^*(\mathcal{M}), \rho^*(\mathcal{M})) \right\}, \quad s \in \mathcal{S}.$$

Following a proof similar to that of Lemma 1 of Burnetas and Katehakis (1997), we can show

$$\Delta\Theta(s, a; \mathcal{M}) = \{(q, \theta) \in \Theta(s, a) \times [\lambda_{\min}, \lambda_{\max}] : \mathcal{L}(s, a; q, \theta, h^*(\mathcal{M}), \rho^*(\mathcal{M})) > h^*(s; \mathcal{M})\} \quad (44)$$

for all (s, a) such that $a \notin O(s; \mathcal{M})$. Recall

$$\phi^*(s, a; \mathcal{M}) = -r(s, a) + \rho^*(\mathcal{M})/\lambda(s, a) - \sum_{j \in \mathcal{S}} p(j|s, a)h^*(j; \mathcal{M}) + h^*(s; \mathcal{M})$$
$$= h^*(s; \mathcal{M}) - \mathcal{L}(s, a; p(\cdot|s, a), \lambda(s, a), h^*(\mathcal{M}), \rho^*(\mathcal{M}),$$

and

$$\mathcal{L}(s, a; q, \theta, h^*(\mathcal{M}), \rho^*(\mathcal{M})) - \mathcal{L}(s, a; p(\cdot|s, a), \lambda(s, a), h^*(\mathcal{M}), \rho^*(\mathcal{M})$$
$$= (q - p(\cdot|s, a)) \cdot h^*(\mathcal{M}) - \rho^*(\mathcal{M})(1/\theta - 1/\lambda(s, a)).$$

Hence it follows from (44) that

$$\Delta\Theta(s, a; \mathcal{M})$$
$$= \{(q, \theta) \in \Theta(s, a) \times [\lambda_{\min}, \lambda_{\max}] : \mathcal{L}(s, a; q, \theta, h^*(\mathcal{M}), \rho^*(\mathcal{M})) > h^*(s; \mathcal{M})\}$$
$$= \{(q, \theta) \in \Theta(s, a) \times [\lambda_{\min}, \lambda_{\max}] :$$
$$(q - p(\cdot|s, a)) \cdot h^*(\mathcal{M}) - \rho^*(\mathcal{M})(1/\theta - 1/\lambda(s, a)) > \phi^*(s, a; \mathcal{M})\}.$$

Thus, for any $(\bar{p}(\cdot|s, a), \bar{\lambda}(s, a)) \in \Delta\Theta(s, a; \mathcal{M}) \neq \emptyset$, we have

$$(\bar{p}(\cdot|s, a) - p(\cdot|s, a)) \cdot h^*(\mathcal{M}) - \rho^*(\mathcal{M})(1/\bar{\lambda}(s, a) - 1/\lambda(s, a)) > \phi^*(s, a; \mathcal{M}).$$

Applying Hölder's inequality and noting $\rho^*(\mathcal{M}) \leq \lambda_{\max}$, we have

$$||\bar{p}(\cdot|s, a) - p(\cdot|s, a)||_1 \cdot ||h^*(\mathcal{M}))||_\infty + |1/\bar{\lambda}(s, a) - 1/\lambda(s, a)| \cdot \lambda_{\max} > \phi^*(s, a; \mathcal{M}).$$

24

Note the bias function in the average reward optimality equation is defined up to an additive constant; so we can assume without loss of generality that $||h^*(\mathcal{M})||_\infty \leq H/2$. Hence,

$$||\bar{p}(\cdot|s,a) - p(\cdot|s,a)||_1 \cdot H/2 + |1/\bar{\lambda}(s,a) - 1/\lambda(s,a)| \cdot \lambda_{\max} > \phi^*(s,a;\mathcal{M}),$$

which implies that

$$||\bar{p}(\cdot|s,a) - p(\cdot|s,a)||_1 + |1/\bar{\lambda}(s,a) - 1/\lambda(s,a)| > \frac{2\phi^*(s,a;\mathcal{M})}{H + 2\lambda_{\max}}. \tag{45}$$

Note that $KL(p(\cdot|s,a), \bar{p}(\cdot|s,a)) \geq \frac{||p(\cdot|s,a) - \bar{p}(\cdot|s,a)||_1^2}{2}$ by Pinsker's inequality. In addition, we have

$$KL(\lambda(s,a), \bar{\lambda}(s,a)) \geq \frac{\lambda_{\min}^3}{2\lambda_{\max}} \cdot |1/\bar{\lambda}(s,a) - 1/\lambda(s,a)|^2, \quad \text{for } \lambda(s,a), \bar{\lambda}(s,a) \in [\lambda_{\min}, \lambda_{\max}],$$

which follows from the fact that $\log(\alpha/\beta) + \beta/\alpha - 1 \geq \frac{\lambda_{\min}}{2\lambda_{\max}}(1 - \beta/\alpha)^2 \geq \frac{\lambda_{\min}^3}{2\lambda_{\max}}(1/\beta - 1/\alpha)^2$ for $\alpha, \beta \in [\lambda_{\min}, \lambda_{\max}]$. As a result,

$$\begin{aligned}
&KL(p(\cdot|s,a), \bar{p}(\cdot|s,a)) + KL(\lambda(s,a), \bar{\lambda}(s,a)) \\
&\geq \frac{||p(\cdot|s,a) - \bar{p}(\cdot|s,a)||_1^2}{2} + \frac{\lambda_{\min}^3}{2\lambda_{\max}} \cdot |1/\bar{\lambda}(s,a) - 1/\lambda(s,a)|^2 \\
&\geq \frac{\lambda_{\min}^3}{2\lambda_{\max}} \cdot \left(||p(\cdot|s,a) - \bar{p}(\cdot|s,a)||_1^2 + |1/\bar{\lambda}(s,a) - 1/\lambda(s,a)|^2 \right) \\
&\geq \frac{\lambda_{\min}^3}{4\lambda_{\max}} \cdot \left(||\bar{p}(\cdot|s,a) - p(\cdot|s,a)||_1 + |1/\bar{\lambda}(s,a) - 1/\lambda(s,a)| \right)^2 \\
&\geq \frac{\lambda_{\min}^3}{4\lambda_{\max}} \cdot \left(\frac{2\phi^*(s,a;\mathcal{M})}{H + 2\lambda_{\max}} \right)^2,
\end{aligned} \tag{46}$$

where the last inequality holds due to (45). Introduce the following optimization problem

$$G(\mathcal{M}) := \inf_{\eta \geq 0} \sum_{(s,a) \in B(\mathcal{M})} \eta(s,a)\phi^*(s,a;\mathcal{M}) \tag{47}$$

$$\text{subject to} \quad \eta(s,a) \cdot \frac{\lambda_{\min}^3}{4\lambda_{\max}} \cdot \left(\frac{2\phi^*(s,a;\mathcal{M})}{H + 2\lambda_{\max}} \right)^2 \geq 1, \quad \text{for all } (s,a) \in B(\mathcal{M}).$$

By (46), the feasible set of this optimization problem is a subset of that of the problem (43). It follows that $C(\mathcal{M}) \leq G(\mathcal{M})$. Moerover, by taking $\eta(s,a) = \frac{4\lambda_{\max}}{\lambda_{\min}^3} \cdot \left(\frac{H + 2\lambda_{\max}}{2\phi^*(s,a;\mathcal{M})} \right)^2$, one finds that

$$\begin{aligned}
G(\mathcal{M}) &= \sum_{(s,a) \in B(\mathcal{M})} \frac{4\lambda_{\max}}{\lambda_{\min}^3} \cdot \left(\frac{H + 2\lambda_{\max}}{2\phi^*(s,a;\mathcal{M})} \right)^2 \cdot \phi^*(s,a;\mathcal{M}) \\
&= \sum_{(s,a) \in B(\mathcal{M})} \frac{\lambda_{\max}}{\lambda_{\min}^3} \cdot \frac{(H + 2\lambda_{\max})^2}{\phi^*(s,a;\mathcal{M})} \\
&\leq \frac{(H + 2\lambda_{\max})^2 SA}{\min_{(s,a) \in B(\mathcal{M})} \phi^*(s,a;\mathcal{M})} \cdot \frac{\lambda_{\max}}{\lambda_{\min}^3}.
\end{aligned}$$

The proof is therefore complete. $\qquad\square$

D Proof of Theorem 2

The proof is lengthy; we divide it into several steps. In Section D.1 we study failing confidence regions. This result is then used in Section D.2 to bound the number of suboptimal decision steps of the CT-UCRL algorithm. With such a bound, we then prove Theorem 2 in Section D.3.

D.1 Failing confidence regions

Lemma 5 (Failing confidence region). *For any episode $k \geq 1$, the probability that the true CTMDP $\mathcal{M}$ is not contained in the set of plausible MDPs $\mathcal{C}_k$ is at most $\frac{\delta}{15t_k^6}$, i.e.*

$$\mathbb{P}(\mathcal{M} \notin \mathcal{C}_k) \leq \frac{\delta}{15t_k^6}. \tag{48}$$

Proof. First, from the proof of Lemma 17 in Jaksch et al. (2010), we have

$$\mathbb{P}\left(\|p(\cdot|s,a) - \hat{p}(\cdot|s,a)\|_1 > \sqrt{\frac{14S\log(2At_k/\delta)}{\max\{1, N_k(s,a)\}}} \right) \leq \frac{\delta}{20t_k^6 SA}. \tag{49}$$

Second, it follows from (15) that given n i.i.d samples of holding times at (s,a), the truncated mean estimator $\frac{1}{\hat{\lambda}(s,a)}$ satisfies

$$\mathbb{P}\left(\left| \frac{1}{\hat{\lambda}(s,a)} - \frac{1}{\lambda(s,a)} \right| \geq \frac{4}{\lambda_{\min}} \cdot \sqrt{\frac{14}{n}\log(2ASt/\delta)} \right) \leq \frac{\delta}{60t^7 SA}, \quad \forall t \geq 1.$$

Since $N_k(s,a) \leq t_k - 1$, we can use a union bound over all possible values of $N_k(s,a) = 1, 2, \ldots, t_k - 1$ and obtain

$$\mathbb{P}\left(\left| \frac{1}{\hat{\lambda}(s,a)} - \frac{1}{\lambda(s,a)} \right| \geq \frac{4}{\lambda_{\min}}\sqrt{\frac{14\log(2ASt_k/\delta)}{\max\{1, N_k(s,a)\}}} \right) \leq \sum_{n=1}^{t_k-1} \frac{\delta}{60t_k^7 SA} \leq \frac{\delta}{60t_k^6 SA}. \tag{50}$$

Finally, in view of (49) and (50), we can sum over all state-action pairs and obtain (48). □

D.2 Bounding the number of suboptimal decision steps of CT-UCRL

In this section, we establish a bound on the number of decision steps in suboptimal episodes for the CT-UCRL algorithm. This bound is critical in the proof of Theorem 2.

Fix the total number of decision steps N. Denote

$$\Delta_k = \sum_{(s,a)} v_k(s,a)(\rho^*/\lambda(s,a) - r(s,a)), \tag{51}$$

where $v_k(s,a)$ denotes the number of visits of the CTMDP to the state-action pair (s,a) in episode k, up to step N. We call Δ_k the *adjusted regret* in episode k, and we will see later that conditional

on $(v_k(s,a))_{(s,a)}$, the expected value of Δ_k corresponds to the expected regret incurred in episode k. We say that an episode k is ϵ-bad if $\frac{\Delta_k}{\ell_k} \geq \epsilon$, where l_k is the number of decision steps in episode k.

We are to provide a bound on the number of decision steps in ϵ-bad episodes for the CT-UCRL algorithm. Theorem 11 in Jaksch et al. (2010) establishes such a bound for the discrete-time MDP. There are essential difficulties in extending its proof to the continuous-time setting due to exponential holding times and a different Bellman optimality equation (2). To overcome these difficulties, we need two new ingredients in our analysis. The first one is the high probability confidence bound for the rate of the holding times; see (50). The second one is the analysis of the extended value iteration (19) for CTMDPs, which has been discussed in Fruit and Lazaric (2017) for SMDPs with different confidence regions.

The following is the main result in this subsection.

Proposition 4. *Let $L_\epsilon(N)$ be the number of decision steps taken by CT-UCRL in $\epsilon-$bad episodes up to step N. Then for any initial state s, $N \geq 2SA, \epsilon > 0$ and $\delta \in (0, 1/2)$, with probability at least $1 - 2\delta$,*

$$L_\epsilon(N) \leq \frac{3}{\epsilon^2} \left(34^2 \lambda_{\max}^2 D(\mathcal{M})^2 S^2 A + 2 \cdot 73^2 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} SA + \frac{24SA}{\lambda_{\min}^2} \right) \cdot \log \left(\frac{N}{\delta} \right),$$

where $D(\mathcal{M})$ is the diameter of the true CTMDP $\mathcal{M}$.

Proof. Fix $N > 1$. Denote by K_ϵ the random set that contains the indices of the ϵ-bad episodes up to step N. Set $\Delta'(\epsilon, N) = \sum_{k \in K_\epsilon} \Delta_k$, and note $L_\epsilon(N) := \sum_{k \in K_\epsilon} \sum_{(s,a)} v_k(s,a)$.

Following the same argument as on p.1580 of Jaksch et al. (2010), we can infer from Lemma 5 that

$$\mathbb{P} \left(\sum_{k \in K_\epsilon} \Delta_k 1_{\mathcal{M} \notin \mathcal{C}_k} > 0 \right) \leq \delta,$$

where the confidence set $\mathcal{C}_k$ is the set of all plausible CTMDPs in episode k. Hence, with probability at least $1 - \delta$,

$$\Delta'(\epsilon, N) \leq \sum_{k \in K_\epsilon} \Delta_k 1_{\mathcal{M} \in \mathcal{C}_k}. \tag{52}$$

Next we bound $\Delta_k 1_{\mathcal{M} \in \mathcal{C}_k}$. By (18) and the assumption that $\mathcal{M} \in \mathcal{C}_k$, we have $\rho^* \leq \rho_k^* + \frac{1}{\sqrt{t_k}}$,

where ρ^* is the optimal average reward of $\mathcal{M}$. Then

$$\Delta_k = \sum_{(s,a)} v_k(s,a)(\rho^*/\lambda(s,a) - r(s,a))$$

$$\leq \sum_{(s,a)} v_k(s,a)(\rho_k^*/\lambda(s,a) - r(s,a)) + \sum_{(s,a)} \frac{v_k(s,a)}{\lambda(s,a)\sqrt{t_k}}$$

$$\leq \sum_{(s,a)} v_k(s,a)(\rho_k^*/\widetilde{\lambda}_k(s,a) - r(s,a)) + \sum_{(s,a)} v_k(s,a)(\rho_k^*/\lambda(s,a) - \rho_k^*/\widetilde{\lambda}_k(s,a))$$

$$+ \sum_{(s,a)} \frac{v_k(s,a)}{\lambda_{\min}\sqrt{t_k}}, \tag{53}$$

where the last inequality follows from the fact that $\lambda(s,a) \geq \lambda_{\min}$ for all s,a. Note in episode k, we execute the optimistic policy $\tilde{\pi}_k$, i.e., $a = \tilde{\pi}_k(s)$. Hence $v_k(s,a) = 0$ if $a \neq \tilde{\pi}_k(s)$.

For the first term of the right hand side of (53), by the extended value iteration (19), we can directly obtain from equation (22) of Fruit and Lazaric (2017) that for all $s \in \mathcal{S}$,

$$\left| \rho_k^* - r(s, \tilde{\pi}_k(s)) \cdot \widetilde{\lambda}_k(s, \tilde{\pi}_k(s)) - \left(\sum_{j \in \mathcal{S}} \widetilde{p}_k(j|s, \tilde{\pi}_k(s)) u_i(j) - u_i(s) \right) \cdot \frac{\widetilde{\lambda}_k(s, \tilde{\pi}_k(s))}{\lambda_{\max}} \right| \leq \frac{1}{\sqrt{t_k}},$$

where the optimistic CTMDP in episode k is $\widetilde{M}_k = (\widetilde{p}_k(\cdot|s,a), \widetilde{\lambda}_k(s,a))_{s,j \in \mathcal{S}, a \in \mathcal{A}}$, and $u_i(s)$ are the state values of the extended value iteration when it stops at iteration i under condition (20). Because $\widetilde{\lambda}_k(s, \tilde{\pi}_k(s)) \geq \lambda_{\min}$, the first term in (53) is bounded above by

$$\frac{1}{\lambda_{\min}} \sum_{s \in \mathcal{S}} v_k(s, \tilde{\pi}_k(s)) \frac{1}{\sqrt{t_k}} + \frac{1}{\lambda_{\max}} \sum_{s \in \mathcal{S}} v_k(s, \tilde{\pi}_k(s)) \left(\sum_{j \in \mathcal{S}} \widetilde{p}_k(j|s, \tilde{\pi}_k(s)) u_i(j) - u_i(s) \right)$$

$$= \frac{1}{\lambda_{\min}} \sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{t_k}} + \frac{1}{\lambda_{\max}} \mathbf{v}_k(\widetilde{P}_k - I) w_k, \tag{54}$$

where $\mathbf{v}_k := (v_k(s, \tilde{\pi}_k(s)))_s$, $\widetilde{P}_k := (\widetilde{p}_k(j|s, \tilde{\pi}_k(s)))_{s,j}$ is the transition matrix of $\tilde{\pi}_k$ on $\widetilde{M}_k$, and $w_k = (w_k(s))_s$ is a column vector defined by

$$w_k(s) = u_i(s) - \frac{\max_{s \in \mathcal{S}} u_i(s) + \min_{s \in \mathcal{S}} u_i(s)}{2}, \quad s \in \mathcal{S}.$$

Here, the equality (54) holds because the matrix $\widetilde{P}_k - I$ right-multiplied by a constant vector is zero. By Lemma 6 of Fruit and Lazaric (2017), one can bound $\|w_k\|_\infty$ by $D = \frac{D(\mathcal{M})\lambda_{\max}^2}{2}$ for all k. Note that

$$\mathbf{v}_k(\widetilde{P}_k - I) w_k = \mathbf{v}_k(\widetilde{P}_k - P_k) w_k + \mathbf{v}_k(P_k - I) w_k, \tag{55}$$

where P_k is the transition matrix associated with the policy π_k under the true model. By Hölder's

inequality, we deduce

$$\mathbf{v}_k(\widetilde{P}_k - P_k)w_k \le \sum_s v_k(s, \tilde{\pi}_k(s))||\widetilde{P}_k(\cdot|s, \pi_k(s)) - P_k(\cdot|s, \tilde{\pi}_k(s))||_1 \cdot ||w_k||_\infty$$

$$\le \sum_s v_k(s, \tilde{\pi}_k(s))\sqrt{\frac{14S \log(2At_k/\delta)}{\max\{1, N_k(s, \tilde{\pi}_k(s))\}}} \cdot D$$

$$\le D\sqrt{14S \log(2AN/\delta)} \sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}}, \tag{56}$$

where the second inequality follows from the fact that the transition matrices of both the optimistic model and the true model lie in the confidence region (16), while the third inequality uses the fact that $t_k \le N$.

For the second term of the right hand side of (53), we first note that the reward function $r(s,a)$ is bounded by 1 for all (s,a); so the long-run average reward per unit time $\rho_k^* \le \lambda_{\max}$. Hence

$$\rho_k^*/\lambda(s,a) - \rho_k^*/\widetilde{\lambda}_k(s,a) \le \lambda_{\max}|1/\lambda(s,a) - 1/\widetilde{\lambda}_k(s,a)|$$

$$\le \lambda_{\max}\left(|1/\lambda(s,a) - 1/\hat{\lambda}_k(s,a)| + |1/\hat{\lambda}_k(s,a) - 1/\widetilde{\lambda}_k(s,a)|\right)$$

$$\le \frac{8\lambda_{\max}}{\lambda_{\min}} \cdot \sqrt{\frac{14\log(2ASt_k/\delta)}{\max\{1, N_k(s,a)\}}},$$

where we use the confidence region (17) for the rates of holding times and the fact that $\mathcal{M}, \widetilde{M}_k \in \mathcal{C}_k$. Thus the second term of (53) is upper bounded by

$$\sum_{(s,a)} v_k(s,a)(\rho_k^*/\lambda(s,a) - \rho_k^*/\widetilde{\lambda}_k(s,a))$$

$$\le \frac{8\lambda_{\max}}{\lambda_{\min}} \cdot \sum_{s,a} v_k(s,a)\sqrt{\frac{14\log(2ASt_k/\delta)}{\max\{1, N_k(s,a)\}}}$$

$$\le \frac{8\lambda_{\max}}{\lambda_{\min}} \cdot \sqrt{14\log(2ASN/\delta)} \cdot \sum_{s,a} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}}. \tag{57}$$

Noting $t_k \ge \max\{1, N_k(s,a)\}$ for all s, a, and combining (53), (54), (55), (56), and (57), we obtain for episode k with $\mathcal{M} \in \mathcal{C}_k$,

$$\Delta_k \le \frac{D}{\lambda_{\max}}\sqrt{14S \log(2AN/\delta)} \cdot \sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} + \frac{1}{\lambda_{\max}}\mathbf{v}_k(P_k - I)w_k$$

$$+ \frac{8\lambda_{\max}}{\lambda_{\min}} \cdot \sqrt{14\log(2ASN/\delta)} \cdot \sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} + \frac{2}{\lambda_{\min}} \sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}}.$$

From equation (27) of Jaksch et al. (2010), we have

$$\sum_{k \in K_\epsilon} \sum_{(s,a)} \frac{v_k(s,a)}{\max\{1, N_k(s,a)\}} \le (1 + \sqrt{2})\sqrt{L_\epsilon(N)SA}.$$

It follows from (52) that with probability at least $1 - \delta$,

$$
\begin{aligned}
\Delta'&(\epsilon, N) \\
&\leq \sum_{k \in K_\epsilon} \Delta_k 1_{\mathcal{M} \in \mathcal{C}_k} \\
&\leq \frac{1}{\lambda_{\max}} \sum_{k \in K_\epsilon} \mathbf{v}_k (P_k - I) w_k 1_{\mathcal{M} \in \mathcal{C}_k} \\
&\quad + \left(\frac{D}{\lambda_{\max}} \sqrt{14 S \log(2AN/\delta)} + \frac{8\lambda_{\max}}{\lambda_{\min}} \cdot \sqrt{14 \log(2ASN/\delta)} \right) \cdot (1 + \sqrt{2})\sqrt{L_\epsilon(N)SA} \\
&\quad + \frac{2}{\lambda_{\min}} \cdot (1 + \sqrt{2})\sqrt{L_\epsilon(N)SA}.
\end{aligned}
\tag{58}
$$

For the first term above, we can use the same argument as in Jaksch et al. (2010) (see p. 1580 - 1581 there) to obtain that, with probability at least $1 - \delta$,

$$
\sum_{k \in K_\epsilon} \mathbf{v}_k (P_k - I) w_k 1_{M \in \mathcal{C}_k} \leq 2D\sqrt{L_\epsilon(N) \log(N/\delta)} + 2DSA \log_2(8N/SA).
$$

Hence (58) yields that with probability at least $1 - 2\delta$,

$$
\begin{aligned}
\Delta'&(\epsilon, N) \\
&\leq 2D(\mathcal{M})\lambda_{\max}\sqrt{L_\epsilon(N)\log(N/\delta)} + 2D(\mathcal{M})\lambda_{\max}SA \log_2(8N/SA) \\
&\quad + \left(\frac{D(\mathcal{M})\lambda_{\max}}{2} \sqrt{14 S \log(2AN/\delta)} + \frac{8\lambda_{\max}}{\lambda_{\min}} \cdot \sqrt{14 \log(2ASN/\delta)} \right) \cdot (1 + \sqrt{2})\sqrt{L_\epsilon(N)SA} \\
&\quad + \frac{2}{\lambda_{\min}} \cdot (1 + \sqrt{2})\sqrt{L_\epsilon(N)SA},
\end{aligned}
$$

noting $D = \frac{D(\mathcal{M})\lambda_{\max}^2}{2}$. In view of equation (32) of Jaksch et al. (2010), this can be simplified to

$$
\begin{aligned}
\Delta'&(\epsilon, N) \\
&\leq 34 D(\mathcal{M})\lambda_{\max}S\sqrt{L_\epsilon(N)A\log(N/\delta)} + \frac{8\lambda_{\max}}{\lambda_{\min}} \cdot \sqrt{14 \log(2ASN/\delta)}(1 + \sqrt{2})\sqrt{L_\epsilon(N)SA} \\
&\quad + \frac{2}{\lambda_{\min}} \cdot (1 + \sqrt{2})\sqrt{L_\epsilon(N)SA}.
\end{aligned}
\tag{59}
$$

Noting by definition $\Delta'(\epsilon, N) = \sum_{k \in K_\epsilon} \Delta_k \geq \epsilon \sum_{k \in K_\epsilon} \ell_k = \epsilon L_\epsilon(N)$, and $8(1 + \sqrt{2})\sqrt{14} \leq 73$, we get

$$
L_\epsilon(N) \leq \frac{3}{\epsilon^2} \left(34^2 \lambda_{\max}^2 D(\mathcal{M})^2 S^2 A \log\left(\frac{N}{\delta}\right) + 73^2 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} SA \log\left(\frac{2ASN}{\delta}\right) + 24SA \frac{1}{\lambda_{\min}^2} \right).
$$

For $N \geq 2SA$, $\log\left(\frac{2ASN}{\delta}\right) \leq 2\log\left(\frac{N}{\delta}\right)$. In addition, $\log\left(\frac{N}{\delta}\right) \geq 1$ for $N \geq 2$ and $\delta < 1/2$. Consequently,

$$
L_\epsilon(N) \leq \frac{3}{\epsilon^2} \left(34^2 \lambda_{\max}^2 D(\mathcal{M})^2 S^2 A + 2 \cdot 73^2 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} SA + \frac{24SA}{\lambda_{\min}^2} \right) \cdot \log\left(\frac{N}{\delta}\right).
$$

The proof is complete. $\qquad\square$

D.3 Proof of Theorem 2

Proof. Given all the preparations, the proof is an adaption of that of Theorem 4 in Jaksch et al. (2010); so we only outline the key steps while highlighting the differences.

We first prove (22). From Proposition 4, it follows that with probability at least $1 - 2\delta$, $L_\epsilon(N) \leq \frac{C \log(N/\delta)}{\epsilon^2}$ for $N \geq 2SA$, where

$$C = 3 \left(34^2 \lambda_{\max}^2 D(\mathcal{M})^2 S^2 A + 2 \cdot 73^2 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} SA + \frac{24SA}{\lambda_{\min}^2} \right).$$

Then we can infer from (59) that the adjusted regret accumulated in ϵ-bad episodes is bounded by

$$\Delta'(\epsilon, N) = \sum_{k \in K_\epsilon} \Delta_k \leq C \cdot \frac{\log(N/\delta)}{\epsilon}$$

with probability at least $1 - 2\delta$. By Assumption 2 and the fact that $r(s,a) \in [0,1]$, we have the following simple bound:

$$\rho^*/\lambda(s,a) - r(s,a) \leq \rho^*/\lambda_{\min} \leq \lambda_{\max}/\lambda_{\min}. \tag{60}$$

Hence, by choosing $\delta = 1/N$, we can bound the expected adjusted regret in $\frac{g}{2}$-bad episodes of the algorithm $\bar{\mathcal{G}} := \text{CT-UCRL}$ up to decision step N as follows:

$$\mathbb{E}_{s_0}\left[\Delta'(\frac{g}{2}, N) \right] = \mathbb{E}_{s_0}\left[\sum_{k \in K_{\frac{g}{2}}} \Delta_k \right] \leq 4C \cdot \frac{\log(N)}{g} + \frac{2\lambda_{\max}}{\lambda_{\min}}. \tag{61}$$

It remains to bound the regret in those episodes k with average adjusted regret smaller than $\frac{g}{2}$, i.e. $k \notin K_{\frac{g}{2}}$. To this end, note that for each policy π there is n_π such that for all $n \geq n_\pi$ the expected average reward after n decision steps is $\frac{g}{2}$-close to the average reward of π. Then applying the same argument as in the proof of Theorem 4 in Jaksch et al. (2010), we have

$$\mathbb{E}_{s_0}\left[\sum_{k \notin K_{\frac{g}{2}}} \Delta_k \right] \leq \frac{\lambda_{\max}}{\lambda_{\min}} \cdot C_1, \quad \text{where } C_1 := \sum_{s,a}[1 + \log_2(\max_{\pi:\pi(s)=a} n_\pi)] \cdot \max_{\pi:\pi(s)=a} n_\pi,$$

where the factor $\frac{\lambda_{\max}}{\lambda_{\min}}$ arises due to the bound in (60) in our setting. Combining with (61), we get the expected regret of UCRL to be

$$\mathbb{E}_{s_0}\left[\sum_{k=1}^m \Delta_k \right] \leq 4C \cdot \frac{\log(N)}{g} + \frac{\lambda_{\max}}{\lambda_{\min}}(2 + C_1), \tag{62}$$

where Δ_k is defined in (51), and m is the number of episodes started up to step N.

Note, however, that the expected regret of the CT-UCRL algorithm $\bar{\mathcal{G}}$ up to step N is given by

$$\mathcal{R}_N^{\bar{\mathcal{G}}}(s_0, \mathcal{M}) = \mathbb{E}_{s_0}[\rho^* \cdot S_N - \sum_{n=0}^{N-1} r(s_n, a_n)] \leq \mathbb{E}_{s_0}\left[\sum_{k=1}^m \hat{\Delta}_k\right],$$

where

$$\hat{\Delta}_k = \sum_{(s,a)} \sum_{j=1}^{v_k(s,a)} (\rho^* \tau_{k_j}(s, a) - r(s, a)),$$

and $(\tau_{k_j}(s, a))_j$ are i.i.d. exponential random variables with rate $\lambda(s, a)$. The difference compared with (62) is $\mathbb{E}_{s_0}[\sum_k(\hat{\Delta}_k - \Delta_k)]$, where Δ_k is defined in (51). Since the mean of $\tau_{k_j}(s, a)$ is $1/\lambda(s, a)$, it follows that given $v_k(s, a)$ after N decision steps, we have

$$\mathbb{E}_{s_0}\left[(\hat{\Delta}_k - \Delta_k)|(v_k(s, a))_{s,a}\right] = 0,$$

which implies that $\mathbb{E}_{s_0}^{\bar{\mathcal{G}}}[\hat{\Delta}_k - \Delta_k] = 0$ for each episode k. Then for $N \geq 2SA$,

$$\mathcal{R}_N^{\bar{\mathcal{G}}}(s_0, \mathcal{M}) \leq \mathbb{E}_{s_0}\left[\sum_{k=1}^m \Delta_k\right] \leq 4C \cdot \frac{\log(N)}{g} + \frac{\lambda_{\max}}{\lambda_{\min}}(2 + C_1).$$

Hence, for any $N \geq 1$,

$$\mathcal{R}_N^{\bar{\mathcal{G}}}(s_0, \mathcal{M}) \leq 4C \cdot \frac{\log(N)}{g} + \frac{\lambda_{\max}}{\lambda_{\min}}(2 + C_1 + 2SA).$$

The proof of (22) is complete by setting

$$C' := \frac{\lambda_{\max}}{\lambda_{\min}}(2 + C_1 + 2SA) = \frac{\lambda_{\max}}{\lambda_{\min}}\left(2 + 2SA + \sum_{s,a}[1 + \log_2(\max_{\pi:\pi(s)=a} n_\pi)] \cdot \max_{\pi:\pi(s)=a} n_\pi\right). \tag{63}$$

We now proceed to prove (24). First, we obtain from (41) that

$$\mathbb{E}_{s_0}[R_T^{\bar{\mathcal{G}}}(s_0, \mathcal{M})] \leq \mathbb{E}_{s_0}[R_{S_{N(T)+1}}^{\bar{\mathcal{G}}}(s_0, \mathcal{M})] + \frac{\rho^*}{\lambda_{\min}} + 1. \tag{64}$$

Hence it suffices to upper bound $\mathbb{E}_{s_0}[R_{S_{N(T)+1}}^{\bar{\mathcal{G}}}(s_0, \mathcal{M})]$. By (22), we have for $N \geq 1$,

$$\mathcal{R}_N^{\bar{\mathcal{G}}}(s_0, \mathcal{M}) = \mathbb{E}_{s_0}[R_{S_N}^{\bar{\mathcal{G}}}(s_0, \mathcal{M})] \leq 4C \cdot \frac{\log(N)}{g} + C'.$$

It follows that

$$\mathbb{E}_{s_0}[R_{S_{N(T)+1}}^{\bar{\mathcal{G}}}(s_0, \mathcal{M})] = \mathbb{E}_{s_0}\left(\mathbb{E}_{s_0}[R_{S_{N(T)+1}}^{\bar{\mathcal{G}}}(s_0, \mathcal{M})|N(T) + 1]\right)$$

$$\leq \frac{4C}{g} \cdot \mathbb{E}_{s_0}(\log([N(T) + 1])) + C'.$$

As already argued in the proof of (12), under any algorithm $\mathcal{G}$, we have $N(t) \leq A_2(t) + 1$ for all t and all sample paths, where $A_2(\cdot)$ is a Poisson process with rate $\lambda_{\max}$. Thus we have

$$\mathbb{E}_{s_0}\left(\log(N(T) + 1)\right) \leq \mathbb{E}\left(\log(A_2(T) + 2)\right) \leq \log(\lambda_{\max}T + 2),$$

where $A_2(T)$ is a Poisson random variable with mean $\lambda_{\max}T$, and the second inequality above is due to Jensen's inequality. Therefore we obtain

$$\mathbb{E}_{s_0}[R_{S_{N(T)+1}}^{\bar{\mathcal{G}}}(s_0, \mathcal{M})] \leq \frac{4C}{g} \cdot \log\left(\lambda_{\max}T + 2\right) + C'.$$

Since $r(s, a) \in [0, 1]$, the long run average reward ρ^* is upper bounded by $\lambda_{\max}$. Thus we can infer from (64) that

$$\mathbb{E}_{s_0}[R_T^{\bar{\mathcal{G}}}(s_0, \mathcal{M})] \leq \frac{4C}{g} \cdot \log\left(\lambda_{\max}T + 2\right) + C' + \frac{\lambda_{\max}}{\lambda_{\min}}.$$

The proof is complete. □