

A Tractable Algorithm For Finite-Horizon Continuous Reinforcement Learning

Phanideep Gampa
 Mathematical Science
 Indian Institute of Technology (BHU),
 Varanasi, India 221005
 gampa.phanideep.mat15@iitbhu.ac.in

Sairam Satwik Kondamudi
 Computer Science and Engineering
 Indian Institute of Technology,
 Hyderabad, India 502285
 satwik7545@gmail.com

K. Lakshmanan
 Computer Science and Engineering
 Indian Institute of Technology (BHU),
 Varanasi, India 221005
 lakshmanank.cse@iitbhu.ac.in

Abstract—We consider the finite horizon continuous reinforcement learning problem. Our contribution is three-fold. First, we give a tractable algorithm based on optimistic value iteration for the problem. Next, we give a lower bound on regret of order $\Omega(T^{2/3})$ for any algorithm discretizes the state space, improving the previous regret bound of $\Omega(T^{1/2})$ of Ortner and Ryabko [1] for the same problem. Next, under the assumption that the rewards and transitions are Hölder Continuous we show that the upper bound on the discretization error is $const.Ln^{-\alpha}T$. Finally, we give some simple experiments to validate our propositions.

Index Terms—Reinforcement Learning, Markov Decision Process (MDP), Regret, Continuous State Space, Bonus, Finite Horizon.

I. INTRODUCTION

In most reinforcement learning algorithms the size of the state and action spaces are assumed to be finite. But many real-world applications have continuous state or action spaces. Though the case of continuous state space has been considered before like in Ortner and Ryabko [1] and Lakshmanan et al. [2] there is no tractable algorithm whose regret has been analyzed. In this work we consider a simpler finite-horizon setting and give a tractable algorithm with near-optimal regret bound.

There have been many assumptions on the rewards and transition function considered before like deterministic transitions in [3] or transition functions that are linear in state and action [4], [5], [6] and [7]. Kakade et al. [8] have considered more general setting of PAC learning in RL with metric state spaces. Another result is by Osband and Roy [9] where bounds on expected regret was derived when reward and transition functions belonged to a class of functions characterized by parameters like eluder dimension and Kolmogorov dimension. In this paper we consider the most general assumption that the reward function and the transition functions are Hölder continuous like the papers [1] and [2]. We derive our bound on the discretization error without any further assumptions. By assuming unbiased sampling, we also derive an upper bound on the regret which is $O(T^{2/3})$ when both transition probabilities and reward function are Lipschitz continuous.

In recent work, Azar et al. [10] have proposed a new algorithm Optimistic Value Iteration (UCBVI). This is different from the UCRL algorithm in Jacksh et al [11] upon which the algorithms of Ortner and Ryabko [1] was based. For episodic problems, it is known that UCBVI algorithm has better regret bounds. We extend this UCBVI algorithm for continuous state

space problems using the state space aggregation used in [1]. It should also be noted that both the algorithms in [1] and [2] are not tractable. Though we have considered an easier episodic setting, our algorithm is tractable. And we show that the discretization error (discretizing the state space to intervals) is bounded above by $const.Ln^{-\alpha}T$ under the assumptions that rewards and transitions are Hölder Continuous.

For infinite horizon setting, in the one-dimensional case with Lipschitz rewards and transition functions the bound by Ortner and Ryabko is of order $\tilde{O}(T^{3/4})$ in [1]. If the transition functions are in addition smooth, a regret bound of $O(T^{2/3})$ is shown in [2].

The best known lower bound for Lipschitz reward and general transition function is by Ortner and Ryabko [1] which is order $\Omega(T^{1/2})$. We improve upon this by giving a lower bound of $\Omega(T^{2/3})$, matching the upper bound for algorithms that discretizes the state-space. We have also implemented the algorithm and give experimental results on one and two dimensional setting. The empirical performance is seen to match with the theoretical results.

II. TERMINOLOGY AND PROBLEM DEFINITION

We define Markov Decision Process (MDP) by state space S , action space A , reward function $r : S \times A \rightarrow [0, 1]$ and transition probability $P : S \times A \times S \rightarrow [0, 1]$. We consider continuous state space $[0, 1]^d$ and finite action MDP. For simplicity we derive results for the one dimensional state space, they can be easily generalized to higher dimensional state spaces. The random rewards given state s and action a are bounded in $[0, 1]$ with mean $r(s, a)$. The probability of going to state s' given state s and action a is given by the transition probability $p(s'|s, a)$. We make the following assumptions like in [1], [2]. This guarantees that rewards and transitions are close in close states.

Assumption 1. There are $L, \alpha > 0$ such that for any two states s, s' and all actions a ,

$$|r(s, a) - r(s', a)| \leq L|s - s'|^\alpha. \quad (1)$$

Assumption 2. There are $L, \alpha > 0$ such that for any two states s, s' and all actions a ,

$$\|p(\cdot|s, a) - p(\cdot|s', a)\|_1 \leq L|s - s'|^\alpha. \quad (2)$$

For simplicity we assume L, α are same for both assumptions.

We consider the finite horizon setting where the agent interacts with the environment in H number of steps per episode. We denote by $[n]$ the set $\{i \in \mathbb{N}, 1 \leq i \leq n\}$. The policy during an episode is expressed as a mapping $\pi : S \times [H] \rightarrow A$. Let $x_{k,h}$ denote the state in the step h of the episode k and π_k denote the policy for the episode k . The value function of each state in k^{th} episode from step h by following a policy π_k is defined by

$$V_h^{\pi_k}(s) = \mathbb{E}_{\pi_k} \left(\sum_{i=h}^H r(x_{k,i}, \pi_k(x_{k,i}, i)) | x_{k,h} = s \right). \quad (3)$$

The *optimal* value function is defined as $V_h^*(x) = \sup_{\pi} V_h^{\pi}(x)$ for all $x \in S$ and $h \geq 1$. The performance of the algorithm is measured according to regret incurred in all the episodes as given by

$$\text{Regret}(K) = \sum_{k=1}^K (V_1^*(x_{k,1}) - V_1^{\pi_k}(x_{k,1})). \quad (4)$$

We discretize the continuous state space into intervals of length $1/n$ like in [1]. Let $I(s)$ denote interval of the state $s \in S$. We define *aggregate rewards* and *aggregate transition probabilities* with respect to $\{I_1, I_2, \dots, I_n\}$ as

$$p^{agg}(I_j | s, a) = \int_{I_j} p(ds' | s, a) \quad (5)$$

$$r^{agg}(I_j, a) = n \cdot \int_{I_j} r(s, a) ds. \quad (6)$$

Here $r^{agg}(I_j, a)$ can be interpreted as the mean reward in the interval I_j . The algorithm treats each interval as single state and the aggregate policy is $\pi_k^{agg}(I(s), i)$. The aggregated value function following this policy is $V_{\pi_k, h}^{agg}(I_j)$. At any state $s \in S$ and step $i \in [H]$, the mapping between the policies are given by

$$\pi_k(s, i) = \pi_k^{agg}(I(s), i) \quad (7)$$

$$V_{\pi_k, h}^{agg}(I) = \mathbb{E}_{\pi_k^{agg}} \left(\sum_{i=h}^H r^{agg}(I_{k,i}, \pi_k^{agg}(I_{k,i}, i)) | I_{k,h} = I \right). \quad (8)$$

For the discretised MDP, the transitions are between intervals, so we define $P^{agg}(I|I', a)$ as the average of $p^{agg}(I|t, a)$ in I'

$$P^{agg}(I|I', a) = n \cdot \int_{I'} p^{agg}(I|t, a) dt \quad (9)$$

III. ALGORITHM

In UCRL [11] based algorithms confidence sets are built around rewards and transition probabilities. But here we build the confidence set around optimal value function V_h^* of the discretized MDP as in Azar et al ([10]). The algorithm proceeds similar to UCBVI in [10]. The only difference is that here we use aggregated rewards, transitions probabilities and value functions. The bonus (Algorithm 2) which is used in calculating the Q-Values is built from the empirical variance

of the estimated next values. This relies on the Bernstein-Freedman's concentration inequality for building the concentration sets.

In UCRL based algorithms [11] we need to find the optimistic MDP and optimal policy for that MDP. This step is not tractable when the state space is continuous as we need knowledge of the bias span H (section 3 of [1]). Though this is not needed for finite-horizon problems, to the best of our knowledge, UCCRL algorithm is not computationally tractable. We note that in our algorithm which is based on UCBVI, this step is not needed. Briefly the algorithm consists of the following parts/sections.

- 1) Initialization part which also includes the discretizing the continuous MDP based on the input parameters. This step is executed only once.
- 2) The next part consists of an iterative flow of three processes. (The number of iterations is equal to the number of episodes K)
 - Estimating the transition probabilities based on the history till that iterate.
 - Finding the Q values using a modified bellman operator (which includes bonus) according to the current transition probabilities. This part can be considered as a simple Dynamic Programming (DP) problem for finding Q values.
 - Execute the current (discrete) policy (according to the Q values found) and record the feedback given by the environment.

A. Bonus

This algorithm follows the heuristic principle known as *optimism in the face of uncertainty*. The algorithms in this paradigm employ optimism to guide exploration. In simple words, the algorithm assigns higher bonus value to that action (given a state/interval) which it is most uncertain about. This encourages the exploration of that action next time when it visits the same state again. Indirectly as the number of times the action chosen increases the algorithm would become more and more certain about it. Given a (state/interval, action) pair, the bonus in the algorithm depends on the variance of value function of the next possible states. Larger the variance, larger the uncertainty of taking that action.

Let us understand this using a simple example. Consider a discretized MDP with 4 possible actions. Given a particular interval I_k , let the Q values and the corresponding bonus found by the algorithm be equal to those given in the figure 1. In the case of not using bonus, the best action would be $a1$ because $(I_k, a1)$ has the highest Q-value in $\{(I_k, a_i) | i = 1, 2, 3, 4\}$. But when bonus is considered, the action $a4$ is the most likely to be chosen. According to the bonus, the decreasing order of uncertainties of actions are $a4, a2, a1, a3$. The combined effect of Q value and the bonus will lead to the exploration of action $a4$. Due to this exploration of action $a4$, the uncertainty about choosing $a4$ reduces.

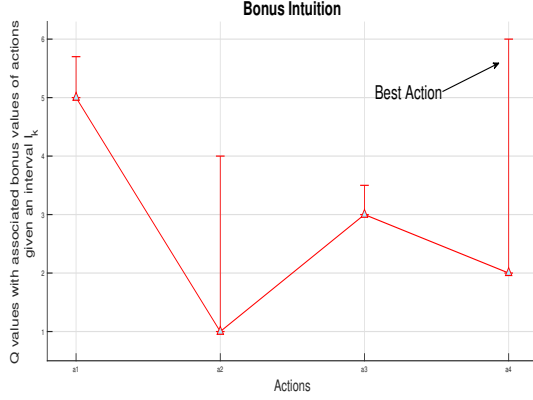


Fig. 1: Q values and bonus of four actions when an interval is fixed for an example MDP

IV. RESULTS

The regret analysis of UCBVI-CRL is straight forward. The regret of the continuous MDP upto K episodes is defined in equation (4). Now add and subtract terms $V_1^{agg^*}(I_{k,1})$, $V_{\pi_{k,1}}^{agg}(I_j)$ to this to get,

$$\begin{aligned} \text{Regret}(K) &= \sum_{k=1}^K ([V_1^*(x_{k,1}) - V_1^{agg^*}(I_{k,1}) \\ &\quad + V_{\pi_{k,1}}^{agg}(I_{k,1}) - V_1^{\pi k}(x_{k,1})] \\ &\quad + \underbrace{[V_1^{agg^*}(I_{k,1}) - V_{\pi_{k,1}}^{agg}(I_{k,1})]}_{\Delta_{agg}}]) \\ \Delta_{error} &= \sum_{k=1}^K ([V_1^*(x_{k,1}) - V_1^{agg^*}(I_{k,1}) \\ &\quad + V_{\pi_{k,1}}^{agg}(I_{k,1}) - V_1^{\pi k}(x_{k,1})] \end{aligned}$$

Here Δ_{agg} is the regret of the discretized MDP and Δ_{error} is the error which is a consequence of discretizing the state space.

A. Bounding Δ_{error}

Lemma 1. Using the discretization technique mentioned above, Δ_{error} is bounded above by $\text{const.} \cdot \ln^{-\alpha} T$.

Proof. Let us split the terms in Δ_{error} into two parts as

$$\begin{aligned} q_1^k &= V_1^*(x_{k,1}) - V_1^{agg^*}(I_{k,1}) \text{ and} \\ q_2^k &= V_{\pi_{k,1}}^{agg}(I_{k,1}) - V_1^{\pi k}(x_{k,1}) \\ \Delta_{error} &\leq \sum_{k=1}^K (|q_1^k| + |q_2^k|) \quad (\text{Triangle Inequality}) \end{aligned}$$

Consider $|q_1^k|$

$$V_1^*(x_{k,1}) = \max_{a \in A} (r(x_{k,1}, a) + \int_S p(ds' | x_{k,1}, a) V_2^*(s')) \quad (\text{Chapter 4 of [12]})$$

Algorithm 1 UCBVI-CRL

Input: State space $[0, 1]$, an action space A , discretization parameter $n \in \mathbb{N}$, Horizon length H , Hölder parameters L, α , confidence parameter δ , reward function r of continuous MDP.

Initialization: Let $I_1 = [0, \frac{1}{n}]$ and $I_j = (\frac{j-1}{n}, \frac{j}{n}]$ for $j = 2, 3, 4, \dots, n$. Set $t = 1$, $\hat{H} = \phi$ and observe the initial state s_1 and interval $I(s_1)$. Set $S^* = \{I_1, I_2, \dots, I_n\}$. For all $x \in S^*$, $N_k(x, a)$ is number of times action a has been chosen in a state $\in x$. Similarly with $N_k(x, a, y)$, $N'_{k,h}(x, a)$ where $x, y \in S^*$.

1: **for** $k = 1, 2, \dots$ **do**

Initialize Episode k:

Compute for all $(x, a, y) \in S^* \times A \times S^*$,

$$N_k(x, a, y) = \sum_{(x', a', y') \in \hat{H}} \mathbb{I}(x' = x, a' = a, y' = y)$$

$$N_k(x, a) = \sum_{y \in S^*} N_k(x, a, y)$$

$$N'_{k,h}(x, a) = \sum_{(x_i, h, a_i, x_{i,h+1}) \in \hat{H}} \mathbb{I}(x_i, h = x, a_i, h = a)$$

Let $\kappa = \{(x, a) \in S^* \times A, N_k(x, a) > 0\}$

Estimate $\hat{P}_k^{agg}(y|x, a) = \frac{N_k(x, a, y)}{N_k(x, a)}$, $\forall (x, a) \in \kappa$, using all samples from states interval y .

Initialize $V_{k, H+1}^{agg}(x) = 0$ for all $x \in S^*$

Calculating Q-values:

2: **for** $h = H, H-1, \dots, 1$ **do**

3: **for** $(x, a) \in S^* \times A$ **do**

4: **if** $(x, a) \in \kappa$ **then**

$$b_{k,h}(x, a) = \text{bonus}(\hat{P}_k^{agg}, V_{k, h+1}^{agg}, N_k, N'_{k,h})$$

$$Q_{k,h}^{agg}(x, a) = \min(Q_{k-1, h}^{agg}(x, a), H, r^{agg}(x, a) + (\hat{P}_k^{agg} V_{k, h+1}^{agg})(x, a) + b_{k,h}(x, a))$$

5: **else**

$$Q_{k,h}^{agg}(x, a) = H$$

6: **end if**

$$V_{k,h}^{agg}(x) = \max_{a \in A} (Q_{k,h}^{agg}(x, a))$$

7: **end for**

8: **end for**

Executing the policy:

9: **for** $h = 1, 2, \dots, H$ **do**

Take action $a_{k,h} = \arg \max_a (Q_{k,h}^{agg}(x_{k,h}, a))$

Update $\hat{H} = \hat{H} \cup (x_{k,h}, a_{k,h}, x_{k, h+1})$

10: **end for**

11: **end for**

$$\begin{aligned} V_1^{agg^*}(I_{k,1}) &= \max_{a \in A} (r^{agg}(I_{k,1}, a) \\ &\quad + \sum_{j=1}^n P^{agg}(I_j | I_{k,1}, a) V_2^{agg^*}(I_j)) \end{aligned} \quad (\text{Bellman Equation})$$

Using the property $|\max f - \max g| \leq \max |f - g|$

$$|q_1^k| \leq \max_{a \in A} |r(x_{k,1}, a) - r^{agg}(I_{k,1}, a)|$$

$$\begin{aligned} &+ \left| \int_S p(ds' | x_{k,1}, a) V_2^*(s') \right. \\ &\quad \left. - \sum_{j=1}^n P^{agg}(I_j | I_{k,1}, a) V_2^{agg^*}(I_j) \right| \end{aligned}$$

Algorithm 2 Bonus

Input: $\hat{P}_k^{agg}(x, a), V_{k,h+1}^{agg}, N_k, N'_{k,h}$, confidence parameter δ .

$$\begin{aligned} \text{Let } b(x, a) &= \sqrt{\frac{8LVar_{Y \sim \hat{P}_k^{agg}(\cdot|x,a)}(V_{k,h+1}^{agg}(Y))}{N_k(x, a)}} \\ &+ \frac{14HL}{3N_k(x, a)} \\ &+ \sqrt{\frac{8 \sum_y \hat{P}_k^{agg}(y|x, a) [\min(\frac{100H^3 S^2 AL^2}{N'_{k,h+1}(y)}, H^2)]}{N_k(x, a)}} \end{aligned}$$

where $L = \ln(\frac{5nAT}{\delta})$

return b

$$\begin{aligned} \text{Let } l_1^k &= \max_{a \in A} |r(x_{k,1}, a) - r^{agg}(I_{k,1}, a)| \text{ and} \\ l_2^k &= \max_{a' \in A} \left| \int_S p(ds'|x_{k,1}, a') V_2^*(s') \right. \\ &\quad \left. - \sum_{j=1}^n P^{agg}(I_j|I_{k,1}, a') V_2^{agg*}(I_j) \right| \end{aligned}$$

$$\begin{aligned} \text{Using the property } \max |f + g| &\leq \max |f| + \max |g| \\ |q_1^k| &\leq (l_1^k + l_2^k) \end{aligned}$$

For any action $a \in A$, l_1^k is bounded by

$$\begin{aligned} |r(x_{k,1}, a) - r^{agg}(I_{k,1}, a)| &= |r(x_{k,1}, a) - n \cdot \int_{I_{k,1}} r(s, a) ds| \\ &\quad \text{(from (6))} \\ &\leq \left| r(x_{k,1}, a) - n \cdot \left(\sum_{i=1}^n m_i \Delta x_i \right) \right| \\ &= \left| r(x_{k,1}, a) - n \cdot \left(\sum_{i=1}^n r(s'_i, a) \frac{1}{n^2} \right) \right| \\ &\leq \frac{1}{n} \cdot \sum_{i=1}^n |r(x_{k,1}, a) - r(s'_i, a)| \leq Ln^{-\alpha}. \end{aligned}$$

The first inequality is obtained by replacing the integral with lower Riemann sum by dividing $I_{k,1}$ into n intervals each of length $\Delta x_i = \frac{1}{n^2}$ and m_i being the infimum in the sub-interval. The second inequality follows since function r is continuous w.r.t s and infimum is the minimum. The last inequality follows from the assumption (1) and length of $I_j = \frac{1}{n}$.

For any action $a \in A$, l_2^k is bounded by

$$\begin{aligned} &\leq |H \sum_{j=1}^n \int_{I_j} p(ds'|x_{k,1}, a) - \sum_{j=1}^n P^{agg}(I_j|I_{k,1}, a) V_2^{agg*}(I_j)| \\ &\quad \text{(rewards } \in [0, 1]) \\ &\leq \sum_{j=1}^n |H \int_{I_j} p(ds'|x_{k,1}, a) - P^{agg}(I_j|I_{k,1}, a) V_2^{agg*}(I_j)| \\ \text{Using } |ab - cd| &\leq |(a+c)(b-d)| + |bc - ad| \\ a =: H \text{ and } b =: &\int_{I_j} p(ds'|x_{k,1}, a) \\ c =: V_2^{agg*}(I_j) \text{ and } d =: &P^{agg}(I_j|I_{k,1}, a) \\ &\leq \sum_{j=1}^n [|(V_2^{agg*}(I_j) + H) \\ &\quad (\int_{I_j} p(ds'|x_{k,1}, a) - P^{agg}(I_j|I_{k,1}, a))| \\ &\quad + |V_2^{agg*}(I_j) \int_{I_j} p(ds'|x_{k,1}, a) - P^{agg}(I_j|I_{k,1}, a)H|] \\ l_2^k &\leq \sum_{j=1}^n [3H \int_{I_j} p(ds'|x_{k,1}, a) - P^{agg}(I_j|I_{k,1}, a)] \\ &\quad (V_2^{agg*}(I_j) \leq H - 1) \end{aligned}$$

From (9), replacing the integral with lower Riemann sum with n sub intervals and using the property of continuity that infimum of a function is equal to minimum in that closed interval.

$$\begin{aligned} l_2^k &\leq \sum_{j=1}^n \left[3H \int_{I_j} p(ds'|x_{k,1}, a) - n \cdot \left(\sum_{i=1}^n p^{agg}(I_j|s'_i, a) \frac{1}{n^2} \right) \right] \\ &\leq \sum_{j=1}^n \left[\frac{3H}{n} \sum_{i=1}^n \left| \int_{I_j} p(ds'|x_{k,1}, a) - p^{agg}(I_j|s'_i, a) \right| \right] \\ &= \sum_{j=1}^n \left[\frac{3H}{n} \sum_{i=1}^n \left| \int_{I_j} p(ds'|x_{k,1}, a) - \int_{I_j} p(ds'|s'_i, a) \right| \right] \\ &= \frac{3H}{n} \sum_{i=1}^n \sum_{j=1}^n \left| \int_{I_j} p(ds'|x_{k,1}, a) - \int_{I_j} p(ds'|s'_i, a) \right| \\ &= \frac{3H}{n} \sum_{i=1}^n \|p(\cdot|x_{k,1}, a) - p(\cdot|s'_i, a)\|_1 \\ &\leq 3HLn^{-\alpha} \quad (x_{k,1}, s'_i \in I_{k,1} \forall i \text{ and using (2)}) \end{aligned}$$

Now we have,

$$|q_2^k| = \left| \mathbb{E}_{\pi^k} \left(\sum_{h=1}^H (r(x_{k,h}, \pi_k) - r^{agg}(I'_{k,h}, \pi_k)) \right) \right|$$

(Using (7))

Hence $|q_2^k|$ can be bound by bounding the difference of the rewards inside the expectation which is similar to bounding l_1^k . For any action a we have,

$$|q_2^k| \leq HLn^{-\alpha}$$

Now Δ_{error} is bounded by

$$\begin{aligned} \Delta_{error} &= \sum_{k=1}^K (|q_1^k| + |q_2^k|) \leq \sum_{k=1}^K (4H + 1) Ln^{-\alpha} \\ &\leq 5Ln^{-\alpha} T \text{ (as } T \geq (KH)) \end{aligned}$$

□

V. ADDITIONAL RESULT: UPPER BOUND ON REGRET

We derived a tighter upper bound on the regret for continuous case (when using the proposed algorithm) when the transition estimates found by the algorithm are unbiased estimates of (9). The below theorem states that.

Theorem 1. *Let M be an MDP with continuous state space $[0, 1]$, A actions, known reward function r and unknown transitions satisfying assumptions 1,2 and the transition estimates are unbiased estimates of (9). Then with probability $1 - \delta$, the regret of UCVBI-CRL after T steps is upper bounded by*

$$5Ln^{-\alpha}T + 30L'\sqrt{nATH} + 2500H^2n^2AL^2 + 4H\sqrt{TL}' \quad (6)$$

where $L' = \ln(5HnAT/\delta)$.

Proof. The regret in the continuous setting can be split into Δ_{error} and Δ_{agg} . The bound on Δ_{error} comes from Lemma 1 and bound on Δ_{agg} is from Theorem 2 of [10]. \square

Corollary 1. *We set $n = T^x$. It can be seen that when $\alpha \in [1/k, 1]$, setting $x < \frac{k}{2k+1}$ first term dominates while the third term $2500H^2n^2AL^2$ dominates when $x > \frac{k}{2k+1}$. Thus the optimal regret of $O(L^2H^2AT^{2k/2k+1})$ is obtained by setting $x = \frac{k}{2k+1}$. This is better than the regret bound of order $T^{\frac{2k+1}{2k+2}}$ for the same in [1]. And in the Lipschitz case i.e., when $\alpha = 1$ we have regret of order $O(T^{2/3})$.*

Remark 1. *The dependence on H , L and A is not optimal. If the algorithms of [1] and [2] are adapted for finite-horizon setting we again wish to emphasize the fact that our algorithm is tractable unlike the algorithms in the other two papers.*

Remark 2. *The regret bound in [2] is also of order $\tilde{O}(T^{2/3})$ for the infinite horizon problem in the Lipschitz case. But there is an additional parameter β which depends on the smoothness of the transition function. And only in the asymptotic case when $\beta \rightarrow \infty$ the regret of $\tilde{O}(T^{2/3})$ is attained.*

Remark 3. *For d -dimensional state space we have n^d states in the last three terms of equation (6). So setting $n = T^{\frac{k}{2kd+1}}$ we get a regret of $\tilde{O}(T^{\frac{2kd}{2kd+1}})$ which is better than the bound of $\tilde{O}(T^{\frac{2kd+1}{2kd+2}})$ in [1] for the case $\alpha \in [1/k, 1]$.*

VI. LOWER BOUNDS

We have the following theorem giving the lower bound.

Theorem 2. *For any algorithm \mathcal{A} using state-space discretization, any natural numbers A and $T \geq A$, the expected regret for \mathcal{A} after T timesteps for the MDP \mathcal{M} with state $S = [0, 1]$, actions A defined above is*

$$E(\Delta_T(\mathcal{A})) \geq L\sqrt{HAT}^{2/3} \quad (9)$$

Proof. We note that any algorithm using state-space discretization also incurs a regret by choosing a wrong action due to discretization error. Hence we have

- 1) Regret in the discretized MDP
- 2) Regret due to the discretization error

The total regret is atleast maximum of these two. The first term is of order \sqrt{nHAT} (See [11]). Now consider the MDP shown in figure 2 with two actions in each state.

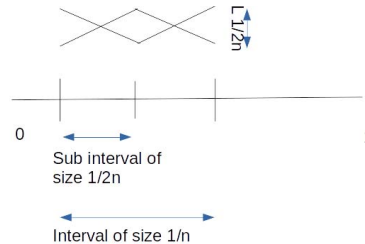


Fig. 2: State space and reward for the MDP

Let the algorithm discretize the state into n intervals. Let us assume that there are two sub-intervals in each interval and the optimal action in each sub-interval is different. Assume for simplicity that all intervals are of equal size. Otherwise divide each interval into sub-intervals of size half of the least interval size and assume constant reward for the remaining length of the intervals.

The mean rewards for actions are as shown in the figure 2, i.e., they are fixed for a state. It can be seen that it is not possible to determine the optimal action for both these sub-intervals simultaneously. And the regret obtained due to discretization is of order $L \cdot \frac{1}{2n}$ as the rewards are Lipschitz continuous. Since this regret is incurred on an average for atleast half of the states visited $T/2$. The total regret is of order $L \cdot \frac{T}{4n}$. Balancing this and the regret in the discretized MDP by taking $n = T^{1/3}$ we get a regret of order $\Omega(L\sqrt{HAT}^{2/3})$. \square

Remark 4. *We note that algorithms like value iteration and non-stationary value iteration algorithms (Chapter 3 of [13]) which do not use state-space discretization cannot be used here as the probability transition function is not known.*

Remark 5. *It is straight forward to show a regret bound of $\Omega(L\sqrt{H_{sp}AT}^{2/3})$ for infinite horizon setting where H_{sp} is the span of the optimal bias function.*

VII. EXPERIMENTS

We have performed our experiments on one dimensional and two dimensional state spaces. For one-dimensional case we implemented the algorithm on a simple problem with $A = 2$ with $[0, 5\pi]$ being the state space S . In this problem, the reward functions for the actions a_1, a_2 are $\sin(x)$ and $-\sin(x)$ respectively. So the optimal policy is, for a state x in interval $[n\pi, (n+1)\pi]$, the one which takes action a_2 if n is odd and

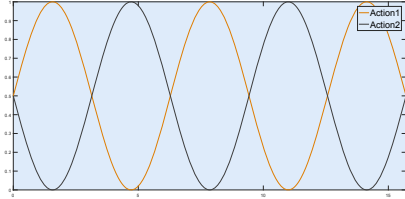


Fig. 3: 1-D Reward Functions

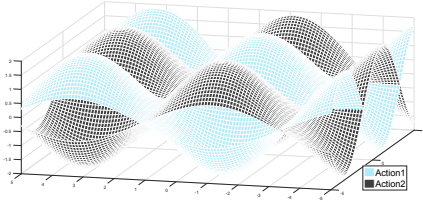
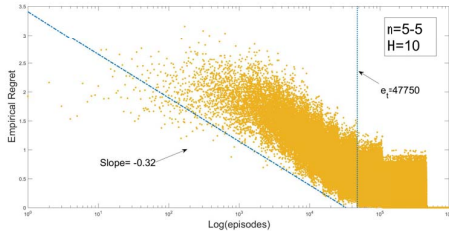
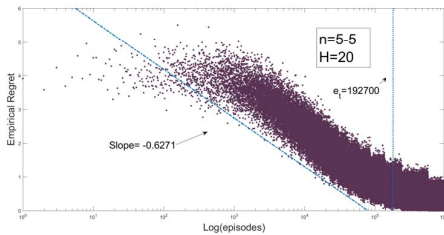


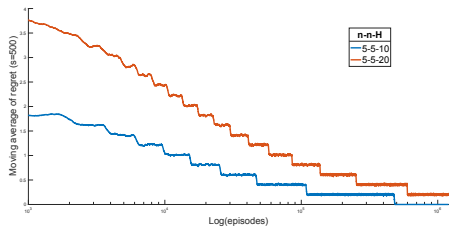
Fig. 4: 2D reward functions plot



(a)



(b)



(c)

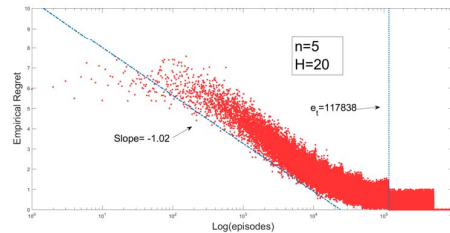
Fig. 5: Plots (a), (b) show the regret on log scale of the two-dimensional problem for different settings. Plot (c) is obtained by taking moving average on the regret.

a_1 if n is even. The reward was scaled to the range $[0, 1]$ to

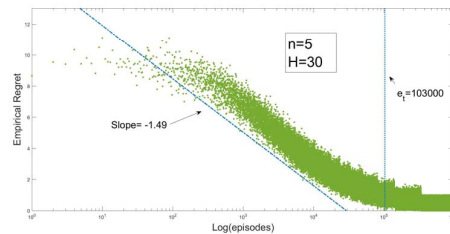
meet the requirements of [10]. From the current state x , after taking an action a_i such that $i \in 1, 2$, the next state is sampled uniformly from the state space S .

We can see from Figure 6 that the empirical regret is converging faster for larger number of intervals, for the same horizon length. The convergence can be better understood by comparing the the points where the empirical regret approaches to 0. The very first point, in every plot, where the empirical regret is zero is indicated by a dashed vertical line showing the episode number. The results match the intuition that for the same problem having more number of states improves the performance.

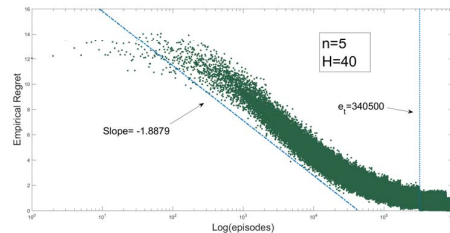
By keeping n constant and varying H it can be seen from the plots that regret approaches zero slower as H is increases. It can also be seen that the regret for larger horizon length at a particular episode was higher than the regret at the same episode for smaller horizon length. Thus the regret increases with the horizon length as expected.



(a)



(b)



(c)

Fig. 6: Plots (a), (b), (c) show the regret for different settings in the one-dimensional case.

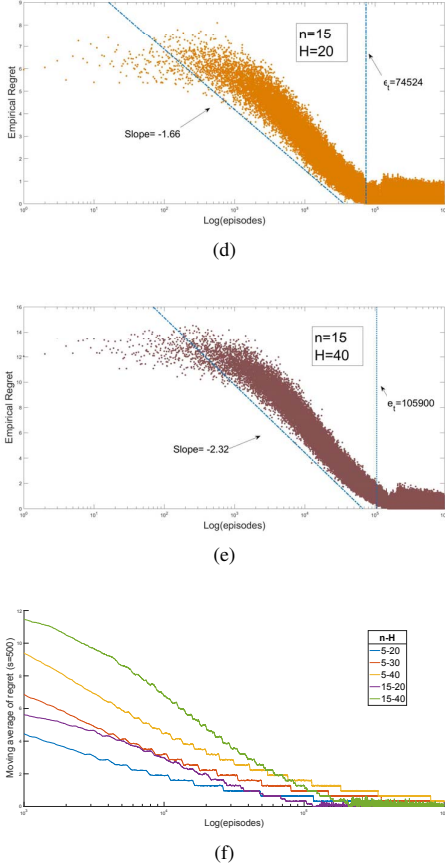


Fig. 6: Plots (d) and (e) show the regret for different settings in the one-dimensional case. Plot (f) is obtained by taking the moving averages of regret for the same case. All the plots are log-scaled along the episode-axis.

Next we conducted experiments on a two-dimensional problem with $A = 2$ and a Lipschitz continuous reward function for the respective actions. The reward functions were $\sin(\frac{x-y}{\sqrt{2}}) + \cos(\frac{x+y}{\sqrt{2}})$ and $\sin(\frac{y-x}{\sqrt{2}}) - \cos(\frac{x+y}{\sqrt{2}})$ for actions a_1 and a_2 respectively. The analysis of this case is also similar to the one-dimensional setting. The algorithm shows similar trend for different values of H for same number of intervals n (see figure 5) i.e, regret approaches zero slower for higher values of H as in one-dimensional setting.

VIII. CONCLUSION

We considered the finite-horizon continuous reinforcement learning problem. We have given an algorithm based on UCBVI for the same. With the only assumption that the reward function and transition probabilities are Lipschitz continuous we show that the upper bound on the discretization error is $const.Ln^{-\alpha}T$. We have shown the matching lower bound under the assumption that the algorithm discretizes the state-space of MDP. In future we would like to show similar bound without this assumption. Also, with an additional assumption

that the sampling is unbiased, we proved that the regret is of order $O(T^{2/3})$, when using our algorithm.

We have also given some experimental results to validate our propositions. In future we would like to extend the algorithm to infinite-horizon case. This seems to be difficult as the UCBVI algorithm needs the horizon length H as input. We also want to improve the dependence on size of action A , Lipschitz constant L and horizon length H on the regret.

REFERENCES

- [1] R. Ortner and D. Ryabko, "Online regret bounds for undiscounted continuous reinforcement learning," in *Advances in Neural Information Processing Systems NIPS*, vol. 25, pp. 1763–1772, 2012.
- [2] K. Lakshmanan, R. Ortner, and D. Ryabko, "Improved regret bounds for undiscounted continuous reinforcement learning," in *Proceedings of 32nd International Conference on Machine Learning (ICML)*, vol. 37, pp. 524–32, 2015.
- [3] A. Bernstein and N. Shimkin, "Adaptive-resolution reinforcement learning with polynomial exploration in deterministic domains," *Machine Learning*, vol. 81, no. 3, pp. 359–397, 2010.
- [4] A. L. Strehl and M. L. Littman, "Online linear regression and its application to model-based reinforcement learning," in *Advances Neural Information Processing Systems NIPS*, vol. 20, pp. 1417–1424, 2008.
- [5] E. Brunskill, B. R. Leffler, L. Li, M. L. Littman, and N. Roy, "Provably efficient learning with typed parametric models," *Journal of Machine Learning Research*, vol. 10, pp. 1955–1988, 2009.
- [6] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Learning Theory, 24th Annual Conference on Learning Theory COLT, JMLR Proceedings Track*, vol. 19, pp. 1–26, 2011.
- [7] M. Ibrahim, A. Javanmard, and B. V. Roy, "Efficient reinforcement learning for high dimensional linear quadratic systems," in *Advances Neural Information Processing Systems NIPS*, vol. 25, pp. 2645–2653, 2012.
- [8] S. Kakade, M. J. Kearns, and J. Langford, "Exploration in metric state spaces," in *Proceedings of 20th International Conference on Machine Learning ICML*, pp. 306–312, 2003.
- [9] I. Osband and B. V. Roy, "Model-based reinforcement learning and the eluder dimension," vol. Preprint, 2014.
- [10] M. G. Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," *arXiv preprint arXiv:1703.05449*, 2017.
- [11] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, pp. 1563–1600, 2010.
- [12] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*, vol. 30. Springer Science & Business Media, 2012.
- [13] O. Hernández-Lerma, *Adaptive Markov control processes*, vol. 79 of *Applied Mathematical Sciences*. Springer, 1989.