
Online Regret Bounds for Undiscounted Continuous Reinforcement Learning

Ronald Ortner*[†]

*Montanuniversitaet Leoben
8700 Leoben, Austria
rortner@unileoben.ac.at

Daniil Ryabko[†]

[†]INRIA Lille-Nord Europe, équipe SequeL
59650 Villeneuve d'Ascq, France
daniil@ryabko.net

Abstract

We derive sublinear regret bounds for undiscounted reinforcement learning in continuous state space. The proposed algorithm combines state aggregation with the use of upper confidence bounds for implementing optimism in the face of uncertainty. Beside the existence of an optimal policy which satisfies the Poisson equation, the only assumptions made are Hölder continuity of rewards and transition probabilities.

1 Introduction

Real world problems usually demand continuous state or action spaces, and one of the challenges for reinforcement learning is to deal with such continuous domains. In many problems there is a natural metric on the state space such that close states exhibit similar behavior. Often such similarities can be formalized as Lipschitz or more generally Hölder continuity of reward and transition functions.

The simplest continuous reinforcement learning problem is the 1-dimensional continuum-armed bandit, where the learner has to choose arms from a bounded interval. Bounds on the regret with respect to an optimal policy under the assumption that the reward function is Hölder continuous have been given in [15, 4]. The proposed algorithms apply the UCB algorithm [2] to a discretization of the problem. That way, the regret suffered by the algorithm consists of the loss by aggregation (which can be bounded using Hölder continuity) plus the regret the algorithm incurs in the discretized setting. More recently, algorithms that adapt the used discretization (making it finer in more promising regions) have been proposed and analyzed [16, 8].

While the continuous bandit case has been investigated in detail, in the general case of continuous state Markov decision processes (MDPs) a lot of work is confined to rather particular settings, primarily with respect to the considered transition model. In the simplest case, the transition function is considered to be deterministic as in [18], and mistake bounds for the respective discounted setting have been derived in [6]. Another common assumption is that transition functions are linear functions of state and action plus some noise. For such settings sample complexity bounds have been given in [22, 7], while $\tilde{O}(\sqrt{T})$ bounds for the regret after T steps are shown in [1]. However, there is also some research considering more general transition dynamics under the assumption that close states behave similarly, as will be considered here. While most of this work is purely experimental [12, 23], there are also some contributions with theoretical guarantees. Thus, [13] considers PAC-learning for continuous reinforcement learning in metric state spaces, when generative sampling is possible. The proposed algorithm is a generalization of the E^3 algorithm [14] to continuous domains. A respective adaptive discretization approach is suggested in [19]. The PAC-like bounds derived there however depend on the (random) behavior of the proposed algorithm.

Here we suggest a learning algorithm for undiscounted reinforcement learning in continuous state space. The proposed algorithm is in the tradition of algorithms like UCRL2 [11] in that it implements

the “optimism in the face of uncertainty” maxim, here combined with state aggregation. Thus, the algorithm does not need a generative model or access to “resets:” learning is done online, that is, in a single continual session of interactions between the environment and the learning policy.

For our algorithm we derive regret bounds of $\tilde{O}(T^{(2+\alpha)/(2+2\alpha)})$ for MDPs with 1-dimensional state space and Hölder-continuous rewards and transition probabilities with parameter α . These bounds also straightforwardly generalize to dimension d where the regret is bounded by $\tilde{O}(T^{(2d+\alpha)/(2d+2\alpha)})$. Thus, in particular, if rewards and transition probabilities are Lipschitz, the regret is bounded by $\tilde{O}(T^{(2d+1)/(2d+2)})$ in dimension d and $\tilde{O}(T^{3/4})$ in dimension 1. We also present an accompanying lower bound of $\Omega(\sqrt{T})$. As far as we know, these are the first regret bounds for a general undiscounted continuous reinforcement learning setting.

2 Preliminaries

We consider the following setting. Given is a Markov decision process (MDP) M with state space $\mathcal{S} = [0, 1]^d$ and finite action space \mathcal{A} . For the sake of simplicity, in the following we assume $d = 1$. However, proofs and results generalize straightforwardly to arbitrary dimension, cf. Remark 5 below. The random rewards in state s under action a are assumed to be bounded in $[0, 1]$ with mean $r(s, a)$. The transition probability distribution in state s under action a is denoted by $p(\cdot|s, a)$.

We will make the natural assumption that rewards and transition probabilities are similar in close states. More precisely, we assume that rewards and transition probabilities are *Hölder continuous*.

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

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

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

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

For the sake of simplicity we will assume that α and L in Assumptions 1 and 2 are the same.

We also assume existence of an optimal policy $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$ which gives optimal average reward $\rho^* = \rho^*(M)$ on M independent of the initial state. A sufficient condition for state-independent optimal reward is geometric convergence of π^* to an invariant probability measure. This is a natural condition which e.g. holds for any communicating finite state MDP. It also ensures (cf. Chapter 10 of [10]) that the Poisson equation holds for the optimal policy. In general, under suitable technical conditions (like geometric convergence to an invariant probability measure μ_π) the *Poisson equation*

$$\rho_\pi + \lambda_\pi(s) = r(s, \pi(s)) + \int_{\mathcal{S}} p(ds'|s, \pi(s)) \cdot \lambda_\pi(s') \quad (1)$$

relates the rewards and transition probabilities under any measurable policy π to its average reward ρ_π and the *bias* function $\lambda_\pi : \mathcal{S} \rightarrow \mathbb{R}$ of π . Intuitively, the bias is the difference in accumulated rewards when starting in a different state. Formally, the bias is defined by the Poisson equation (1) and the normalizing equation $\int_{\mathcal{S}} \lambda_\pi d\mu_\pi = 0$ (cf. e.g. [9]). The following result follows from the bias definition and Assumptions 1 and 2 (together with results from Chapter 10 of [10]).

Proposition 3. *Under Assumptions 1 and 2, the bias of the optimal policy is bounded.*

Consequently, it makes sense to define the *bias span* $H(M)$ of a continuous state MDP M satisfying Assumptions 1 and 2 to be $H(M) := \sup_s \lambda_{\pi^*}(s) - \inf_s \lambda_{\pi^*}(s)$. Note that since $\inf_s \lambda_{\pi^*}(s) \leq 0$ by definition of the bias, the bias function λ_{π^*} is upper bounded by $H(M)$.

We are interested in algorithms which can compete with the optimal policy π^* and measure their performance by the *regret* (after T steps) defined as $T\rho^*(M) - \sum_{t=1}^T r_t$, where r_t is the random reward obtained by the algorithm at step t . Indeed, within T steps no *canonical* or even *bias optimal* optimal policy (cf. Chapter 10 of [10]) can obtain higher accumulated reward than $T\rho^* + H(M)$.

3 Algorithm

Our algorithm UCCRL, shown in detail in Figure 1, implements the “optimism in the face of uncertainty maxim” just like UCRL2 [11] or REGAL [5]. It maintains a set of plausible MDPs \mathcal{M} and

Algorithm 1 The UCCRL algorithm

Input: State space $\mathcal{S} = [0, 1]$, action space \mathcal{A} , confidence parameter $\delta > 0$, aggregation parameter $n \in \mathbb{N}$, upper bound H on the bias span, Lipschitz parameters L, α .

Initialization:

- ▷ Let $I_1 := [0, \frac{1}{n}]$, $I_j := (\frac{j-1}{n}, \frac{j}{n}]$ for $j = 2, 3, \dots, n$.
- ▷ Set $t := 1$, and observe the initial state s_1 and interval $I(s_1)$.

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

- ▷ Let $N_k(I_j, a)$ be the number of times action a has been chosen in a state $\in I_j$ prior to episode k , and $v_k(I_j, a)$ the respective counts in episode k .

Initialize episode k :

- ▷ Set the start time of episode k , $t_k := t$.
- ▷ Compute estimates $\hat{r}_k(s, a)$ and $\hat{p}_k^{\text{agg}}(I_j|s, a)$ for rewards and transition probabilities, using all samples from states in the same interval $I(s)$, respectively.

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

- ▷ Let \mathcal{M}_k be the set of plausible MDPs \tilde{M} with $H(\tilde{M}) \leq H$ and rewards $\tilde{r}(s, a)$ and transition probabilities $\tilde{p}(\cdot|s, a)$ satisfying

$$|\tilde{r}(s, a) - \hat{r}_k(s, a)| \leq Ln^{-\alpha} + \sqrt{\frac{7 \log(2nAt_k/\delta)}{2 \max\{1, N_k(I(s), a)\}}}, \quad (2)$$

$$\left\| \tilde{p}^{\text{agg}}(\cdot|s, a) - \hat{p}_k^{\text{agg}}(\cdot|s, a) \right\|_1 \leq Ln^{-\alpha} + \sqrt{\frac{56n \log(2At_k/\delta)}{\max\{1, N_k(I(s), a)\}}}. \quad (3)$$

- ▷ Choose policy $\tilde{\pi}_k$ and $\tilde{M}_k \in \mathcal{M}_k$ such that

$$\rho_{\tilde{\pi}_k}(\tilde{M}_k) = \arg \max\{\rho^*(M) \mid M \in \mathcal{M}_k\}. \quad (4)$$

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

while $v_k(I(s_t), \tilde{\pi}_k(s_t)) < \max\{1, N_k(I(s_t), \tilde{\pi}_k(s_t))\}$ **do**

- ▷ Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward r_t , and observe next state s_{t+1} .
- ▷ Set $t := t + 1$.

end while

end for

chooses optimistically an MDP $\tilde{M} \in \mathcal{M}$ and a policy $\tilde{\pi}$ such that the average reward $\rho_{\tilde{\pi}}(\tilde{M})$ is maximized, cf. (4). Whereas for UCRL2 and REGAL the set of plausible MDPs is defined by confidence intervals for rewards and transition probabilities for each individual state-action pair, for UCCRL we assume an MDP to be plausible if its *aggregated* rewards and transition probabilities are within a certain range. This range is defined by the aggregation error (determined by the assumed Hölder continuity) and respective confidence intervals, cf. (2), (3). Correspondingly, the estimates for rewards and transition probabilities for some state action-pair (s, a) are calculated from all sampled values of action a in states close to s .

More precisely, for the aggregation UCCRL partitions the state space into intervals $I_1 := [0, \frac{1}{n}]$, $I_k := (\frac{k-1}{n}, \frac{k}{n}]$ for $k = 2, 3, \dots, n$. The corresponding aggregated transition probabilities are defined by

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

Generally, for a (transition) probability distribution $p(\cdot)$ over \mathcal{S} we write $p^{\text{agg}}(\cdot)$ for the aggregated probability distribution with respect to $\{I_1, I_2, \dots, I_n\}$. Now, given the aggregated state space $\{I_1, I_2, \dots, I_n\}$, estimates $\hat{r}(s, a)$ and $\hat{p}^{\text{agg}}(\cdot|s, a)$ are calculated from all samples of action a in states in $I(s)$, the interval I_j containing s . (Consequently, the estimates are the same for states in the same interval.)

As UCRL2 and REGAL, UCCRL proceeds in episodes in which the chosen policy remains fixed. Episodes are terminated when the number of times an action has been sampled from some interval I_j has been doubled. Only then estimates are updated and a new policy is calculated.

Since all states in the same interval I_j have the same confidence intervals, finding the optimal pair $\tilde{M}_k, \tilde{\pi}_k$ in (4) is equivalent to finding the respective optimistic discretized MDP \tilde{M}_k^{agg} and an optimal policy $\tilde{\pi}_k^{\text{agg}}$ on \tilde{M}_k^{agg} . Then $\tilde{\pi}_k$ can be set to be the extension of $\tilde{\pi}_k^{\text{agg}}$ to \mathcal{S} , that is, $\tilde{\pi}_k(s) := \tilde{\pi}_k^{\text{agg}}(I(s))$ for all s . However, due to the constraint on the bias even in this finite case efficient computation of \tilde{M}_k^{agg} and $\tilde{\pi}_k^{\text{agg}}$ is still an open problem. We note that the REGAL.C algorithm [5] selects optimistic MDP and optimal policy in the same way as UCCRL.

While the algorithm presented here is the first modification of UCRL2 to continuous reinforcement learning problems, there are similar adaptations to online aggregation [20] and learning in finite state MDPs with some additional similarity structure known to the learner [21].

4 Regret Bounds

For UCCRL we can derive the following bounds on the regret.

Theorem 4. *Let M be an MDP with continuous state space $[0, 1]$, A actions, rewards and transition probabilities satisfying Assumptions 1 and 2, and bias span upper bounded by H . Then with probability $1 - \delta$, the regret of UCCRL (run with input parameters n and H) after T steps is upper bounded by*

$$\text{const} \cdot nH \sqrt{AT \log\left(\frac{T}{\delta}\right)} + \text{const}' \cdot HLn^{-\alpha}T. \quad (6)$$

Therefore, setting $n = T^{1/(2+2\alpha)}$ gives regret upper bounded by

$$\text{const} \cdot HL \sqrt{A \log\left(\frac{T}{\delta}\right)} \cdot T^{(2+\alpha)/(2+2\alpha)}.$$

With no known upper bound on the bias span, guessing H by $\log T$ one still obtains an upper bound on the regret of $\tilde{O}(T^{(2+\alpha)/(2+2\alpha)})$.

Intuitively, the second term in the regret bound of (6) is the discretization error, while the first term corresponds to the regret on the discretized MDP. A detailed proof of Theorem 4 can be found in Section 5 below.

Remark 5 (d -dimensional case). The general d -dimensional case can be handled as described for dimension 1, with the only difference being that the discretization now has n^d states, so that one has n^d instead of n in the first term of (6). Then choosing $n = T^{1/(2d+2\alpha)}$ bounds the regret by $\tilde{O}(T^{(2d+\alpha)/(2d+2\alpha)})$.

Remark 6 (unknown horizon). If the horizon T is unknown then the doubling trick (executing the algorithm in rounds $i = 1, 2, \dots$ guessing $T = 2^i$ and setting the confidence parameter to $\delta/2^i$) gives the same bounds.

Remark 7 (unknown Hölder parameters). The UCCRL algorithm receives (bounds on) the Hölder parameters L as α as inputs. If these parameters are not known, then one can still obtain sublinear regret bounds albeit with worse dependence on T . Specifically, we can use the model-selection technique introduced in [17]. To do this, fix a certain number J of values for the constants L and α ; each of these values will be considered as a model. The model selection consists in running UCCRL with each of these parameter values for a certain period of τ_0 time steps (*exploration*). Then one selects the model with the highest reward and uses it for a period of τ'_0 time steps (*exploitation*), while checking that its average reward stays within (6) of what was obtained in the exploitation phase. If the average reward does not pass this test, then the model with the second-best average reward is selected, and so on. Then one switches to exploration with longer periods τ_1 , etc. Since there are no guarantees on the behavior of UCCRL when the Hölder parameters are wrong, none of the models can be discarded at any stage. Optimizing over the parameters τ_i and τ'_i as done in [17], and increasing the number J of considered parameter values, one can obtain regret bounds of $\tilde{O}(T^{(2+2\alpha)/(2+3\alpha)})$, or $\tilde{O}(T^{4/5})$ in the Lipschitz case. For details see [17]. Since in this model-selection process UCCRL is used in a “black-box” fashion, the exploration is rather wasteful, and thus we think that this bound is suboptimal. Recent developments suggest that results of [17] can be improved and consequently the dependence on T in the regret bound in the case of unknown Hölder parameters can be made the same as in the known parameters case.

The following is a complementing lower bound on the regret for continuous state reinforcement learning.

Theorem 8. For any $A, H > 1$ and any reinforcement learning algorithm there is a continuous state reinforcement learning problem with A actions and bias span H satisfying Assumption 1 such that the algorithm suffers regret of $\Omega(\sqrt{HAT})$.

Proof. Consider the following reinforcement learning problem with state space $[0, 1]$. The state space is partitioned into n intervals I_j of equal size. The transition probabilities for each action a are on each of the intervals I_j concentrated and equally distributed on the same interval I_j . The rewards on each interval I_j are also constant for each a and are chosen as in the lower bounds for a multi-armed bandit problem [3] with nA arms. That is, giving only one arm slightly higher reward, it is known [3] that regret of $\Omega(\sqrt{nAT})$ can be forced upon any algorithm on the respective bandit problem. Adding another action giving no reward and equally distributing over the whole state space, the bias span of the problem is n and the regret $\Omega(\sqrt{HAT})$. \square

Remark 9. Note that Assumption 2 does not hold in the example used in the proof of Theorem 8. However, the transition probabilities are piecewise constant (and hence Lipschitz) and known to the learner. Actually, it is straightforward to deal with piecewise Hölder continuous rewards and transition probabilities where the finitely many points of discontinuity are known to the learner. If one makes sure that the intervals of the discretized state space do not contain any discontinuities, it is easy to adapt UCCRL and Theorem 4 accordingly.

Remark 10 (comparison to bandits). The bounds of Theorems 4 and 8 cannot be directly compared to bounds for the continuous-armed bandit problem [15, 4, 16, 8], because the latter is no special case of learning MDPs with continuous state space (and rather corresponds to a continuous action space). Thus, in particular one cannot freely sample an arbitrary state of the state space as assumed in continuous-armed bandits.

5 Proof of Theorem 4

For the proof of the main theorem we adapt the proof of the regret bounds for finite MDPs in [11] and [5]. Although the state space is now continuous, due to the finite horizon T , we can reuse some arguments, so that we keep the structure of the original proof of Theorem 2 in [11]. Some of the necessary adaptations made are similar to techniques used for showing regret bounds for other modifications of the original UCRL2 algorithm [20, 21], which however only considered finite-state MDPs.

5.1 Splitting into Episodes

Let $v_k(s, a)$ be the number of times action a has been chosen in episode k when being in state s , and denote the total number of episodes by m . Then setting $\Delta_k := \sum_{s,a} v_k(s, a)(\rho^* - r(s, a))$, with probability at least $1 - \frac{\delta}{12T^{5/4}}$ the regret of UCCRL after T steps is upper bounded by (cf. Section 4.1 of [11]),

$$\sqrt{\frac{5}{8}T \log\left(\frac{8T}{\delta}\right)} + \sum_{k=1}^m \Delta_k. \quad (7)$$

5.2 Failing Confidence Intervals

Next, we consider the regret incurred when the true MDP M is not contained in the set of plausible MDPs \mathcal{M}_k . Thus, fix a state-action pair (s, a) , and recall that $\hat{r}(s, a)$ and $\hat{p}^{\text{agg}}(\cdot|s, a)$ are the estimates for rewards and transition probabilities calculated from all samples of state-action pairs contained in the same interval $I(s)$. Now assume that at step t there have been $N > 0$ samples of action a in states in $I(s)$ and that in the i -th sample a transition from state $s_i \in I(s)$ to state s'_i has been observed ($i = 1, \dots, N$).

First, concerning the rewards one obtains as in the proof of Lemma 17 in Appendix C.1 of [11] — but now using Hoeffding for independent and not necessarily identically distributed random variables — that

$$\Pr\left\{\left|\hat{r}(s, a) - \mathbb{E}[\hat{r}(s, a)]\right| \geq \sqrt{\frac{7}{2N} \log\left(\frac{2nAt}{\delta}\right)}\right\} \leq \frac{\delta}{60nAt^7}. \quad (8)$$

Concerning the transition probabilities, we have for a suitable $x \in \{-1, 1\}^n$

$$\begin{aligned}
& \left\| \hat{p}^{\text{agg}}(\cdot|s, a) - \mathbb{E}[\hat{p}^{\text{agg}}(\cdot|s, a)] \right\|_1 = \sum_{j=1}^n \left| \hat{p}^{\text{agg}}(I_j|s, a) - \mathbb{E}[\hat{p}^{\text{agg}}(I_j|s, a)] \right| \\
& = \sum_{j=1}^n \left(\hat{p}^{\text{agg}}(I_j|s, a) - \mathbb{E}[\hat{p}^{\text{agg}}(I_j|s, a)] \right) x(I_j) \\
& = \frac{1}{N} \sum_{i=1}^N \left(x(I(s'_i)) - \int_{\mathcal{S}} p(ds'|s_i, a) \cdot x(I(s')) \right). \tag{9}
\end{aligned}$$

For any $x \in \{-1, 1\}^n$, $X_i := x(I(s'_i)) - \int_{\mathcal{S}} p(ds'|s_i, a) \cdot x(I(s'))$ is a martingale difference sequence with $|X_i| \leq 2$, so that by Azuma-Hoeffding inequality (e.g., Lemma 10 in [11]), $\Pr\{\sum_{i=1}^N X_i \geq \theta\} \leq \exp(-\theta^2/8N)$ and in particular

$$\Pr\left\{\sum_{i=1}^N X_i \geq \sqrt{56nN \log\left(\frac{2At}{\delta}\right)}\right\} \leq \left(\frac{\delta}{2At}\right)^{7n} \leq \frac{\delta}{2^n 20nAt^7}.$$

A union bound over all sequences $x \in \{-1, 1\}^n$ then yields from (9) that

$$\Pr\left\{\left\|\hat{p}^{\text{agg}}(\cdot|s, a) - \mathbb{E}[\hat{p}^{\text{agg}}(\cdot|s, a)]\right\|_1 \geq \sqrt{\frac{56n}{N} \log\left(\frac{2At}{\delta}\right)}\right\} \leq \frac{\delta}{20nAt^7}. \tag{10}$$

Another union bound over all t possible values for N , all n intervals and all actions shows that the confidence intervals in (8) and (10) hold at time t with probability at least $1 - \frac{\delta}{15t^6}$ for the actual counts $N(I(s), a)$ and all state-action pairs (s, a) . (Note that the equations (8) and (10) are the same for state-action pairs with states in the same interval.)

Now, by linearity of expectation $\mathbb{E}[\hat{r}(s, a)]$ can be written as $\frac{1}{N} \sum_{i=1}^N r(s_i, a)$. Since the s_i are assumed to be in the same interval $I(s)$, it follows that $|\mathbb{E}[\hat{r}(s, a)] - r(s, a)| < Ln^{-\alpha}$. Similarly, $\left\|\mathbb{E}[\hat{p}^{\text{agg}}(\cdot|s, a)] - p^{\text{agg}}(\cdot|s, a)\right\|_1 < Ln^{-\alpha}$. Together with (8) and (10) this shows that with probability at least $1 - \frac{\delta}{15t^6}$ for all state-action pairs (s, a)

$$|\hat{r}(s, a) - r(s, a)| < Ln^{-\alpha} + \sqrt{\frac{7 \log(2nAt/\delta)}{2 \max\{1, N(I(s), a)\}}}, \tag{11}$$

$$\left\|\hat{p}^{\text{agg}}(\cdot|s, a) - p^{\text{agg}}(\cdot|s, a)\right\|_1 < Ln^{-\alpha} + \sqrt{\frac{56n \log(2At/\delta)}{\max\{1, N(I(s), a)\}}}. \tag{12}$$

This shows that the true MDP is contained in the set of plausible MDPs $\mathcal{M}(t)$ at step t with probability at least $1 - \frac{\delta}{15t^6}$, just as in Lemma 17 of [11]. The argument that

$$\sum_{k=1}^m \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} \leq \sqrt{T} \tag{13}$$

with probability at least $1 - \frac{\delta}{12T^{5/4}}$ then can be taken without any changes from Section 4.2 of [11].

5.3 Regret in Episodes with $M \in \mathcal{M}_k$

Now for episodes with $M \in \mathcal{M}_k$, by the optimistic choice of \tilde{M}_k and $\tilde{\pi}_k$ in (4) we can bound

$$\begin{aligned}
\Delta_k & = \sum_s v_k(s, \tilde{\pi}_k(s)) (\rho^* - r(s, \tilde{\pi}_k(s))) \\
& \leq \sum_s v_k(s, \tilde{\pi}_k(s)) (\tilde{\rho}_k^* - r(s, \tilde{\pi}_k(s))) \\
& = \sum_s v_k(s, \tilde{\pi}_k(s)) (\tilde{\rho}_k^* - \tilde{r}_k(s, \tilde{\pi}_k(s))) + \sum_s v_k(s, \tilde{\pi}_k(s)) (\tilde{r}_k(s, \tilde{\pi}_k(s)) - r(s, \tilde{\pi}_k(s))).
\end{aligned}$$

Any term $\tilde{r}_k(s, a) - r(s, a) \leq |\tilde{r}_k(s, a) - \hat{r}_k(s, a)| + |\hat{r}_k(s, a) - r(s, a)|$ is bounded according to (2) and (11), as we assume that $\tilde{M}_k, M \in \mathcal{M}_k$, so that summarizing states in the same interval I_j

$$\Delta_k \leq \sum_s v_k(s, \tilde{\pi}_k(s)) (\tilde{\rho}_k^* - \tilde{r}_k(s, \tilde{\pi}_k(s))) + 2 \sum_{j=1}^n \sum_{a \in \mathcal{A}} v_k(I_j, a) \left(Ln^{-\alpha} + \sqrt{\frac{7 \log(2nAt_k/\delta)}{2 \max\{1, N_k(I_j, a)\}}} \right).$$

Since $\max\{1, N_k(I_j, a)\} \leq t_k \leq T$, setting $\tau_k := t_{k+1} - t_k$ to be the length of episode k we have

$$\begin{aligned} \Delta_k &\leq \sum_s v_k(s, \tilde{\pi}_k(s)) (\tilde{\rho}_k^* - \tilde{r}_k(s, \tilde{\pi}_k(s))) \\ &\quad + 2Ln^{-\alpha} \tau_k + \sqrt{14 \log \left(\frac{2nAT}{\delta} \right)} \sum_{j=1}^n \sum_{a \in \mathcal{A}} \frac{v_k(I_j, a)}{\sqrt{\max\{1, N_k(I_j, a)\}}}. \end{aligned} \quad (14)$$

We continue analyzing the first term on the right hand side of (14). By the Poisson equation (1) for $\tilde{\pi}_k$ on \tilde{M}_k , denoting the respective bias by $\tilde{\lambda}_k := \tilde{\lambda}_{\tilde{\pi}_k}$ we can write

$$\begin{aligned} &\sum_s v_k(s, \tilde{\pi}_k(s)) (\tilde{\rho}_k^* - \tilde{r}_k(s, \tilde{\pi}_k(s))) \\ &= \sum_s v_k(s, \tilde{\pi}_k(s)) \left(\int_{\mathcal{S}} \tilde{p}_k(ds'|s, \tilde{\pi}_k(s)) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s) \right) \\ &= \sum_s v_k(s, \tilde{\pi}_k(s)) \left(\int_{\mathcal{S}} p(ds'|s, \tilde{\pi}_k(s)) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s) \right) \end{aligned} \quad (15)$$

$$+ \sum_s v_k(s, \tilde{\pi}_k(s)) \sum_{j=1}^n \int_{I_j} \left(\tilde{p}_k(ds'|s, \tilde{\pi}_k(s)) - p(ds'|s, \tilde{\pi}_k(s)) \right) \cdot \tilde{\lambda}_k(s'). \quad (16)$$

5.4 The True Transition Functions

Now $\|\tilde{p}_k^{\text{agg}}(\cdot|s, a) - p^{\text{agg}}(\cdot|s, a)\|_1 \leq \|\tilde{p}_k^{\text{agg}}(\cdot|s, a) - \hat{p}_k^{\text{agg}}(\cdot|s, a)\|_1 + \|\hat{p}_k^{\text{agg}}(\cdot|s, a) - p^{\text{agg}}(\cdot|s, a)\|_1$ can be bounded by (3) and (12), because we assume $\tilde{M}_k, M \in \mathcal{M}_k$. Hence, since by definition of the algorithm H bounds the bias function $\tilde{\lambda}_k$, the term in (16) is bounded by

$$\begin{aligned} &\sum_s v_k(s, \tilde{\pi}_k(s)) \sum_{j=1}^n \int_{I_j} \tilde{\lambda}_k(s') \left(\tilde{p}_k(ds'|s, \tilde{\pi}_k(s)) - p(ds'|s, \tilde{\pi}_k(s)) \right) \\ &\leq \sum_s v_k(s, \tilde{\pi}_k(s)) \cdot H \cdot \sum_{j=1}^n \left(\tilde{p}_k^{\text{agg}}(I_j|s, \tilde{\pi}_k(s)) - p^{\text{agg}}(I_j|s, \tilde{\pi}_k(s)) \right) \\ &\leq \sum_s v_k(s, \tilde{\pi}_k(s)) \cdot H \cdot 2 \left(Ln^{-\alpha} + \sqrt{\frac{56n \log(2AT/\delta)}{\max\{1, N_k(I(s), a_t)\}}} \right) \\ &= 2HLn^{-\alpha} \tau_k + 4H \sqrt{14n \log \left(\frac{2AT}{\delta} \right)} \sum_{j=1}^n \sum_{a \in \mathcal{A}} \frac{v_k(I_j, a)}{\sqrt{\max\{1, N_k(I_j, a)\}}}, \end{aligned} \quad (17)$$

while for the term in (15)

$$\begin{aligned} &\sum_s v_k(s, \tilde{\pi}_k(s)) \left(\int_{\mathcal{S}} p(ds'|s, \tilde{\pi}_k(s)) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s) \right) \\ &= \sum_{t=t_k}^{t_{k+1}-1} \left(\int_{\mathcal{S}} p(ds'|s_t, a_t) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s_t) \right) \\ &= \sum_{t=t_k}^{t_{k+1}-1} \left(\int_{\mathcal{S}} p(ds'|s_t, a_t) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s_{t+1}) \right) + \tilde{\lambda}_k(s_{t_{k+1}}) - \tilde{\lambda}_k(s_{t_k}). \end{aligned}$$

Let $k(t)$ be the index of the episode time step t belongs to. Then the sequence $X_t := \int_{\mathcal{S}} p(ds'|s_t, a_t) \cdot \tilde{\lambda}_{k(t)}(s') - \tilde{\lambda}_{k(t)}(s_{t+1})$ is a sequence of martingale differences so that Azuma-Hoeffding inequality shows (cf. Section 4.3.2 and in particular eq. (18) in [11]) that after summing over all episodes we have

$$\begin{aligned} &\sum_{k=1}^m \left(\sum_{t=t_k}^{t_{k+1}-1} \left(\int_{\mathcal{S}} p(ds'|s_t, a_t) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s_{t+1}) \right) + \tilde{\lambda}_k(s_{t_{k+1}}) - \tilde{\lambda}_k(s_{t_k}) \right) \\ &\leq H \sqrt{\frac{5}{2} T \log \left(\frac{8T}{\delta} \right)} + HnA \log_2 \left(\frac{8T}{nA} \right), \end{aligned} \quad (18)$$

where the second term comes from an upper bound on the number of episodes, which can be derived analogously to Appendix C.2 of [11].

5.5 Summing over Episodes with $M \in \mathcal{M}_k$

To conclude, we sum (14) over all the episodes with $M \in \mathcal{M}_k$, using (15), (17), and (18). This yields that with probability at least $1 - \frac{\delta}{12T^{5/4}}$

$$\begin{aligned} \sum_{k=1}^m \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} &\leq 2HLn^{-\alpha}T + 4H\sqrt{14n \log\left(\frac{2AT}{\delta}\right)} \cdot \sum_{k=1}^m \sum_{j=1}^n \sum_{a \in \mathcal{A}} \frac{v_k(I_j, a)}{\sqrt{\max\{1, N_k(I_j, a)\}}} \\ &+ H\sqrt{\frac{5}{2}T \log\left(\frac{8T}{\delta}\right)} + HnA \log_2\left(\frac{8T}{nA}\right) \\ &+ 2Ln^{-\alpha}T + \sqrt{14 \log\left(\frac{2nAT}{\delta}\right)} \sum_{k=1}^m \sum_{j=1}^n \sum_{a \in \mathcal{A}} \frac{v_k(I_j, a)}{\sqrt{\max\{1, N_k(I_j, a)\}}}. \end{aligned} \quad (19)$$

Analogously to Section 4.3.3 and Appendix C.3 of [11], one can show that

$$\sum_{j=1}^n \sum_{a \in \mathcal{A}} \sum_k \frac{v_k(I_j, a)}{\sqrt{\max\{1, N_k(I_j, a)\}}} \leq (\sqrt{2} + 1)\sqrt{nAT},$$

and we get from (19) after some simplifications that with probability $\geq 1 - \frac{\delta}{12T^{5/4}}$

$$\begin{aligned} \sum_{k=1}^m \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} &\leq H\sqrt{\frac{5}{2}T \log\left(\frac{8T}{\delta}\right)} + HnA \log_2\left(\frac{8T}{nA}\right) \\ &+ \left((4H + 1)\sqrt{14n \log\left(\frac{2AT}{\delta}\right)}\right)(\sqrt{2} + 1)\sqrt{nAT} + 2(H + 1)Ln^{-\alpha}T. \end{aligned} \quad (20)$$

Finally, evaluating (7) by summing Δ_k over all episodes, by (13) and (20) we have with probability $\geq 1 - \frac{\delta}{4T^{5/4}}$ an upper bound on the regret of

$$\begin{aligned} &\sqrt{\frac{5}{8}T \log\left(\frac{8T}{\delta}\right)} + \sum_{k=1}^m \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} + \sum_{k=1}^m \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} \\ &\leq \sqrt{\frac{5}{8}T \log\left(\frac{8T}{\delta}\right)} + \sqrt{T} + H\sqrt{\frac{5}{2}T \log\left(\frac{8T}{\delta}\right)} + HnA \log_2\left(\frac{8T}{nA}\right) \\ &+ \left((4H + 1)\sqrt{14n \log\left(\frac{2AT}{\delta}\right)}\right)(\sqrt{2} + 1)\sqrt{nAT} + 2(H + 1)Ln^{-\alpha}T. \end{aligned}$$

A union bound over all possible values of T and further simplifications as in Appendix C.4 of [11] finish the proof. \square

6 Outlook

We think that a generalization of our results to continuous action space should not pose any major problems. In order to improve over the given bounds, it may be promising to investigate more sophisticated discretization patterns.

The assumption of Hölder continuity is an obvious, yet not the only possible assumption one can make about the transition probabilities and reward functions. A more general problem is to assume a set \mathcal{F} of functions, find a way to measure the “size” of \mathcal{F} , and derive regret bounds depending on this size of \mathcal{F} .

Acknowledgments

The authors would like to thank the three anonymous reviewers for their helpful suggestions and Rémi Munos for useful discussion which helped to improve the bounds. This research was funded by the Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and FEDER (Contrat de Projets Etat Region CPER 2007-2013), ANR projects EXPLO-RA (ANR-08-COSI-004), Lampada (ANR-09-EMER-007) and CoAdapt, and by the European Community’s FP7 Program under grant agreements n° 216886 (PASCAL2) and n° 270327 (CompLACS). The first author is currently funded by the Austrian Science Fund (FWF): J 3259-N13.

References

- [1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. *COLT 2011, JMLR Proceedings Track*, 19:1–26, 2011.
- [2] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47:235–256, 2002.
- [3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32:48–77, 2002.
- [4] Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007*, pages 454–468, 2007.
- [5] Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, pages 25–42, 2009.
- [6] Andrey Bernstein and Nahum Shimkin. Adaptive-resolution reinforcement learning with polynomial exploration in deterministic domains. *Mach. Learn.*, 81(3):359–397, 2010.
- [7] Emma Brunskill, Bethany R. Leffler, Lihong Li, Michael L. Littman, and Nicholas Roy. Provably efficient learning with typed parametric models. *J. Mach. Learn. Res.*, 10:1955–1988, 2009.
- [8] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. Online optimization of χ -armed bandits. In *Advances in Neural Information Processing Systems 22, NIPS 2009*, pages 201–208, 2010.
- [9] Onésimo Hernández-Lerma and Jean Bernard Lasserre. *Discrete-time Markov control processes*, volume 30 of *Applications of mathematics*. Springer, 1996.
- [10] Onésimo Hernández-Lerma and Jean Bernard Lasserre. *Further topics on discrete-time Markov control processes*, volume 42 of *Applications of mathematics*. Springer, 1999.
- [11] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.
- [12] Nicholas K. Jong and Peter Stone. Model-based exploration in continuous state spaces. In *Abstraction, Reformulation, and Approximation, 7th International Symposium, SARA 2007*, pages 258–272. Springer, 2007.
- [13] Sham Kakade, Michael J. Kearns, and John Langford. Exploration in metric state spaces. In *Machine Learning, Proc. 20th International Conference, ICML 2003*, pages 306–312, 2003.
- [14] Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. *Mach. Learn.*, 49:209–232, 2002.
- [15] Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances Neural Information Processing Systems 17, NIPS 2004*, pages 697–704, 2005.
- [16] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proc. 40th Annual ACM Symposium on Theory of Computing, STOC 2008*, pages 681–690, 2008.
- [17] Odalric-Ambrym Maillard, Rémi Munos, and Daniil Ryabko. Selecting the state-representation in reinforcement learning. In *Advances Neural Processing Systems 24, NIPS 2011*, pages 2627–2635, 2012.
- [18] Gerhard Neumann, Michael Pfeiffer, and Wolfgang Maass. Efficient continuous-time reinforcement learning with adaptive state graphs. In *Machine Learning: ECML 2007, 18th European Conference on Machine Learning*, pages 250–261, 2007.
- [19] Ali Nouri and Michael L. Littman. Multi-resolution exploration in continuous spaces. In *Advances in Neural Information Processing Systems 21, NIPS 2008*, pages 1209–1216, 2009.
- [20] Ronald Ortner. Adaptive aggregation for reinforcement learning in average reward Markov decision processes. *Ann. Oper. Res.*, 2012. doi:10.1007/s10479-12-1064-y, to appear.
- [21] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless Markov bandits. In *Proc. 23rd Conference on Algorithmic Learning Theory, ALT 2012*, pages 214–228, 2012.
- [22] Alexander L. Strehl and Michael L. Littman. Online linear regression and its application to model-based reinforcement learning. In *Advances Neural Information Processing Systems 20, NIPS 2007*, pages 1417–1424, 2008.
- [23] William T. B. Uther and Manuela M. Veloso. Tree based discretization for continuous state space reinforcement learning. In *Proc. 15th National Conference on Artificial Intelligence and 10th Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98*, pages 769–774, 1998.