Journal of Artificial Intelligence Research 67 (2020) 115-128 Submitted 01/2019; published 01/2020

Regret Bounds for Reinforcement Learning via Markov Chain Concentration

Ronald Ortner RORTNER@UNILEOBEN.AC.AT
Department Mathematik und Informationstechnolgie
Montanuniversität Leoben
Franz-Josef-Strasse 18
8700 Leoben, Austria

Abstract

We give a simple optimistic algorithm for which it is easy to derive regret bounds of $\tilde{O}(\sqrt{t_{\rm mix}SAT})$ after T steps in uniformly ergodic Markov decision processes with S states, A actions, and mixing time parameter $t_{\rm mix}$. These bounds are the first regret bounds in the general, non-episodic setting with an optimal dependence on all given parameters. They could only be improved by using an alternative mixing time parameter.

1. Introduction

Starting with Burnetas and Katehakis (1997), regret bounds for reinforcement learning have addressed the question of how difficult it is to learn optimal behavior in an unknown Markov decision process (MDP). Some of these bounds —like the one derived by Burnetas and Katehakis (1997)— depend on particular properties of the underlying MDP, typically some kind of gap that specifies the distance between an optimal and a sub-optimal action or policy (see e.g. Ok, Proutière, & Tranos, 2018, for a recent refinement of such bounds). The first so-called problem independent bounds that have no dependence on any gap-parameter were obtained by Jaksch, Ortner, and Auer (2010). For MDPs with S states, A actions and diameter D the regret of the UCRL algorithm was shown to be $\tilde{O}(DS\sqrt{AT})$ after any T steps. A corresponding lower bound of $\Omega(\sqrt{DSAT})$ left the open question of the true dependence of the regret on the parameters S and D. Recently, regret bounds of $\tilde{O}(D\sqrt{SAT})$ have been claimed by Agrawal and Jia (2017), however there seems to be a gap in the proof, cf. Sec. 38.9 of Lattimore and Szepesvári (2019), so that the original bounds of Jaksch et al. (2010) are still the best known bounds.

In the simpler episodic setting, the gap between the upper and the lower bounds has been closed by Azar, Osband, and Munos (2017), showing that the regret is of order $\tilde{O}(\sqrt{HSAT})$, where H is the length of an episode. However, while bounds for the non-episodic setting can be easily transferred to the episodic setting, the reverse is not true. We also note that another kind of regret bounds that appears in the literature assumes an MDP sampled from some distribution (see e.g. Osband & Roy, 2017, for a recent contribution). Regret bounds in this Bayesian setting cannot be turned into bounds for the worst case setting as considered here.

There is also quite some work on bounds on the number of samples from a generative model necessary to approximate the optimal policy by an error of at most ε. Obviously, having access to a generative model makes learning the optimal policy easier than in the online setting considered here. However, for ergodic MDPs (on which we will focus in this note) it could be argued that any policy reaches any state so that in this case sample complexity bounds could in principle be turned into regret bounds. We first note that this seems difficult for bounds in the discounted setting, which make up the majority in the literature. Bounds in the discounted setting (see e.g. Azar, Munos, & Kappen, 2013b; Sidford, Wang, Wu, Yang, & Ye, 2018, for more recent contributions obtaining near-optimal bounds) depend on the term $1-\gamma$, where γ is the discount factor, and it is not clear how this term translates into a mixing time parameter in the average reward case. For the few results in the average reward setting the best sample complexity bound we are aware of is the bound of $\tilde{O}\big(\frac{\tau^2 t_{\text{mix}}^2 SA}{\varepsilon^2}\big)$ of Wang (2017), where t_{mix} is a mixing time parameter like ours (cf. below) and τ characterizes the range of stationary distributions across policies. Translated into respective regret bounds, these would have a worse (i.e., linear) dependence on the mixing time and would depend on the additional parameter $\tau > 1$, which does not appear in the bounds we are going to present below.

Starting with Kearns and Singh (2002) and Brafman and Tennenholtz (2002) there are also sample complexity bounds in the literature that were derived for settings without generative sampling model. Although this is obviously harder, there are bounds for the discounted case where the dependence with respect to S, A, and ε is the same as for the case with a generative sampling model (Szita & Szepesvári, 2010). However, we are not aware of any such bounds for the undiscounted setting that would translate into online regret bounds optimal in S, A, and T.

In this note, we present a simple algorithm that allows the derivation of regret bounds of $\tilde{O}(\sqrt{t_{\text{mix}} SAT})$ for uniformly ergodic MDPs with mixing time t_{mix}, a parameter that measures how long it takes to approximate the stationary distribution induced by any policy. These bounds are optimal with respect to the parameters S, A, T, and t_{mix}. The only possible improvement is a replacement of t_{mix} by a parameter that may be smaller for some MDPs, such as the diameter (Jaksch et al., 2010) or the bias span (Bartlett & Tewari, 2009; Fruit, Pirotta, Lazaric, & Ortner, 2018b). We note, however, that it is easy to give MDPs for which t_{mix} is basically of the same size as the mentioned alternative parameters.[1] Accordingly, the obtained bound basically closes the gap between the upper and the lower bound on the regret for a subclass of MDPs.

Algorithmically, the algorithm we propose works like an optimistic bandit algorithm such as UCB (Auer, Cesa-Bianchi, & Fischer, 2002). Such algorithms have been proposed before for MDP settings with a limited set of policies (Azar, Lazaric, & Brunskill, 2013a). The main difference to the latter approach is that due to the re-use of samples we obtain regret bounds that do not scale with the number of policies but with the number of state-action pairs. We note however that as the approach of Azar et al. (2013a) our algorithm needs to evaluate each policy independently, which makes it impractical. The proof of the regret bound is much simpler than for bounds achieved before and relies on concentration results for Markov chains.

[1] For a discussion of various transition parameters used in the literature we refer to Jaksch et al. (2010) and Bartlett and Tewari (2009).

2. Setting

We consider reinforcement learning in an average reward *Markov decision process* (*MDP*) with finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$. We assume that each stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ induces a uniformly ergodic[2] Markov chain on the state space. In such MDPs, which we call *uniformly ergodic*, the chain induced by a policy π has a unique stationary distribution μ_π. Then writing the mean reward for choosing an action a in a state s as $r(s,a)$, the (state-independent) average reward ρ_π can be written as $\rho_\pi = \boldsymbol{\mu}_\pi^\top \mathbf{r}_\pi$, where $\boldsymbol{\mu}_\pi = (\mu_\pi(s))_s$ and $\mathbf{r}_\pi = (r(s,\pi(s))_s$ are the (column) vectors for the stationary distribution and the average reward under π, respectively. We assume in the following that the reward distribution for each state-action pair (s,a) has support in $[0,1]$.

The maximal average reward is known (cf. Puterman, 1994) to be achieved by a stationary policy π^* that gives average reward $\rho^* := \rho_{\pi^*}$. We are interested in the *regret* accumulated by an algorithm after any number of T steps defined as[3]

$$R_T := T\rho^* - \sum_t r_t,$$

where r_t are the (random) rewards collected by the algorithm at each step t.

3. Preliminaries on Markov Chains

In this section, we give some definitions and results about Markov chain concentration that we will use in the following.

3.1 Mixing Times

For two distributions P, Q over the same state space $(\mathcal{S}, \mathcal{F})$ with σ-algebra $\mathcal{F}$, let

$$d_{TV}(P,Q) := \sup_{A\in\mathcal{F}} |P(A) - Q(A)|$$

be the *total variational distance* between P and Q. A Markov chain with a transition kernel p and a stationary distribution μ is said to be *uniformly ergodic*, if there are a $\theta < 1$ and a finite L such that

$$\sup_{s\in\mathcal{S}} d_{TV}(p^n(s,\cdot),\mu) \;\leq\; L\theta^n.$$

Furthermore, the *mixing time* t_{mix} of the Markov chain is defined as

$$t_{\mathrm{mix}} := \min\big\{n \mid \sup_{s\in\mathcal{S}} d_{TV}(p^n(s,\cdot),\mu) \leq \tfrac{1}{4}\big\}.$$

For a uniformly ergodic MDP we set the mixing time t^π_{mix} of a policy π to be the mixing time of the Markov chain induced by π, and define the *mixing time of the MDP* to be $t_{\mathrm{mix}} := \max_\pi t^\pi_{\mathrm{mix}}$.

[2] See Section 3 for definitions.

[3] Since we are only interested in upper bounds on this quantity we ignore the dependence on the initial state to keep things simpler. For a more detailed discussion we refer to Jaksch et al. (2010).

3.2 McDiarmid's Inequality for Markov Chains

Our results mainly rely on the following version of McDiarmid's inequality for Markov chains due to Paulin (2015).

Lemma 1. *(Corollary 2.10 and the following Remark 2.11 of Paulin, 2015)*
Consider a uniformly ergodic Markov chain $X_1, \ldots, X_n$ with state space $\mathcal{S}$ and mixing time t_{mix}. Let $f : \mathcal{S}^n \to \mathbb{R}$ with

$$f(s_1, \ldots, s_n) - f(s'_1, \ldots, s'_n) \;\leq\; \sum_i c_i \mathbb{1}[s_i \neq s'_i]. \tag{1}$$

Then

$$\mathbb{P}\Big\{\big|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]\big| \geq \varepsilon\Big\} \leq 2\exp\left(-\frac{2\varepsilon^2}{9\,\|c\|_2^2\, t_{\text{mix}}}\right).$$

Lemma 1 can be used to obtain a concentration result for the empirical average reward of any policy π in an MDP. This works analogously to the concentration bounds for the total variational distance between the empirical and the stationary distribution (Proposition 2.18 of Paulin, 2015).

Corollary 2. *Consider an MDP and a policy π that induces a uniformly ergodic Markov chain with mixing time t_{mix}. Using (column) vector notation $\boldsymbol{\mu} := (\mu_\pi(s))_s$ and $\mathbf{r} := (r(s, \pi(s))_s$ for the stationary distribution and the reward function under π, and writing $\hat{\boldsymbol{\mu}}^n$ for the empirical distribution after n steps defined as $\hat{\mu}^n(s) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i = s\}$, it holds that*

$$\mathbb{P}\Big\{\big|\hat{\boldsymbol{\mu}}^{n\top}\mathbf{r} - \boldsymbol{\mu}^\top\mathbf{r}\big| \geq \varepsilon\Big\} \leq 2\exp\left(-\frac{2\varepsilon^2 n}{9t_{\text{mix}}}\right).$$

Proof. Setting $f(X_1, \ldots, X_n) := \frac{1}{n}\big(r(X_1, \pi(X_1)) + \ldots + r(X_n, \pi(X_n))\big)$, condition (1) holds choosing $c_i = \frac{1}{n}$ for $i = 1, \ldots, n$ and the claim follows from Lemma 1. □

Choosing the error probability to be δ, we obtain the following confidence interval that will be used by our algorithm.

Corollary 3. *Using the same assumptions and notation of Corollary 2, with probability at least $1 - \delta$,*

$$\big|\hat{\boldsymbol{\mu}}^{n\top}\mathbf{r} - \boldsymbol{\mu}^\top\mathbf{r}\big| \;\leq\; \sqrt{\frac{9t_{\text{mix}}\log\frac{2}{\delta}}{2n}}.$$

3.3 Concentration of the Empirical Distribution

We will also use the following results on the concentration of the empirical state distribution of Markov chains given by Paulin (2015). In the following, consider a uniformly ergodic Markov chain $X_1, \ldots, X_n$ with a stationary distribution μ and a mixing time t_{mix}. Let $\hat{\mu}^n$ be the empirical distribution after performing n steps in the chain.

Lemma 4. *(Proposition 2.18 of Paulin, 2015)*

$$\mathbb{P}\Big\{\big|d_{TV}(\mu, \hat{\mu}^n) - \mathbb{E}[d_{TV}(\mu, \hat{\mu}^n)]\big| \geq \varepsilon\Big\} \leq 2\exp\left(-\frac{2\varepsilon^2 n}{9t_{\text{mix}}}\right).$$

Lemma 5. *(Proposition 3.16 and the following remark of Paulin, 2015)*

$$\mathbb{E}[d_{TV}(\mu, \hat{\mu}^n)] \leq \sum_{s \in \mathcal{S}} \min \left(\sqrt{\frac{8\mu(s)}{n\beta}}, \mu(s) \right),$$

where β is the pseudo-spectral gap[4] *of the chain.*

Lemma 6. *(Proposition 3.4 of Paulin, 2015) In uniformly ergodic Markov chains, the pseudo-spectral gap β can be bounded via the mixing time t_{mix} as*

$$\tfrac{1}{\beta} \leq 2t_{\text{mix}}.$$

We summarize these results in the following corollary.

Corollary 7. *With probability at least $1 - \delta$,*

$$d_{TV}(\mu, \hat{\mu}^n) \leq \sqrt{\frac{38 S t_{\text{mix}} \log \frac{2}{\delta}}{n}}.$$

Proof. Using the bound of Lemma 6 in Lemma 5 and setting the error probability in Lemma 4 to δ, one obtains by Jensen's inequality

$$d_{TV}(\mu, \hat{\mu}^n) \leq \sqrt{\frac{16 S t_{\text{mix}} \mu(s)}{n}} + \sqrt{\frac{9 t_{\text{mix}} \log \frac{2}{\delta}}{2n}},$$

and the claim of the corollary follows immediately. □

4. Algorithm

At the core, the Osp algorithm we propose works like the UCB algorithm in the bandit setting (Auer et al., 2002). In our case, each policy corresponds to an arm, and the concentration results of the previous chapter are used to obtain suitable confidence intervals for the MDP setting.

Osp (shown in detail as Algorithm 1) does not evaluate the policies at each time step. Instead, it proceeds in phases[5] (cf. line 3 of Osp), where in each phase k an optimistic policy π_k is selected (line 8). This is done (cf. line 5) by first constructing for each policy π a sample path $\mathcal{P}_\pi = \big((s_t, \pi(s_t), r_t, s_{t+1})\big)_{t=1}^n$ from the observations so far. Accordingly, the algorithm keeps a record of all observations. That is, after choosing in a state s an action a, obtaining the reward r, and observing a transition to the next state s', the respective observation (s, a, r, s') is appended to the sequence of observations $\mathcal{O}$ (cf. line 10).

The sample path $\mathcal{P}_\pi$ constructed from the observation sequence $\mathcal{O}$ contains each observation from $\mathcal{O}$ at most once. Further, the path $\mathcal{P}_\pi = \big((s_t, \pi(s_t), r_t, s_{t+1})\big)_{t=1}^n$ is such that there

[4] The pseudo-spectral gap is defined as $\max_k \left\{ \frac{\gamma(\mathbf{P}^{*k}\mathbf{P}^k)}{k} \right\}$, where $\mathbf{P}$ is the transition kernel interpreted as linear operator, $\mathbf{P}^*$ is the adjoint of $\mathbf{P}$, and $\gamma(\mathbf{P}^{*k}\mathbf{P}^k)$ is the spectral gap of the self-adjoint operator $\mathbf{P}^{*k}\mathbf{P}^k$. For more details see Paulin (2015). Here we do not make direct use of this quantity and only use the bound given in Lemma 6.

[5] We emphasize that we consider *non-episodic* reinforcement learning and that these phases are internal to the algorithm.

Algorithm 1 Optimistic Sample Path (Osp)

```
1:  Input: confidence $\delta$, horizon $T$, (upper bound on) mixing time $t_{\mathrm{mix}}$
    //Initialization:
2:  Set $t := 1$ and let the sequence $\mathcal{O}$ of observations $(s, a, r, s')$ be empty.
    // Compute sample paths for policies
3:  for phases $k = 1, 2, \dots$ do
4:      for each policy $\pi : \mathcal{S} \to \mathcal{A}$ do
5:          Use Algorithm 2 to construct a non-extendible sample path $\mathcal{P}_\pi$ from $\mathcal{O}$.
6:          Let $\hat{\rho}_\pi := \frac{1}{|\mathcal{P}_\pi|} \sum_{(s,\pi(s),r,s') \in \mathcal{P}_\pi} r$, and set $\tilde{\rho}_\pi := \hat{\rho}_\pi + \sqrt{\frac{8 t_{\mathrm{mix}} \log \frac{16tT}{\delta}}{|\mathcal{P}_\pi|}}$.
7:      end for
        // Choose optimistic policy
8:      Choose $\pi_k := \arg\max_\pi \tilde{\rho}_\pi$ and set $n_{<k} := |\mathcal{P}_{\pi_k}|$.
        // Execute optimistic policy $\pi_k$
9:      for $\tau = 1, \dots, n_k := \max\left\{ n_{<k}, \sqrt{\frac{T}{SA}} \right\}$ do
10:         Choose action $a_t = \pi_k(s_t)$, obtain reward $r_t$, and observe $s_{t+1}$.
            Set $t := t + 1$ and append the observation $(s_t, a_t, r_t, s_{t+1})$ to $\mathcal{O}$.
11:     end for
12: end for
```

Algorithm 2 Path Construction

```
Input: Observation sequence $\mathcal{O}$, policy $\pi$, initial state $s_1$
Set $t = 1$ and let path $\mathcal{P}_\pi$ be empty.
while $\mathcal{O}$ contains an unused observation of the form $(s_t, \pi(s_t), \cdot, \cdot)$ do
    Choose the first unused occurrence $o_t := (s_t, \pi(s_t), r, s)$ of such an observation.
    Append $o_t$ to $\mathcal{P}_\pi$.
    Mark $o_t$ in $\mathcal{O}$ as used.
    Set $s_{t+1} := s$ and $t := t + 1$.
end while
Mark all observations in $\mathcal{O}$ as unused.
Output: sample path $\mathcal{P}_\pi$
```

is no unused observation $(s_{n+1}, \pi(s_{n+1}), r, s)$ in $\mathcal{O}$ that could be used to extend the path by appending the observation. In the following, we say that such a path is *non-extendible.* Algorithm 2 provides an algorithm for constructing a non-extendible path from a given set of observations. Alternative constructions could be used for obtaining non-extendible paths as well.

For each possible policy π the algorithm computes an estimate of the average reward ρ_π from the sample path $\mathcal{P}_\pi$ and considers an optimistic upper confidence value $\tilde{\rho}_\pi$ (cf. line 6 of Osp) using the concentration results of Section 3. The policy with the maximal $\tilde{\rho}_\pi$ is chosen for use in phase k. The length n_k of phase k, in which the chosen policy π_k is executed, depends on the length $n_{<k} := |\mathcal{P}_{\pi_k}|$ of the sample path $\mathcal{P}_{\pi_k}$. That is, π_k is usually played for $n_{<k}$ steps, but at least for $\sqrt{\frac{T}{SA}}$ steps (cf. line 9).

Note that at the beginning, all sample paths are empty in which case we set the confidence intervals to be ∞, and the algorithm chooses an arbitrary policy. The initial state of the sample paths can be chosen to be the current state, but this is not necessary. Note that by the Markov property the outcomes of all samples are independent of each other. The way Algorithm 2 extracts observations from $\mathcal{O}$ is analogous to when having access to a generative sampling model as e.g. assumed in work on sample complexity bounds (e.g. Azar et al., 2013b). In both settings the algorithm can request a sample for a particular state-action pair (s, a). The only difference is that in our case at some point there are no suitable samples available anymore, when the construction of the sample path is terminated.

As the goal of this paper is to demonstrate an easy way to obtain optimal regret bounds, we do not elaborate in detail on computational aspects of the algorithm. A brief discussion is however in order. First, note that it is obviously not necessary to construct sample paths from scratch in each phase. It is sufficient to extend the path for each policy with new and previously unused samples. Further, while the algorithm as given is exponential in nature (as it loops over all A^S policies), it may be possible to find the optimistic policy by some kind of optimistic policy gradient algorithm (Lazaric, 2018). We note that policies in ergodic MDPs exhibit a particular structure (see Section 3 of Ortner, 2007, that could be exploited by such an algorithm). However, at the moment this is not more than an idea for future research and the details of such an algorithm are yet to be developed.

5. Regret Analysis

The following theorem is the main result of this note.

Theorem 1. *In uniformly ergodic MDPs, with probability at least* $1 - \delta$ *the regret of* Osp *is bounded by*

$$R_T \;\leq\; 10 \log(\tfrac{16T^2}{\delta}) \sqrt{t_{\text{mix}} SAT},$$

provided that $T \geq S^3 A \left(\frac{152 t_{\text{mix}} \log \frac{16T^2}{\delta}}{\mu_{\min}^2} \right)^2$, *where* $\mu_{\min} := \min_{\pi, s: \mu_\pi(s) > 0} \mu_\pi(s)$.

The improvement with respect to previously known bounds can be achieved due to the fact that the confidence intervals for our algorithm are computed on the policy level and not on the level of rewards and transition probabilities as for UCRL (Jaksch et al., 2010). This avoids the problem of having rectangular confidence intervals that lead to an additional factor of $\sqrt{S}$ in the regret bounds for UCRL, cf. the discussion of Osband and Roy (2017).

To keep the exposition simple, we have chosen confidence intervals which give a high probability bound for each horizon T. It is easy to adapt the confidence intervals to gain a high probability bound that holds for all T simultaneously (cf. Jaksch et al., 2010).

The mixing time parameter in our bounds is different from the transition parameters in the regret bounds of Jaksch et al. (2010) or the bias span used by Bartlett and Tewari (2009) and Fruit et al. (2018b). We note however that for *reversible* Markov chains, t_{mix} is linearly bounded in the diameter (i.e., the hitting time) of the chain, cf. Section 10.5 of Levin, Peres, and Wilmer (2009). It follows from the lower bounds on the regret of

Jaksch et al. (2010) that the upper bound of Theorem 1 is best possible with respect to the appearing parameters. Mixing times have also been used for sample complexity bounds in reinforcement learning (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002), however not for a fixed constant $\frac{1}{4}$ as in our case but with respect to the required accuracy. It would be desirable to replace the upper bound t_{mix} on all mixing times by the mixing time of the optimal policy as done by Azar et al. (2013a). However, the technique of Azar et al. (2013a) comes at the price of an additional dependence on the number of considered policies, which in our case obviously would deteriorate the bound.

The parameter T can be guessed using a standard doubling scheme giving the same regret bounds with a slightly larger constant. Guessing t_{mix} is more costly. For example, using $\log T$ as a guess for t_{mix}, the additional regret is an additive constant exponential in t_{mix}. We note however, that it is an open problem whether it is possible to get regret bounds depending on a different parameter than the diameter (such as the bias span) without having a larger bound on the quantity, cf. the discussion in Appendix A of Fruit, Pirotta, and Lazaric (2018a).

5.1 Proof of Theorem 1

Recall that π_k is the policy applied in phase k for n_k steps. The respective optimistic estimate $\tilde{\rho}_{\pi_k}$ has been computed from a sample path of length $n_{<k}$.

5.1.1 Estimates $\tilde{\rho}_\pi$ are optimistic

We start showing that the values $\tilde{\rho}_\pi$ computed by our algorithm from the sample paths of any policy π are indeed optimistic. This holds in particular for the employed policies π_k.

Lemma 8. *With probability at least $1-\frac{\delta}{4}$, for all phases k it holds that*

$$\tilde{\rho}_{\pi_k} \;\geq\; \rho^* \;\geq\; \rho_{\pi_k}.$$

Proof. Let us first consider an arbitrary fixed policy π and some time step t. Using (column) vector notation $\boldsymbol{\mu} := (\mu_\pi(s))_s$ and $\mathbf{r} := (r(s,\pi(s))_s$ for the stationary distribution and the reward function under π, and writing $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{r}}$ for the respective estimated values at step t, we have

$$\begin{aligned} \rho_\pi - \hat{\rho}_\pi &= \boldsymbol{\mu}^\top \mathbf{r} - \hat{\boldsymbol{\mu}}^\top \hat{\mathbf{r}} \\ &= (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \mathbf{r} + \hat{\boldsymbol{\mu}}^\top (\mathbf{r} - \hat{\mathbf{r}}). \end{aligned} \tag{2}$$

Let n be the length of the sample path $\mathcal{P}_\pi$ from which the estimates are computed. Then the first term of (2) can be bounded by Corollary 3 as

$$|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \mathbf{r}| \leq \sqrt{\frac{9 t_{\mathrm{mix}} \log \frac{16tT}{\delta}}{2n}} \tag{3}$$

with probability at least $1-\frac{\delta}{8T}$ (using a union bound over all t possible values for n). The second term of (2) can be written as

$$|\hat{\boldsymbol{\mu}}^\top (\mathbf{r} - \hat{\mathbf{r}})| = \tfrac{1}{n} \cdot \Big| \sum_{(s,\pi(s),r,s') \in \mathcal{P}_\pi} (r(s,\pi(s)) - r) \Big|.$$

Since the sum is a martingale difference sequence, we obtain by Azuma-Hoeffding inequality (cf. Lemma A.7 of Cesa-Bianchi & Lugosi, 2006) and another union bound that with probability $1 - \frac{\delta}{8T}$

$$|\hat{\boldsymbol{\mu}}^\top(\mathbf{r} - \hat{\mathbf{r}})| \leq \sqrt{\frac{\log \frac{16tT}{\delta}}{2n}}. \tag{4}$$

Summarizing, we get from (2)–(4) that for any policy π the estimate $\hat{\rho}_\pi$ computed at time step t satisfies with probability at least $1 - \frac{\delta}{4T}$

$$|\rho_\pi - \hat{\rho}_\pi| \;\leq\; \sqrt{\frac{8t_{\mathrm{mix}} \log \frac{16tT}{\delta}}{n}}. \tag{5}$$

This holds in particular for an optimal policy π^*, so that by a union bound over all time steps t we have that with probability at least $1-\frac{\delta}{4}$ at each time step it holds that $\tilde{\rho}_{\pi^*} \;\geq\; \rho_{\pi^*}$. Then by definition of the algorithm (in particular line 8) and optimality of π^* it follows that with probability at least $1 - \frac{\delta}{4}$,

$$\tilde{\rho}_{\pi_k} \;\geq\; \tilde{\rho}_{\pi^*} \;\geq\; \rho_{\pi^*} \;\geq\; \rho_{\pi_k} \tag{6}$$

for all episodes k. □

5.1.2 Splitting regret into phases

In the following, let $\hat{\rho}_{\pi_k}^{<k}$ be the empirical average reward of π_k computed from the sample path *before* episode k, and $\hat{\rho}_{\pi_k}^{(k)}$ the empirical average reward of π_k *in* episode k. We write the regret as a sum over the regret in the single phases as

$$R_T \;=\; T\rho^* - \sum_{t=1}^{T} r_t \;=\; \sum_k n_k(\rho^* - \hat{\rho}_{\pi_k}^k). \tag{7}$$

Now we can distinguish between two kinds of phases: The length of most phases is $n_k = n_{<k}$. However, there are also a few phases where the sample path for the chosen policy π_k is shorter than $\sqrt{\frac{T}{SA}}$, when the length is $n_k = \sqrt{\frac{T}{SA}} > n_{<k}$. Let $\mathcal{K}^- := \{k \,|\, n_k > n_{<k}\}$ be the set of these latter phases and set $K^- := |\mathcal{K}^-|$. The regret for each phase in $\mathcal{K}^-$ is simply bounded by $\sqrt{\frac{T}{SA}}$, so that we obtain from (7) that

$$R_T \;\leq\; K^- \sqrt{\frac{T}{SA}} + \sum_{k \notin \mathcal{K}^-} n_k(\rho^* - \hat{\rho}_{\pi_k}^k). \tag{8}$$

For episodes $k \notin \mathcal{K}^-$ we note that[6] $n_k \leq n_{<k}$. Hence, by Lemma 8 and the definition of $\tilde{\rho}_{\pi_k}$ the respective regret with probability at least $1 - \frac{\delta}{4}$ is bounded by

$$\begin{aligned}
\sum_{k \notin \mathcal{K}^-} n_k(\rho^* - \hat{\rho}_{\pi_k}^k) \;&\leq\; \sum_{k \notin \mathcal{K}^-} n_k(\tilde{\rho}_{\pi_k} - \hat{\rho}_{\pi_k}^k) \\
&\leq\; \sum_{k \notin \mathcal{K}^-} \sqrt{8 n_k t_{\mathrm{mix}} \log \tfrac{16T^2}{\delta}} + \sum_{k \notin \mathcal{K}^-} n_k(\hat{\rho}_{\pi_k}^{<k} - \hat{\rho}_{\pi_k}^k).
\end{aligned} \tag{9}$$

[6] The final phase may be shorter than $n_{<k}$.

The difference in the empirical means in the last term of (9) can be bounded by the confidence interval in (5) and a union bound over all time steps. Combining this with (8) and (9) yields that with probability at least $1 - \frac{3\delta}{4}$,

$$R_T \;\leq\; K^{-}\sqrt{\frac{T}{SA}} + 3 \sum_{k \notin \mathcal{K}^{-}} \sqrt{8 n_k t_{\mathrm{mix}} \log \tfrac{16T^2}{\delta}}. \tag{10}$$

It remains to bound the number of phases (not) in $\mathcal{K}^{-}$. A bound on K^{-} obviously gives a bound on the first term in (10), while a bound on the number K^{+} of phases not in $\mathcal{K}^{-}$ allows to bound the second term, as by Jensen's inequality we have due to $\sum_{k \notin \mathcal{K}^{-}} n_k \leq T$ that

$$\sum_{k \notin \mathcal{K}^{-}} \sqrt{8 n_k t_{\mathrm{mix}} \log \tfrac{16T^2}{\delta}} \leq \sqrt{8 T K^{+} t_{\mathrm{mix}} \log \tfrac{16T^2}{\delta}}. \tag{11}$$

5.1.3 Bounding the number of phases

The following lemma gives a bound on the total number of phases that can be used as a bound on K^{-} and K^{+} to conclude the proof of Theorem 1.

Lemma 9. *With probability at least $1 - \frac{\delta}{4}$, the number of phases up to step T is bounded by*

$$K \leq SA \log_{\frac{4}{3}} \left(\tfrac{T}{SA}\right),$$

provided that $T \geq S^3 A \left(\frac{152 t_{\mathrm{mix}} \log \frac{16T^2}{\delta}}{\mu_{\min}^2}\right)^2$, where $\mu_{\min} := \min_{\pi, s: \mu_\pi(s) > 0} \mu_\pi(s)$.

Proof. Let $n_{<k}(s,a)$ be the number of visits to (s,a) before phase k. Note that the sample path for each policy π in general will not use all samples of $(s, \pi(s))$, so that we also introduce the notation $n^{\pi}_{<k}(s)$ for the number of samples of $(s, \pi(s))$ used in the sample path of π computed before phase k. Recall that by definition of the algorithm, sample paths are non-extendible, so that for each π there is a state s^{-} for which all samples are used,[7] that is, $n_{<k}(s^{-}, \pi(s^{-})) = n^{\pi}_{<k}(s^{-})$. We write $\hat{\mu}_{<k}$ and $\hat{\mu}_k$ for the empirical distributions of the policy π_k in the sample path for phase k and in phase k, respectively.

Note that for each phase k we have

$$d_{TV}(\mu_{\pi_k}, \hat{\mu}_{<k}) \;\;\leq\;\; \sqrt{\frac{38 S t_{\mathrm{mix}} \log \frac{16T^2}{\delta}}{n_{<k}}} \quad \text{and} \tag{12}$$

$$d_{TV}(\mu_{\pi_k}, \hat{\mu}_{k}) \;\;\leq\;\; \sqrt{\frac{38 S t_{\mathrm{mix}} \log \frac{16T^2}{\delta}}{n_{k}}}, \tag{13}$$

each with probability at least $1 - \frac{\delta}{8T}$ by Corollary 7 and a union bound over all possible values of n_k and $n_{<k}$, respectively.[8] By another union bound over the at most T phases,

[7] In particular, this holds for the last state of the sample path.

[8] We note that instead of Corollary 7 it would also be sufficient to use Corollary 2 to derive a result similar to Lemma 9 where the sufficient size of T would have a smaller constant but an additional S in the log-term due to a necessary union bound over all states.

(12) and (13) hold for all phases k with probability at least $1-\frac{\delta}{4}$. In the following, we assume that the confidence intervals of (12) and (13) hold, so that all following results hold with probability $1-\frac{\delta}{4}$.

Each phase k has length at least $n_k \geq \sqrt{\frac{T}{SA}}$. Consequently, if $T \geq S^3 A \Big(\frac{152 t_{\mathrm{mix}} \log \frac{16T^2}{\delta}}{\mu_{\min}^2}\Big)^2$, then it is guaranteed by (13) that in each phase k it holds that

$$d_{TV}(\mu_{\pi_k}, \hat{\mu}_k) \leq \sqrt{\frac{38 S t_{\mathrm{mix}} \log \frac{16T^2}{\delta}}{n_k}} \leq \frac{\mu_{\min}}{2} \leq \frac{\mu_{\pi_k}(s)}{2}, \tag{14}$$

and therefore for each state s

$$\frac{\mu_{\pi_k}(s)}{2} \leq \hat{\mu}_k(s). \tag{15}$$

Now consider an arbitrary phase k and let s^- be the state for which $n_{<k}(s^-, \pi_k(s^-)) = n_{<k}^{\pi_k}(s^-)$, so that in particular $\hat{\mu}_{<k}(s^-)\, n_{<k} = n_{<k}^{\pi_k}(s^-)$. We are going to show that the number of visits to $(s^-, \pi_k(s^-))$ is increased by (at least) a factor $\frac{4}{3}$ in phase k. By (12)–(15) and using that[9] $n_k \geq n_{<k}$ we have

$$\begin{aligned} n_{<k}(s^-, \pi_k(s^-)) &= \hat{\mu}_{<k}(s^-)\, n_{<k} \\ &\leq \mu_{\pi_k}(s^-)\, n_{<k} + \sqrt{38 n_{<k} S t_{\mathrm{mix}} \log \tfrac{16T^2}{\delta}} \\ &\leq 2\hat{\mu}_k(s^-)\, n_{<k} + \sqrt{38 n_{<k} S t_{\mathrm{mix}} \log \tfrac{16T^2}{\delta}} \\ &\leq 2\hat{\mu}_k(s^-)\, n_k + \sqrt{38 n_k S t_{\mathrm{mix}} \log \tfrac{16T^2}{\delta}} \\ &\leq 2\hat{\mu}_k(s^-)\, n_k + \frac{\mu_{\pi_k}(s^-)}{2} n_k \\ &\leq 3\hat{\mu}_k(s^-)\, n_k, \end{aligned}$$

so that abbreviating $a^- := \pi_k(s^-)$

$$n_{<k+1}(s^-, a^-) = n_{<k}(s^-, a^-) + \hat{\mu}_k(s^-)\, n_k \geq \tfrac{4}{3}\, n_{<k}(s^-, a^-).$$

Hence in each phase there is a state-action pair for which the number of visits is increased by a factor of $\frac{4}{3}$. This can be used to show that the total number of phases K within T steps is upper bounded as

$$K \leq SA \log_{\frac{4}{3}} \left(\tfrac{T}{SA}\right). \tag{16}$$

The proof of (16) can be rewritten from Proposition 3 of Ortner (2010), with the only difference that the factor 2 is replaced by $\frac{4}{3}$. □

Finally, combining (8), (11), and Lemma 9, using that $K^-, K^+ \leq K$, we obtain that with probability at least $1-\delta$

$$\begin{aligned} R_T &\leq K^- \sqrt{\frac{T}{SA}} + 3\sqrt{8TK^+ t_{\mathrm{mix}} \log \tfrac{16T^2}{\delta}} \\ &\leq \sqrt{SAT} \log_{\frac{4}{3}} \left(\tfrac{T}{SA}\right) + 3\sqrt{8 t_{\mathrm{mix}} SAT \log_{\frac{4}{3}} \left(\tfrac{T}{SA}\right) \log\left(\tfrac{16T^2}{\delta}\right)}, \end{aligned}$$

which completes the proof of the theorem. □

[9] As already mentioned, this may not hold for the last episode, which is however not relevant here.

6. Discussion and Conclusion

While we were able to close the gap between lower and upper bound on the regret for uniformly ergodic MDPs, there are still quite a few open questions. First of all, the concentration results we use are only available for uniformly ergodic Markov chains, so a generalization of our approach to more general communicating MDPs seems not easy. An improvement over the parameter t_{mix} may be possible by considering more specific concentration results for Markov reward processes. These might depend not so much on the mixing time than the bias span (Fruit et al., 2018b). However, even if one achieves such bounds, the resulting regret bounds would depend on the maximum bias span over all policies. Obtaining a dependence on the bias span of the optimal policy instead seems not easily possible. Finally, another topic for future research is to develop an optimistic policy gradient algorithm that computes the optimistic policy more efficiently than by an iteration over all policies.

Acknowledgments

This work has been supported by the Austrian Science Fund (FWF): I 3437-N33 in the framework of the CHIST-ERA ERA-NET (DELTA project). The author would like to thank Ronan Fruit, Alessandro Lazaric, and Matteo Pirotta for discussion as well as Sadegh Talebi, Chen-Yu Wei, and three anonymous reviewers for pointing out errors in a previous version of the paper.

References

Agrawal, S., & Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems 30, NIPS 2017*, pp. 1184–1194.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, *47*, 235–256.

Azar, M. G., Lazaric, A., & Brunskill, E. (2013a). Regret bounds for reinforcement learning with policy advice. In *Machine Learning and Knowledge Discovery in Databases – European Conference, ECML PKDD 2013, Proceedings, Part I*, pp. 97–112.

Azar, M. G., Munos, R., & Kappen, H. J. (2013b). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, *91*(3), 325–349.

Azar, M. G., Osband, I., & Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, Vol. 70 of *JMLR Workshop and Conference Proceedings*, pp. 263–272.

Bartlett, P. L., & Tewari, A. (2009). REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, pp. 25–42.

Brafman, R. I., & Tennenholtz, M. (2002). R-MAX – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, *3*, 213–231.

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

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games.* Cambridge University Press.

Fruit, R., Pirotta, M., & Lazaric, A. (2018a). Near optimal exploration-exploitation in non-communicating markov decision processes. In *Advances in Neural Information Processing Systems 31, NeurIPS 2018*, pp. 2998–3008.

Fruit, R., Pirotta, M., Lazaric, A., & Ortner, R. (2018b). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, Vol. 80 of *JMLR Workshop and Conference Proceedings*, pp. 1573–1581.

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

Kearns, M. J., & Singh, S. P. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, *49*, 209–232.

Lattimore, T., & Szepesvári, C. (2019). *Bandit Algorithms.* Draft from 27th June 2019, available at http://banditalgs.com/.

Lazaric, A. (2018). personal communication, August and October 2018.

Levin, D. A., Peres, Y., & Wilmer, E. L. (2009). *Markov chains and mixing times.* American Mathematical Society.

Ok, J., Proutière, A., & Tranos, D. (2018). Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems 31, NeurIPS 2018*, pp. 8888–8896.

Ortner, R. (2007). Linear dependence of stationary distributions in ergodic Markov decision processes. *Operation Research Letters*, *35*, 619–626.

Ortner, R. (2010). Online regret bounds for Markov decision processes with deterministic transitions. *Theoretical Computer Science*, *411*(29–30), 2684–2695.

Osband, I., & Roy, B. V. (2017). Why is posterior sampling better than optimism for reinforcement learning?. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, Vol. 70 of *JMLR Workshop and Conference Proceedings*, pp. 2701–2710.

Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, *20*(79), 1–32.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., New York, NY, USA.

Sidford, A., Wang, M., Wu, X., Yang, L., & Ye, Y. (2018). Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems 31, NeurIPS 2018*, pp. 5192–5202.

Szita, I., & Szepesvári, C. (2010). Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning, ICML 2010*, pp. 1031–1038.

Wang, M. (2017). Primal-dual π learning: Sample complexity and sublinear run time for ergodic Markov decision problems. *CoRR*, *abs/1710.06100*.