Journal of Machine Learning Research volume (year) pages Submitted submitted; Published published

Variance-Aware Regret Bounds for Undiscounted Reinforcement Learning in MDPs

Mohammad Sadegh Talebi* MSTMS@KTH.SE
KTH Royal Institute of Technology, Stockholm, Sweden

Odalric-Ambrym Maillard† ODALRIC.MAILLARD@INRIA.FR
INRIA Lille – Nord Europe, Villeneuve d'Ascq, France

Editor: Mehryar Mohri and Karthik Sridharan

Abstract

The problem of reinforcement learning in an unknown and discrete Markov Decision Process (MDP) under the average-reward criterion is considered, when the learner interacts with the system in a single stream of observations, starting from an initial state without any reset. We revisit the minimax lower bound for that problem by making appear the local variance of the bias function in place of the diameter of the MDP. Furthermore, we provide a novel analysis of the KL-UCRL algorithm establishing a high-probability regret bound scaling as $\widetilde{\mathcal{O}}\Big(\sqrt{S\sum_{s,a}\mathbf{V}^{\star}_{s,a}T}\Big)$ for this algorithm for ergodic MDPs, where S denotes the number of states and where $\mathbf{V}^{\star}_{s,a}$ is the variance of the bias function with respect to the next-state distribution following action a in state s. The resulting bound improves upon the best previously known regret bound $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ for that algorithm, where A and D respectively denote the maximum number of actions (per state) and the diameter of MDP. We finally compare the leading terms of the two bounds in some benchmark MDPs indicating that the derived bound can provide an order of magnitude improvement in some cases. Our analysis leverages novel variations of the transportation lemma combined with Kullback-Leibler concentration inequalities, that we believe to be of independent interest.

Keywords: Undiscounted Reinforcement Learning, Markov Decision Processes, Concentration Inequalities, Regret Minimization, Bellman Optimality

1. Introduction

In this paper, we consider Reinforcement Learning (RL) in an unknown and discrete Markov Decision Process (MDP) under the average-reward criterion, when the learner interacts with the system in a single stream of observations, starting from an initial state without any reset. More formally, let $M = (\mathcal{S}, \mathcal{A}, \nu, P)$ denote an MDP where $\mathcal{S}$ is a finite set of states and $\mathcal{A}$ is a finite set of actions available at any state, with respective cardinalities S and A. The reward function and the transition kernel is respectively denoted by ν and P. The game goes as follows: the learner starts in some state $s_1 \in \mathcal{S}$ at time $t = 1$. At each time step $t \in \mathbb{N}$, the learner chooses one action $a \in \mathcal{A}$ in her current state $s \in \mathcal{S}$ based on her past decisions and observations. When executing action a in state s, the learner receives a random reward r drawn independently from distribution $\nu(s, a)$ with support $[0, 1]$ and mean $\mu(s, a)$. The state then transits to a next state $s' \in \mathcal{S}$ sampled with probability $p(s'|s, a)$, and a new

∗. The authors contributed equally.
†. The authors contributed equally.

decision step begins. As the transition probabilities and reward functions are unknown, the learner has to learn them by trying different actions and recording the realized rewards and state transitions. We refer to standard textbooks (Sutton and Barto, 1998; Puterman, 2014) for background material on RL and MDPs.

The performance of the learner can be quantified through the notion of regret, which compares the reward collected by the learner (or the algorithm) to that obtained by an oracle always following an optimal policy, where a policy is a mapping from states to actions. More formally, let $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ denote a possibly stochastic policy. We further introduce the notation $p(s'|s,\pi(s)) = \mathbb{E}_{Z\sim\pi(s)}[p(s'|s,Z)]$, and $P_\pi f$ to denote the function $s \mapsto \sum_{s'\in\mathcal{S}} p(s'|s,\pi(s))f(s')$. Likewise, let $\mu_\pi(s) = \mathbb{E}_{Z\sim\pi(s)}[\mu(s,Z)]$ denote the mean reward after choosing action $\pi(s)$ in step s.

Definition 1 (Expected cumulative reward) *The expected cumulative reward of policy π when run for T steps from initial state s_1 is defined as*

$$R_{\pi,T}(s_1) = \mathbb{E}\bigg[\sum_{t=1}^{T} r(s_t,a_t)\bigg] = \mu_\pi(s_1) + (P_\pi\mu_\pi)(s_1) + \cdots = \sum_{t=1}^{T}(P_\pi^{t-1}\mu_\pi)(s_1)\,.$$

where $a_t \sim \pi(s_t)$, $s_{t+1}\sim p(\cdot|s_t,a_t)$, and finally $r(s,a)\sim\nu(s,a)$.

Definition 2 (Average gain and bias) *Let us introduce the average transition operator $\overline{P}_\pi = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} P_\pi^{t-1}$. The average gain g_π and bias function b_π are defined by*

$$g_\pi(s_1) = \lim_{T\to\infty}\frac{1}{T}R_{\pi,T}(s_1) = (\overline{P}_\pi\mu_\pi)(s_1)\,, \qquad b_\pi(s) = \sum_{t=1}^{\infty}\Big((P_\pi^{t-1} - \overline{P}_\pi)\mu_\pi\Big)(s)\,.$$

The previous definition requires some mild assumption on the MDP for the limits to makes sense. It is shown (see, e.g., (Puterman, 2014)) that the average gain achieved by executing a stationary policy π in a communicating MDP M is well-defined and further does not depend on the initial state, i.e., $g_\pi(s_1) = g_\pi$. For this reason, we restrict our attention to such MDPs in the rest of this paper. Furthermore, let $\star$ denote an optimal policy, that is[1] $g_\star = \max_\pi g_\pi$.

Definition 3 (Regret) *We define the regret of any learning algorithm $\mathbb{A}$ after T steps as*

$$\mathrm{Regret}_{\mathbb{A},T} := \sum_{t=1}^{T} r(s_t^\star,\star(s_t^\star)) - \sum_{t=1}^{T} r(s_t,a_t) \quad \textit{where } a_t = \mathbb{A}(s_t,(\{s_{t'},a_{t'},r_{t'}\})_{t'<t})\,,$$

and $s_{t+1}^\star\sim p(\cdot|s_t^\star,\star(s_t^\star))$ with $s_1^\star = s_1$ is a sequence generated by the optimal strategy.

By an application of Azuma-Hoeffding's inequality for bounded random martingales, it is immediate to show that with probability higher than $1-\delta$,

$$\begin{aligned}\mathrm{Regret}_{\mathbb{A},T} &\leqslant \sum_{t=1}^{T}\Big(P_\star^{t-1}\mu_\star - P_{a_t}^{t-1}\mu_{a_t}\Big) + \sqrt{2T\log(2/\delta)}\\ &= \sum_{t=1}^{T}(P_\star^{t-1} - \overline{P}_\star)\mu_\star + \bigg[Tg_\star - \sum_{t=1}^{T} P_{a_t}^{t-1}\mu_{a_t}\bigg] + \sqrt{2T\log(2/\delta)}\,.\end{aligned}$$

Thus, following (Jaksch et al., 2010), it makes sense to focus on the control of the middle term in brackets only. This leads us to consider the following notion of regret, which we may call *effective regret*:

$$\mathfrak{R}_{\mathbb{A},T} := Tg^\star - \sum_{t=1}^{T} r(s_t,a_t)\,.$$

1. The maximum is reached since there are only finitely many deterministic policies.

To date, several algorithms have been proposed in order to minimize the regret based on the *optimism in the face of uncertainty* principle, coming from the literature on stochastic multi-armed bandits (see (Robbins, 1952)). Algorithms designed based on this principle typically maintain confidence bounds on the unknown reward and transition distributions, and choose an optimistic model that leads to the highest average long-term reward. One of the first algorithms based on this principle for MDPs is due to (Burnetas and Katehakis, 1997), which is shown to be asymptotically optimal. Their proposed algorithm uses the Kullback-Leibler (KL) divergence to define confidence bounds for transition probabilities. Subsequent studies by (Tewari and Bartlett, 2008), (Auer and Ortner, 2007), (Jaksch et al., 2010), and (Bartlett and Tewari, 2009) propose algorithms that maintain confidence bounds on transition kernel defined by L_1 or total variation norm. The use of L_1 norm, instead of KL-divergence, allows one to describe the uncertainty of the transition kernel by a polytope, which in turn brings computational advantages and ease in the regret analysis. On the other hand, such polytopic models are typically known to provide poor representations of underlying uncertainties; we refer to the literature on the robust control of MDPs with uncertain transition kernels, e.g., (Nilim and El Ghaoui, 2005), and more appropriately to (Filippi et al., 2010). Indeed, as argued in (Filippi et al., 2010), optimistic models designed by L_1 norm suffer from two shortcomings: (i) the L_1 optimistic model could lead to inconsistent models by assigning a zero mass to an already observed element, and (ii) due to polytopic shape of L_1-induced confidence bounds, the maximizer of a linear optimization over L_1 ball could significantly vary for a small change in the value function, thus resulting in sub-optimal exploration (we refer to the discussion and illustrations on pages 120–121 in (Filippi et al., 2010)).

Both of these shortcomings are avoided by making use of the KL-divergence and properties of the corresponding KL-ball. In (Filippi et al., 2010), the authors introduce the `KL-UCRL` algorithm that modifies the `UCRL2` algorithm of (Jaksch et al., 2010) by replacing L_1 norms with KL divergences in order to define the confidence bound on transition probabilities. Further, they provide an efficient way to carry out linear optimization over the KL-ball, which is necessary in each iteration of the Extended Value Iteration. Despite these favorable properties and the strictly superior performance in numerical experiments (even for very short time horizons), the best known regret bound for `KL-UCRL` matches that of `UCRL2`. Hence, from a theoretical perspective, the potential gain of use of KL-divergence to define confidence bounds for transition function has remained largely unexplored. The goal of this paper is to investigate this gap.

Main contributions. In this paper we provide a new regret bound for `KL-UCRL` scaling as $\widetilde{\mathcal{O}}\big(\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}T} + D\sqrt{T}\big)$ for ergodic MDPs with S states, A actions, and diameter D. Here, $\mathbf{V}^\star_{s,a} := \mathbb{V}_{p(\cdot|s,a)}(b^\star)$ denotes the variance of the optimal bias function $b^\star$ of the true (unknown) MDP with respect to next state distribution under state-action (s,a). This bound improves over the best previous bound of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ for `KL-UCRL` as $\sqrt{\mathbf{V}^\star_{s,a}} \leqslant D$. Interestingly, in several examples $\sqrt{\mathbf{V}^\star_{s,a}} \ll D$ and actually $\sqrt{\mathbf{V}^\star_{s,a}}$ is comparable to $\sqrt{D}$. Our numerical experiments on typical MDPs further confirm that $\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}}$ could be orders of magnitude smaller than $DS\sqrt{A}$. To prove this result, we provide novel transportation concentration inequalities inspired by the transportation method that relate the so-called transportation cost under two discrete probability measures to the KL-divergence

between the two measures and the associated variances. To the best of our knowledge, these inequalities are new and of independent interest. To complete our result, we provide a new minimax regret lower bound of order $\Omega(\sqrt{SA\mathbf{V}_{\max}T})$, where $\mathbf{V}_{\max} := \max_{s,a} \mathbf{V}^{\star}_{s,a}$. In view of the new minimax lower bound, the reported regret bound for `KL-Ucrl` can be improved by only a factor $\sqrt{S}$.

Related work. RL in unknown MDPs under average-reward criterion dates back to the seminal papers by (Graves and Lai, 1997), and (Burnetas and Katehakis, 1997), followed by (Tewari and Bartlett, 2008). Among these studies, for the case of ergodic MDPs, (Burnetas and Katehakis, 1997) derive an asymptotic MDP-dependent lower bound on the regret and provide an asymptotically optimal algorithm. Algorithms with finite-time regret guarantees and for wider class of MDPs are presented by (Auer and Ortner, 2007), (Jaksch et al., 2010; Auer et al., 2009), (Bartlett and Tewari, 2009), (Filippi et al., 2010), and (Maillard et al., 2014).

`Ucrl2` and `KL-Ucrl` achieve a $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ regret bound with high probability in communicating MDPs, for any unknown time horizon. `Regal` obtains a $\widetilde{\mathcal{O}}(D'S\sqrt{AT})$ regret with high probability in the larger class of weakly communicating MDPs, provided that we know an upper bound D' on the span of the bias function. It is however still an open problem to incorporate this knowledge into an implementable algorithm. The `TSDE` algorithm by Ouyang et al. (Ouyang et al., 2017) achieves a regret growing as $\widetilde{\mathcal{O}}(D'S\sqrt{AT})$ for the class of weakly communicating MDPs, where D' is a given bound on the span of the bias function. In a recent study, (Agrawal and Jia, 2017) propose an algorithm based on posterior sampling for the class of communicating MDPs. Under the assumption of known reward function and *known time horizon*, their algorithm enjoys a regret bound scaling as $\widetilde{\mathcal{O}}(D\sqrt{SAT})$, which constitutes the best known regret upper bound for learning in communicating MDPs and has a tight dependencies on S and A.

We finally mention that some studies consider regret minimization in MDPs in the *episodic* setting, where the length of each episode is fixed and known; see, e.g., (Osband et al., 2013), (Gheshlaghi Azar et al., 2017), and (Dann et al., 2017). Although RL in the episodic setting bears some similarities to the average-reward setting, the techniques developed in these paper strongly rely on the fixed length of the episode, which is assumed to be small, and do not directly carry over to the case of undiscounted RL considered here.

2. Background Material and The `KL-Ucrl` Algorithm

In this section, we recall some basic material on undiscounted MDPs and then detail the `KL-Ucrl` algorithm.

Lemma 4 (Bias and Gain) *The gain and bias function satisfy the following relations*

$$\begin{aligned} &(\textit{Bellman equation}) && b_\pi + g_\pi &&= \mu_\pi + P_\pi b_\pi \\ &(\textit{Fundamental matrix}) && b_\pi &&= [I - P_\pi + \overline{P_\pi}]^{-1}[I - \overline{P}_\pi]\mu_\pi \,. \end{aligned}$$

This result is an easy consequence of the fact that $\overline{P}_\pi$ (see Definition 2) satisfies $\overline{P}_\pi P_\pi = P_\pi \overline{P}_\pi = \overline{P}_\pi \overline{P}_\pi = \overline{P}_\pi$ (see (Puterman, 2014) as well as Appendix E for details).

According to the standard terminology, we say a policy is $b_\star$-improving if it satisfies $\pi(s) = \mathrm{argmax}_{a\in\mathcal{A}}\, \mu(s,a) + (P_a b_\star)(s)$. Applying the theory of MDPs (see, e.g., (Puterman,

2014)), it can be shown that any $b_\star$-improving policy is optimal and thus that we can choose $\star$ to satisfy[2] the following fundamental identity[3]

$$\text{(Bellman optimality equation)} \quad \forall s \in \mathcal{S},\, b_\star(s) + g_\star = \max_{a \in \mathcal{A}} \Big(\mu(s,a) + \sum_{y \in \mathcal{S}} p(y|s,a) b_\star(y) \Big) \,.$$

We now recall the definition of diameter and mixing time as we consider only MDPs with finite diameter or mixing time.

Definition 5 (Diameter (Jaksch et al., 2010)) *Let $T_\pi(s'|s)$ denote the first hitting time of state s' when following stationary policy π from initial state s. The diameter D of an MDP M is defined as*

$$D := \max_{s \neq s'} \min_\pi \mathbb{E}[T_\pi(s'|s)].$$

Definition 6 (Mixing time (Auer and Ortner, 2007)) *Let $\mathcal{C}_\pi$ denote the Markov chain induced by the policy π in an ergodic MDP M and let $T_{\mathcal{C}_\pi}$ represent the hitting time of $\mathcal{C}_\pi$. The mixing time T_M of M is defined as*

$$T_M := \max_\pi T_{\mathcal{C}_\pi} \,.$$

For convenience, we also introduce, for any function f defined on $\mathcal{S}$, its span defined by $\mathbb{S}(f) := \max_{s \in \mathcal{S}} f(s) - \min_{s \in \mathcal{S}} f(s)$. It actually acts as a semi-norm (see (Puterman, 2014)).

We finally introduce the following quantity that appears in the known problem-dependent lower-bounds on the regret, and plays the analogue of the mean gap in the bandit literature.

Definition 7 (Sub-optimality gap) *The sub-optimality of action a at state s is*

$$\varphi(s,a) = \mu(s,\star(s)) - \mu(s,a) + (p(\cdot|s,\star(s)) - p(\cdot|s,a))^\top b_\star \,. \tag{1}$$

Note importantly that φ is defined in terms of the bias $b_\star$ of the optimal policy $\star$. Indeed, it can be shown that minimizing the effective regret (in expectation) is essentially equivalent to minimizing the quantity $\sum_{s,a} \varphi(s,a)\mathbb{E}[N_T(s,a)]$, where $N_T(s,a)$ is the total number of steps when action a has been played in state s. More precisely, it is not difficult to show (see Appendix E for completeness) that for any stationary policy π and all t,

$$\mathbb{E}[\mathfrak{R}_{\pi,t}] = \sum_{s,a} \varphi(s,a)\mathbb{E}[N_t(s,a)] + \big((P_\pi^{t-1} - I)b_\star\big)(s_1) \leqslant \sum_{s,a} \varphi(s,a)\mathbb{E}[N_t(s,a)] + D \,. \tag{2}$$

The `KL-Ucrl` algorithm. The `KL-UCRL` algorithm (Filippi et al., 2010; Filippi, 2010) is a model-based algorithm inspired by `UCRL2` (Jaksch et al., 2010). To present the algorithm, we first describe how it defines, at each given time t, the set of plausible MDPs based on the observation available at that time. To this end, we introduce the following notations. Under a given algorithm and for a state-action pair (s,a), let $N_t(s,a)$ denote the number of visits, up to time t, to (s,a): $N_t(s,a) = \sum_{t'=0}^{t-1} \mathbb{I}\{s_{t'} = s, a_{t'} = a\}$. Then, let $N_t(s,a)^+ = \max\{N_t(s,a), 1\}$. Similarly, $N_t(s,a,s')$ denotes the number of visits to (s,a), up to time

2. The solution to this fixed-point equation is defined only up to an additive constant. Some people tend to use this equation in order to define $b_\star$ and $g_\star$, but this is a bad habit that we avoid here.
3. Throughout this paper, we may use $g^\star$ (resp. $b^\star$) and $g_\star$ (resp. $b_\star$) interchangeably.

t, followed by a visit to state s': $N_t(s,a,s') = \sum_{t'=0}^{t-1} \mathbb{I}\{s_{t'} = s, a_{t'} = a, s_{t'+1} = s'\}$. We introduce the empirical estimates of transition probabilities and rewards:

$$\hat{\mu}_t(s,a) = \frac{\sum_{t'=0}^{t-1} r_t \mathbb{I}\{s_{t'} = s, a_{t'} = a\}}{N_t(s,a)^+}, \quad \hat{p}_t(s'|s,a) = \frac{N_t(s,a,s')}{N_t(s,a)^+}\,.$$

KL-Ucrl, as an optimistic model-based approach, considers the set $\mathcal{M}_t$ as a collection of all MDPs $M' = (\mathcal{S}, \mathcal{A}, \nu', P')$, whose transition kernels and reward functions satisfy:

$$\mathtt{KL}(\hat{p}_t(\cdot|s,a), p'(\cdot|s,a)) \leqslant C_p/N_t(s,a)\,, \tag{3}$$

$$|\hat{\mu}_t(s,a) - \mu'(s,a)| \leqslant \sqrt{C_\mu/N_t(s,a)}\,, \tag{4}$$

where μ' denotes the mean of ν', and where $C_p := C_p(T,\delta) = S\,(B + \log(G)(1 + 1/G))$, with $B = B(T,\delta) := \log(2eS^2A\log(T)/\delta)$ and $G = B + 1/\log(T)$, and $C_\mu := C_\mu(T,\delta) = \log(4SA\log(T)/\delta)/1.99$. Importantly, as proven in (Filippi et al., 2010, Proposition 1), with probability at least $1 - 2\delta$, the true MDP M belongs to the set $\mathcal{M}_t$ uniformly over all time steps $t \leqslant T$.

Similarly to Ucrl2, KL-Ucrl proceeds in episodes of varying lengths; see Algorithm 1. We index an episode by $k \in \mathbb{N}$. The starting time of the k-th episode is denoted t_k, and by a slight abuse of notation, let $\mathcal{M}_k := \mathcal{M}_{t_k}$, $N_k := N_{t_k}$, $\hat{\mu}_k = \hat{\mu}_{t_k}$, and $\hat{p}_k := \hat{p}_{t_k}$. At $t = t_k$, the algorithm forms the set of plausible MDPs $\mathcal{M}_k$ based on the observations gathered so far. It then defines an extended MDP $M_{\mathtt{ext},k} = (\mathcal{S}, \mathcal{A} \times \mathcal{M}_k, \mu_{\mathtt{ext}}, P_{\mathtt{ext}})$, where for an extended action $a_{\mathtt{ext}} = (a, M')$, it defines $\mu_{\mathtt{ext}}(s, a_{\mathtt{ext}}) = \mu'(s,a)$ and $p_{\mathtt{ext}}(s'|s, a_{\mathtt{ext}}) = p'(s'|s,a)$. Then, a $\frac{1}{\sqrt{t_k}}$-optimal extended policy $\pi_{\mathtt{ext},k}$ is computed in the form $\pi_{\mathtt{ext},k}(s) = (\tilde{M}_k, \tilde{\pi}_k(s))$, in the sense that it satisfies

$$\tilde{g}_k \overset{\text{def}}{=} g_{\tilde{\pi}_k}(\tilde{M}_k) \geqslant \max_{M' \in \mathcal{M}_k, \pi} g_\pi(M') - \frac{1}{\sqrt{t_k}}\,,$$

where $g_\pi(M)$ denotes the gain of policy π in MDP M. $\tilde{M}_k$ and $\tilde{\pi}_k$ are respectively called the optimistic MDP and the optimistic policy. Finally, an episode stops at the first step $t = t_{k+1}$ when the number of local counts $v_{k,t}(s,a) = \sum_{t'=t_k}^{t} \mathbb{I}\{s_{t'} = s, a_{t'} = a\}$ exceeds $N_{t_k}(s,a)$ for some (s,a). We denote with some abuse $v_k = v_{k,t_{k+1}-1}$.

Remark 8 *The value $1/\sqrt{t_k}$ is a parameter of Extended Value Iteration and is only here for computational reasons: with sufficient computational power, it could be replaced with* 0.

Algorithm 1 KL-Ucrl (Filippi et al., 2010), with input parameter $\delta \in (0,1]$

Initialize: For all (s,a), set $N_0(s,a) = 0$ and $v_0(s,a) = 0$. Set $t = 1$, $k = 1$, and observe initial state s_1
for episodes $k \geqslant 1$ **do**
  Set $t_k = t$
  Set $N_k(s,a) = N_{k-1}(s,a) + v_{k-1}(s,a)$ for all (s,a)
  Find a $\frac{1}{\sqrt{t_k}}$-optimal policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ using Extended Value Iteration
  while $v_k(s_t, a_t) \geqslant N_k(s_t, a_t)$ **do**
    Play action $a_t = \tilde{\pi}_k(s_t)$, and observe the next state s_{t+1} and reward $r(s_t, a_t)$
    Update $N_k(s,a,x)$ and $v_k(s,a)$ for all actions a and states s, x
  end while
end for

3. Regret Lower Bound

In order to motivate the dependence of the regret on the local variance, we first provide the following minimax lower bound that makes appear this scaling.

Theorem 9 *There exists an MDP M with S states and A actions with $S, A \geqslant 10$, such that the expected regret under any algorithm $\mathbb{A}$ after $T \geqslant DSA$ steps for any initial state satisfies*

$$\mathbb{E}[\mathfrak{R}_{\mathbb{A},T}] \geqslant 0.0123\sqrt{\mathbf{V}_{\max} SAT}, \qquad \textit{where} \qquad \mathbf{V}_{\max} := \max_{s,a} \mathbb{V}_{p(\cdot|s,a)}(b^\star)\,.$$

Let us recall that (Jaksch et al., 2010) present a minimax lower bound on the regret scaling as $\Omega(\sqrt{DSAT})$. Their lower bound follows by considering a family of *hard-to-learn* MDPs. To prove the above theorem, we also consider the same MDP instances as in (Jaksch et al., 2010) and leverage their techniques. We however show that choosing a slightly different choice of transition probabilities for the problem instance leads to a lower bound scaling as $\Omega(\sqrt{\mathbf{V}_{\max} SAT})$, which does not depend on the diameter (the details are provided in the appendix).

We also remark that for the considered problem instance, easy calculations show that for any state-action pair (s, a), the variance of bias function satisfies $c_1\sqrt{D} \leqslant \mathbb{V}_{p(\cdot|s,a)}(b^\star) \leqslant c_2 D$ for some constants c_1 and c_2. Hence, the lower bound in Theorem 9 can serve as an alternative minimax lower bound without any dependence on the diameter.

4. Concentration Inequalities and The Kullback-Leibler Divergence

Before providing the novel regret bound for the `KL-UCRL` algorithm, let us discuss some important tools that we use for the regret analysis. We believe that these results, which could also be of independent interest beyond RL, shed light on some of the challenges of the regret analysis.

Let us first recall a powerful result from mathematical statistics (we provide the proof in Appendix B for completeness) known as the transportation lemma; see, e.g., (Boucheron et al., 2013, Lemma 4.18):

Lemma 10 (Transportation lemma) *For any function f, let us introduce $\varphi_f : \lambda \mapsto \log \mathbb{E}_P[\exp(\lambda(f(X) - \mathbb{E}_P[f]))]$. Whenever φ_f is defined on some possibly unbounded interval I containing 0, define its dual $\varphi_{\star,f}(x) = \sup_{\lambda \in I} \lambda x - \varphi_f(\lambda)$. Then it holds*

$$\begin{aligned}
\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \varphi_{+,f}^{-1}(\mathtt{KL}(Q,P)) \quad \textit{where} \quad \varphi_{+,f}^{-1}(t) = \inf\{x \geqslant 0 : \varphi_{\star,f}(x) > t\}\,, \\
\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\geqslant \varphi_{-,f}^{-1}(\mathtt{KL}(Q,P)) \quad \textit{where} \quad \varphi_{-,f}^{-1}(t) = \sup\{x \leqslant 0 : \varphi_{\star,f}(x) > t\}\,.
\end{aligned}$$

This result is especially interesting when Q is the empirical version of P built from n i.i.d. observations, since in that case it enables to *decouple* the concentration properties of the distribution from the specific structure of the considered function. Further, it shows that controlling the KL divergence between Q and P induces a concentration result valid for all (nice enough) functions f, which is especially useful when we do not know in advance the function f we want to handle (such as bias function $b_\star$).

The quantities $\varphi_{+,f}^{-1}$, $\varphi_{-,f}^{-1}$ may look complicated. When $f(X)$ (where $X \sim P$) is Gaussian, they coincide with $t \mapsto \pm\sqrt{2\mathbb{V}_P(f)t}$. Controlling them in general is challenging. However for bounded functions, a Bernstein-type relaxation can be derived that uses the variance $\mathbb{V}_P(f)$ and the span $\mathbb{S}(f)$:

Corollary 11 (Bernstein transportation) *For any function f such that $\mathbb{V}_P[f]$ and $\mathbb{S}(f)$ are finite,*

$$\begin{aligned}\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \sqrt{2\mathbb{V}_P[f]\mathtt{KL}(Q,P)} + \frac{2}{3}\mathbb{S}(f)\mathtt{KL}(Q,P)\,,\\ \forall Q \ll P, \quad \mathbb{E}_P[f] - \mathbb{E}_Q[f] &\leqslant \sqrt{2\mathbb{V}_P[f]\mathtt{KL}(Q,P)}\,.\end{aligned}$$

We also provide below another variation of this result that is especially useful when the bounds of Corollary 11 cannot be handled, and that seems to be new (up to our knowledge):

Lemma 12 (Transportation method II) *Let $P \in \mathcal{P}(\mathcal{X})$ be a probability distribution on a finite alphabet $\mathcal{X}$. Then, for any real-valued function f defined on $\mathcal{X}$, it holds that*

$$\begin{aligned}\forall P \ll Q, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \left(\sqrt{\mathcal{V}_{P,Q}(f)} + \sqrt{\mathcal{V}_{Q,P}(f)}\right)\sqrt{2\mathtt{KL}(P,Q)} + \mathbb{S}(f)\mathtt{KL}(P,Q)\,,\\ \text{where} \qquad \mathcal{V}_{P,Q}(f) &:= \sum_{x\in\mathcal{X}:P(x)\geqslant Q(x)} P(x)(f(x)-\mathbb{E}_P[f])^2\,.\end{aligned}$$

When P is the transition law under a state-action pair (s,a) and Q is its empirical estimates up to time t, i.e. $Q = \hat{p}_t(\cdot|s,a)$ and $P = p(\cdot|s,a)$, the first assertion in Corollary 11 can be used to decouple $\mathbb{E}_Q[f] - \mathbb{E}_P[f]$ from specific structure of f. In particular, if f is some bias function, then f has a bounded span D, and since $\mathtt{KL}(Q,P) = \widetilde{\mathcal{O}}(N_t^{-1})$, the first order terms makes appear the variance of f. This would result in a term scaling as $\widetilde{\mathcal{O}}(\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}T})$ in our regret bound, where $\widetilde{\mathcal{O}}(\cdot)$ hides poly-logarithmic terms.

Now, for the case when $Q = \hat{p}_t(\cdot|s,a)$ and $P = \tilde{p}_t(\cdot|s,a)$ is the optimistic transition law at time t, the second inequality in Corollary 11 allows us to bound $\mathbb{E}_P[f] - \mathbb{E}_Q[f]$ by the variance of f under law $\tilde{p}(\cdot|s,a)$, which itself is controlled by the variance of f under the true law $p(\cdot|s,a)$. Using such an approach would lead to a term scaling as $\widetilde{\mathcal{O}}(\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}T} + DS^2T^{1/4})$. We can remove the term scaling as $\widetilde{\mathcal{O}}(T^{1/4})$ in our regret analysis by resorting to Lemma 12 instead, in combination with the following property of the operator $\mathcal{V}$:

Lemma 13 *Consider two distributions $P, Q \in \mathcal{P}(\mathcal{X})$ with $|\mathcal{X}| \geqslant 2$. Then, for any real-valued function f defined on $\mathcal{X}$, it holds that*

$$\begin{aligned}(i) \quad \mathcal{V}_{P,Q}(f) &\leqslant \mathbb{V}_P(f)\,,\\ (ii) \quad \sqrt{\mathcal{V}_{P,Q}(f)} &\leqslant \sqrt{2\mathbb{V}_Q(f)} + 3\mathbb{S}(f)\sqrt{|\mathcal{X}|\mathtt{KL}(Q,P)}\,.\end{aligned}$$

5. Variance-Aware Regret Bound for KL-Ucrl

In this section, we present a regret upper bound for KL-UCRL that leverages the results presented in the previous section. Let $\Psi := \mathbb{S}(b^\star)$ denote the span of the bias function, and

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ define $\mathbf{V}^{\star}_{s,a} := \mathbb{V}_{p(\cdot|s,a)}(b^{\star})$ as the variance of the bias function under law $p(\cdot|s,a)$.

Let $\tilde{\star}_k$ denote the optimal policy in the extended MDP $\mathcal{M}_k$, whose gain $\tilde{g}_{\tilde{\star}_k}$ satisfies $\tilde{g}_{\tilde{\star}_k} = \max_{M' \in \mathcal{M}_k, \pi} g_{\pi}(M')$. We consider a variant of KL-UCRL, which computes, in every episode k, a policy $\tilde{\pi}_k$ satisfying: $\max_s |\tilde{b}_k(s) - \tilde{b}_{\tilde{\star}_k}(s)| \leqslant \frac{1}{\sqrt{t_k}}$, and $\tilde{g}_k \geqslant \tilde{g}_{\tilde{\star}_k} - \frac{1}{\sqrt{t_k}}$.[4]

In the following theorem, we provide a refined regret bound for KL-UCRL:

Theorem 14 (Variance-aware regret bound for KL-Ucrl) *With probability at least* $1-6\delta$*, the regret under* KL-UCRL *for any ergodic MDP* M *and for any initial state satisfies*

$$\begin{aligned}\mathfrak{R}_{\text{KL-UCRL},T} \leqslant &\Big(31\sqrt{S\textstyle\sum_{s,a}\mathbf{V}^{\star}_{s,a}} + 35S\sqrt{A} + \sqrt{2}D + 1\Big)\sqrt{TB(T,\delta)} \\ &+ \widetilde{\mathcal{O}}\Big(SA(T_M SA + D + S^{3/2})\log(T)\Big)\,,\end{aligned}$$

where $\widetilde{\mathcal{O}}$ *hides the terms scaling as* $\mathrm{polylog}(\log(T)/\delta)$*. Hence, with probability at least* $1-\delta$*,*

$$\mathfrak{R}_{\text{KL-UCRL},T} = \mathcal{O}\Big(\big[\sqrt{S\textstyle\sum_{s,a}\mathbf{V}^{\star}_{s,a}} + D\big]\sqrt{T\log(\log(T)/\delta)}\Big)\,.$$

Remark 15 *If the cardinality of the set* $\mathcal{S}^{+}_{s,a} := \{s' : p(s'|s,a) > 0\}$ *for state-action* (s,a) *is known, then one can use the following improved confidence bound for the pair* (s,a) *(instead of (3)):*

$$N_t(s,a)\mathtt{KL}(\hat{p}_t(\cdot|s,a), p'(\cdot|s,a)) \leqslant C_p^{s,a}\,, \tag{5}$$

where $C_p^{s,a} = \frac{|\mathcal{S}^{+}_{s,a}|}{S} C_p$ *(see, e.g., (Filippi, 2010, Proposition 4.1) for the corresponding concentration result). As a result, if* $|\mathcal{S}^{+}_{s,a}|$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$ *is known, it is then straightforward to show that the corresponding variant of* KL-UCRL*, which relies on (5), achives a regret growing as* $\widetilde{\mathcal{O}}\big(\sqrt{\sum_{s,a}|\mathcal{S}^{+}_{s,a}|\mathbf{V}^{\star}_{s,a}T} + D\sqrt{T}\big)$.

The regret bound provided in the aforementioned remark is of particular importance in the case of *sparse MDPs*, where most states transit to only a few next-states under various actions. We would like to stress that to get an improvement of a similar flavour for UCRL2, to the best of our knowledge, one has to know the sets $\mathcal{S}^{+}_{s,a}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ rather than their cardinalities.

Sketch of proof of Theorem 14. The detailed proof of this result is provided in Appendix C. In order to better understand it, we now provide a high level sketch of proof explaining the main steps of the analysis.

First note that by an application of Azuma-Hoeffding inequality, the effective regret is upper bounded by $\mathfrak{R}_{\mathbb{A},T} \leqslant Tg_{\star} - \sum_{t=1}^{T}\mu(s_t,a_t) + \sqrt{T\log(1/\delta)/2}$, with probability at least $1-\delta$. We proceed by decomposing the term $Tg_{\star} - \sum_{t=1}^{T}\mu(s_t,a_t)$ on the episodes $k = 1, \ldots, m(T)$, where $m(T)$ is the total number of episodes after T steps. Introducing

4. We study such a variant to facilitate the analysis and presentation of the proof. This variant of KL-UCRL may be computationally less efficient than Algorithm 1. We stress however that, in view of the number of episodes (growing as $SA\log(T)$) as well as Remark 8, with sufficient computational power such an algorithm could be practical.

$v_k(s,a)$ as the number of visits to (s,a) during episode k for any (s,a) and k, with probability at least $1-\delta$ we have

$$\mathfrak{R}_{\mathbb{A},T} \leqslant \sum_{k=1}^{m(T)} \Delta_k + \sqrt{T\log(1/\delta)/2} = \sum_{k=1}^{m(T)} \Delta_k \quad \text{where } \Delta_k = \sum_{s,a} v_k(s,a)(g^\star - \mu(s,a))\,.$$

We focus on episodes such that $M \in \mathcal{M}_k$, corresponding to valid confidence intervals, up to losing a probability only 2δ. In order to control $\Delta_k \mathbb{I}\{M \in \mathcal{M}_k\}$, we use the decomposition

$$\sum_{s,a} v_k(s,a)(g^\star - \mu(s,a)) = \sum_{s,a} v_k(s,a)(\tilde{g}_k - \mu(s,a) + (g^\star - \tilde{g}_k))\,.$$

We refrain from using the fact that $g^\star - \tilde{g}_k \leqslant 1/\sqrt{t_k}$ and instead use it as a slack later in the proof. We then introduce the bias function from the identity $\tilde{g}_k - \tilde{\mu}_k = (\widetilde{P}_k - I)\tilde{b}_k$, and thus get

$$\Delta_k \;=\; \sum_{s,a} v_k(s,a)\Big(\underbrace{(\widetilde{P}_k - P_k)b_\star}_{(a)} + \underbrace{(P_k - I)\tilde{b}_k}_{(b)} + \underbrace{(\widetilde{P}_k - P_k)(\tilde{b}_k - b_\star) + (g^\star - \tilde{g}_k)}_{(c)} \Big)$$

Term (a). The first term is controlled thanks to our variance-aware concentration inequalities:

$$\begin{aligned}
(\widetilde{P}_k - P_k)b_\star &= (\widehat{P}_k - P_k)b_\star + (\widetilde{P}_k - \widehat{P}_k)b_\star\,, \qquad \text{where} \\
\forall s, \quad ((\widehat{P}_k - P_k)b_\star)(s) &\leqslant \sqrt{2\mathbf{V}^\star_{s,\tilde{\pi}_k(s)}\mathtt{KL}(\hat{p}_k,p)} + \frac{2}{3}\mathbb{S}(b_\star)\mathtt{KL}(\hat{p}_k,p) \qquad \text{and} \\
\forall s, \quad ((\widetilde{P}_k - \widehat{P}_k)b_\star)(s) &\leqslant (1+\sqrt{2})\sqrt{2\mathbb{V}_{\hat{p}_k}(b_\star)\mathtt{KL}(\hat{p}_k,\tilde{p}_k)} + \mathbb{S}(b_\star)(1+3\sqrt{2S})\mathtt{KL}(\hat{p}_k,\tilde{p}_k)\,.
\end{aligned}$$

The first inequality is obtained by Corollary 11 while the second one by a combination of Lemma 12 together with Lemma 13. We then relate $\sqrt{\mathbb{V}_{\hat{p}_k}(b_\star)}$ to $\sqrt{\mathbb{V}_p(b_\star)}$ thanks to:

Lemma 16 *For any episode $k \geqslant 1$ such that $M \in \mathcal{M}_k$, it holds that for any pair (s,a),*

$$\sqrt{\mathbb{V}_{\hat{p}_k(\cdot|s,a)}(f)} \leqslant \sqrt{2\mathbb{V}_{p(\cdot|s,a)}(f)} + \frac{6S\mathbb{S}(f)B}{\sqrt{N_k(s,a)}} \qquad \textit{with probability at least } 1-\delta.$$

It is then not difficult to show that this first term, when summed over all episodes, contributes to the regret as $\widetilde{\mathcal{O}}(\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}}\sqrt{T\log(\log(T)/\delta)})$, where the $\log(\log(T))$ terms comes from the use of time-uniform concentration inequalities.

Term (b). We then turn to Term (b) and observe that it makes appear a martingale difference structure. Following the same reasoning as in (Jaksch et al., 2010) or (Filippi et al., 2010), the right way to control it is however to sum this contribution over all episodes and make appear a martingale difference sequence of T deterministic terms, bounded by the deterministic quantity D, since $\mathbb{S}(\tilde{b}_k) \leqslant D$. This comes at the price of losing a constant error D per episode. Now, since it can be shown that $m(T) \leqslant SA\log_2(8T/SA)$ as for `UCRL2`, we deduce that with probability higher than $1-\delta$,

$$\sum_{k=1}^{m(T)} \sum_{s,a} v_k(s,a)(P_k - I)\tilde{b}_k \leqslant D\sqrt{2T\log(1/\delta)} + 2DSA\log_2(8T/SA)\,.$$

Term (c). It thus remains to handle Term (c). To this end, we first partition the states into $\mathcal{S}_s^+ = \{x \in \mathcal{S} : \widetilde{P}_k(s,x) > P_k(s,x)\}$ and its complementary set $\mathcal{S}_s^-$, and get

$$\begin{aligned} v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b_\star) &= \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^+} (\widetilde{P}_k(s,x) - P_k(s,x))(\tilde{b}_k(x) - b_\star(x)) \\ &+ \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (\widetilde{P}_k(s,x) - P_k(s,x))(\tilde{b}_k(x) - b_\star(x)) \,. \end{aligned}$$

We thus need to control the difference of bias from above and from below. To that end, we note that by property of the bias function, it holds that

$$\tilde{b}_k - b_\star = \underbrace{\tilde{g}_{\tilde{\star}_k} - \tilde{g}_k + (\tilde{\mu}_k - \mu_k) + (\widetilde{P}_k - P_k)b_\star}_{(d)} - \varphi_k + \widetilde{P}_k(\tilde{b}_k - b_\star) \,.$$

Owing to the fact that $\tilde{g}_{\tilde{\star}_k} - \tilde{g}_k \leqslant 1/\sqrt{t_k}$ and by the previous results on concentration inequalities, the term (d) can be shown to be scaling as $\widetilde{\mathcal{O}}\Big(\sqrt{\frac{S\mathbf{V}^\star_{s,a}}{N_k(s,a)}}\Big)$. Thus, this means that provided that for all s,a, $N_k(s,a) \gtrsim \frac{S\mathbf{V}^\star_{s,a}}{\varphi(s,a)^2}$, then $(d) - \varphi(s,a) \leqslant 0$, and thus $\tilde{b}_k - b_\star \leqslant 0 + \widetilde{P}_k(0 + \ldots) \leqslant 0$. On the other hand, for the control of the last term, we first note that for an $\tilde{b}_{\tilde{\star}_k}$-improving policy (which is optimal in the extended MDP), then for all $J \in \mathbb{N}$ it holds

$$b_\star - \tilde{b}_{\tilde{\star}_k} \leqslant (\tilde{g}_{\tilde{\star}_k} - g_\star) + P_\star(b_\star - \tilde{b}_{\tilde{\star}_k}) \leqslant J(\tilde{g}_{\tilde{\star}_k} - g_\star) + P_\star^J(b_\star - \tilde{b}_{\tilde{\star}_k}) \,.$$

Thus, we obtain that

$$\begin{aligned} v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b_\star) + v_k(g_\star - \tilde{g}_{\tilde{\star}_k})\mathbf{1} &\leqslant \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))\big(P_\star^J(b_\star - \tilde{b}_{\tilde{\star}_k})\big)(x) \\ &+ \sum_s v_k(s,\tilde{\pi}_k(s))\Big[1 - J\sum_{x\in\mathcal{S}_s^-}(P_k(s,x) - \widetilde{P}_k(s,x))\Big](g_\star - \tilde{g}_{\tilde{\star}_k}) + \eta_k \,, \end{aligned} \tag{6}$$

where η_k is controlled by the error of computing $\tilde{b}_k$ in episode k (which, for the considered variant of the algorithm, is bounded by $\sqrt{32SB}\sum_{s,a}\frac{v_k(s,a)}{N_k(s,a)^+}$). In order to handle the remaining terms in (6), and choose J, we use the fact that $P_\star$ is γ-contracting for some $\gamma < 1$. Thus, choosing $J = \frac{\log(D)}{\log(1/\gamma)}$ ensures that contribution of the first term in (6) is less than $\sqrt{32SB}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}}$. Furthermore, provided that $N_k(s,a) \gtrsim SBJ^2$ for all s and a, we observe that the term in brackets is non-negative, and hence the second term in (6) becomes negative (later on we consider the case where this condition is not satisfied). Putting together, we get $(c) \leqslant (2\sqrt{32SB}+1)\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}}$.

Finally, it remains to handle the case where some state-action pair is not sufficiently sampled, that is there exists (s,a) such that $N_k(s,a) < \ell_{s,a}$, where

$$\ell_{s,a} = \ell_{s,a}(T,\delta) := \widetilde{\mathcal{O}}\Big(SB\max\Big\{\frac{\Psi}{\varphi(s,a)}, \frac{\log(D)}{\log(1/\gamma)}\Big\}^2\Big), \quad \forall s,a.$$

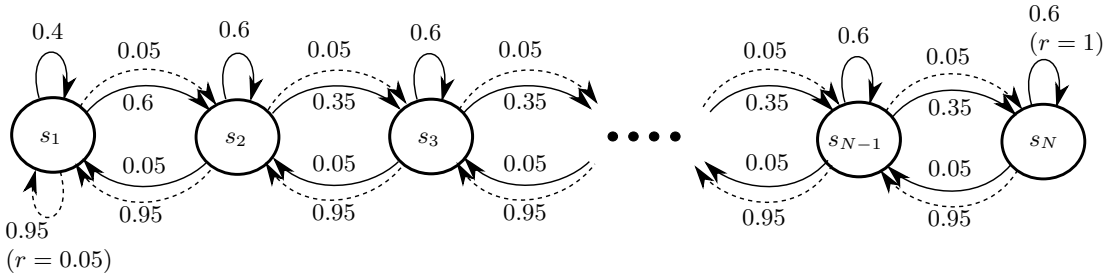

Figure 1: The N-state Ergodic *RiverSwim* MDP

Borrowing some arguments from (Auer and Ortner, 2007), we show that a given state-action pair (s, a), which is not sufficiently sampled, contributes to the regret (until it becomes sufficiently sampled) by at most $\mathcal{O}(T_M \max(\ell_{s,a}, \log(SA/\delta)))$ with probability at least $1 - \frac{\delta}{SA}$. Summing over (s, a) gives the total contribution to regret. At this point, the proof is essentially done, up to some technicalities and careful handling of second order terms.

Remark 17 *Most steps in the proof of Theorem 14 carries over to the case of communicating MDPs without restriction (up to considering the fact that for a communicating MDP, $P_\star$ may not induce a contractive mapping. Yet there exists some integer $\beta \geqslant 1$ such that $P_\star^\beta$ induces a contractive mapping; this will only affect the terms scaling as $\widetilde{\mathcal{O}}(\log(T))$ in the regret bound). It is however not clear how to appropriately bound the regret when some state-action pair is not sufficiently sampled.*

Illustrative numerical experiments. In order to better highlight the magnitude of the main terms in Theorem 14 when compared to other existing results, we consider a standard class of environments for which we compute them explicitly.

For the sake of illustration, we consider the *RiverSwim* MDP, introduced in (Strehl and Littman, 2008), as our benchmark environment. In order to satisfy ergodicity, here we consider a slightly modified version of the original *RiverSwim* (see Figure 1). Furthermore, to convey more intuition about the potential gains, we consider varying number of states. The benefits of KL-UCRL have already been studied experimentally in (Filippi et al., 2010), and we compute in Table 1 features that we believe explain the reason behind this. In particular, it is apparent that while $\Psi\sqrt{SA} \leqslant D\sqrt{SA}$ grows very large as S increases, $\mathbf{V}^\star_{s,a}$ is very small, on all tested environments, and does not change as S increases. Further, even on this simple environment, we see that $\sqrt{\sum_{s,a} \mathbf{V}^\star_{s,a}}$ is an order or magnitude smaller than $\Psi\sqrt{SA}$. We believe that these computations highlight the fact that the regret bound of Theorem 14 captures a massive improvement over the initial analysis of KL-UCRL in (Filippi et al., 2010), and over alternative algorithms such as UCRL2.

6. Conclusion

In this paper, we revisited the analysis of KL-UCRL as well as the lower bound on the regret in ergodic MDPs, in order to make appear the local variance of the bias function of the MDP. Our findings show that, owing to properties of the Kullback-Leibler divergence,

S	Ψ	$\max_{s,a} \mathbf{V}^\star_{s,a}$	$\Psi\sqrt{SA}$	$\sqrt{\sum_{s,a} \mathbf{V}^\star_{s,a}}$
6	6.3	0.6322	21.9	1.8
12	14.9	0.6327	72.9	2.8
20	26.3	0.6327	166.4	3.7
40	54.9	0.6327	490.9	5.3
70	97.7	0.6327	1156.5	7.1
100	140.6	0.6327	1988.3	8.5

Table 1: Comparison of span and variance for S-state *Ergodic RiverSwim*.

the leading term $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ obtained for the regret of `KL-UCRL` and `UCRL2` can be reduced to $\widetilde{\mathcal{O}}\Big(\sqrt{S\sum_{s,a} \mathbf{V}^\star_{s,a} T}\Big)$, while the lower bound for any algorithm can be shown to be $\Omega(\sqrt{SA\mathbf{V}_{\max} T})$, where $\mathbf{V}_{\max} := \max_{s,a} \mathbf{V}^\star_{s,a}$. Computations of these terms in some illustrative MDP show that the reported upper bound may improve an order of magnitude over the existing ones (as observed experimentally in (Filippi, 2010)), thus highlighting the fact that trading the diameter of the MDP to the local variance of the bias function may result in huge improvements.

We note that this improvement often corresponds to a gain of a factor $\mathcal{O}(\sqrt{D})$. A natural question is whether the $\sqrt{S}$ gap between the upper and lower bounds can be filled in. In the simpler setting of episodic reinforcement learning with known horizon H, several papers have shown that by taking advantage of this knowledge, it is possible to design strategies for which the regret bound does not lose a $\sqrt{S}$ factor. However, such strategies do not apply straightforwardly to undiscounted reinforcement learning. Nonetheless, we believe that combining techniques of such studies with the tools that we have developed is a fruitful research direction.

Acknowledgment

M. S. Talebi acknowledges the Ericsson Research Foundation for supporting his visit to INRIA Lille Nord – Europe. This work has been supported by CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, the French Ministry of Higher Education and Research, INRIA, and the French Agence Nationale de la Recherche (ANR), under grant ANR-16- CE40-0002 (project BADASS).

References

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 1184–1194, 2017.

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

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 22 (NIPS)*, pages

89–96, 2009.

Peter L Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 35–42, 2009.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5711–5721, 2017.

Sarah Filippi. *Stratégies optimistes en apprentissage par renforcement*. PhD thesis, Ecole nationale supérieure des telecommunications-ENST, 2010.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122, 2010.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint arXiv:1602.07182*, 2016.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 263–272, 2017.

Todd L. Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *The Journal of Machine Learning Research*, 11:1563–1600, 2010.

Odalric-Ambrym Maillard, Timothy A. Mann, and Shie Mannor. How hard is my MDP? "the distribution-norm to the rescue". In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 1835–1843, 2014.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Ian Osband, Dan Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 3003–3011, 2013.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown Markov decision processes: A Thompson sampling approach. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 1333–1342, 2017.

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

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT Press Cambridge, 1998.

Ambuj Tewari and Peter L. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 1505–1512, 2008.

Flemming Topsøe. Some bounds for the logarithmic function. *Inequality theory and applications*, 4:137, 2006.

Appendix A. Regret Lower Bound

The proof of Theorem 9 mainly relies on the problem instance for the derivation of the minimax lower bound in (Jaksch et al., 2010) and related arguments there. For the sake of completeness, we first recall their problem instance and then compute the variance of the corresponding bias function.

To get there, we first consider the two-state MDP M' shown in Figure 2, where there are two states $\{s_0, s_1\}$, each having $A' = \lfloor \frac{A-1}{2} \rfloor$ actions. We consider deterministic rewards defined as $r(s_0, a) = 0$ and $r(s_1, a) = 1$ for all $a \in \mathcal{A}$. The learner knows the rewards but not the transition probabilities. Let $\delta := \frac{4}{D}$, where D is the diameter of the MDP for which we derive the lower bound. Under any action a, $p(s_0|s_1, a) = \delta$. In state s_0, there is a unique optimal action $a^\star$, which will be referred to as the *good* action. For any $a \neq a^\star$, we have $p(s_1|s_0, a) = \delta$ whereas $p(s_1|s_0, a^\star) = \delta + \varepsilon$ for some $\varepsilon \in (0, \frac{\delta}{2})$ that will be determined later. Note that the diameter D' of M' satisfies: $D' = \frac{1}{\delta} = \frac{D}{4}$.

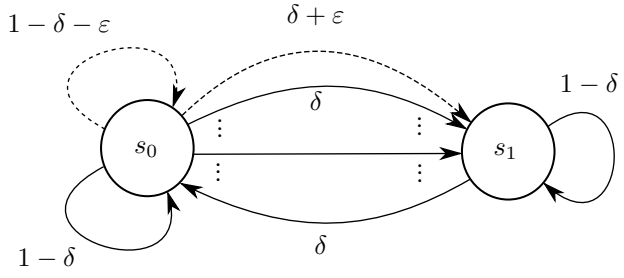

Figure 2: The MDP M' for lower bound (Jaksch et al., 2010)

We consider $\delta \in (0, \frac{1}{3})$.[5] After straightforward calculations, one finds that the average reward in M' is given by

$$g^\star = \frac{1/\delta}{1/\delta + 1/(\delta+\varepsilon)} = \frac{\delta+\varepsilon}{2\delta+\varepsilon} \, .$$

Furthermore, from Bellman optimality equation we obtain

$$b^\star(s_0) + \frac{\delta+\varepsilon}{2\delta+\varepsilon} = (\delta+\varepsilon) b^\star(s_1) + (1-\delta-\varepsilon) b^\star(s_0) \, ,$$

thus giving $\Psi := \mathbb{S}(b^\star) = b^\star(s_1) - b^\star(s_0) = \frac{1}{2\delta+\varepsilon}$. Consider $a \neq a^\star$ and let $p = p(\cdot|s_0, a)$. It follows that:

$$\begin{aligned}
\mathbb{E}_p[b^\star] &= \delta b^\star(s_1) + (1-\delta) b^\star(s_0) = b^\star(s_0) + \delta\Psi \, , \\
\mathbb{V}_p(b^\star) &= \delta (b^\star(s_1) - \mathbb{E}_p[b^\star])^2 + (1-\delta)(b^\star(s_0) - \mathbb{E}_p[b^\star])^2 = \delta(1-\delta)\Psi^2 \, .
\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}
\mathbb{V}_{p(\cdot|s_0,a^\star)}(b^\star) &= (\delta+\varepsilon)(1-\delta-\varepsilon)\Psi^2 \, , \\
\mathbb{V}_{p(\cdot|s_1,a)}(b^\star) &= \delta(1-\delta)\Psi^2 \, , \quad \forall a.
\end{aligned}$$

Hence, using the facts that $x \mapsto x(1-x)$ is increasing for $x \in [0, \frac{1}{2}]$ and $\varepsilon + \delta \leqslant \frac{1}{2}$, we obtain

$$\mathbf{V}_{\max} := \max_{s,a} \mathbb{V}_{p(\cdot|s,a)}(b^\star) = (\delta+\varepsilon)(1-\delta-\varepsilon)\Psi^2 \, .$$

A.0.1 The composite MDP

We now build a composite MDP M as considered in (Jaksch et al., 2010), as a concatenation of $k := \lfloor \frac{S}{2} \rfloor$ copies of M' in the form of an A'-ary tree, where only one copy contains the good action $a^\star$ (see Figure 3). To this end, we first add $A'+1$ additional actions so that M has at most A actions per state. For any state s_0, one of these new actions connects s_0 to the root, and the rest connect s_0 to the leaves. Whereas for any state s_1, all new actions make a transition to the same state s_1. By construction, the diameter of the composite MDP M does not exceed $2(\frac{D}{4} + \log_{A'} k)$, so that MDP M has $2\lfloor \frac{S}{2} \rfloor \leqslant S$ states, $\lfloor \frac{A'-1}{2} \rfloor + \lfloor \frac{A'-1}{2} \rfloor + 1 \leqslant A$ actions, and a diameter less than D.

A.1 Proof of Theorem 9

To derive the claimed result, we derive a lower bound on the regret for the composite MDP presented above. Our analysis is largely built on the techniques used in the proof of (Jaksch et al., 2010, Theorem 5). We also closely follow the notations used in (Jaksch et al., 2010).

Let us assume, as in the proof of (Jaksch et al., 2010, Theorem 5), that all states s_0 are identified so that M is equivalent to an MDP M' with kA' actions (note that following the same argument as in (Jaksch et al., 2010), despite the same maximal average reward, learning in M' is easier than in M, and so any regret lower bound for M' implies a lower

5. The case of $\delta > 1/3$ can be handled similarly to the analysis of (Jaksch et al., 2010).

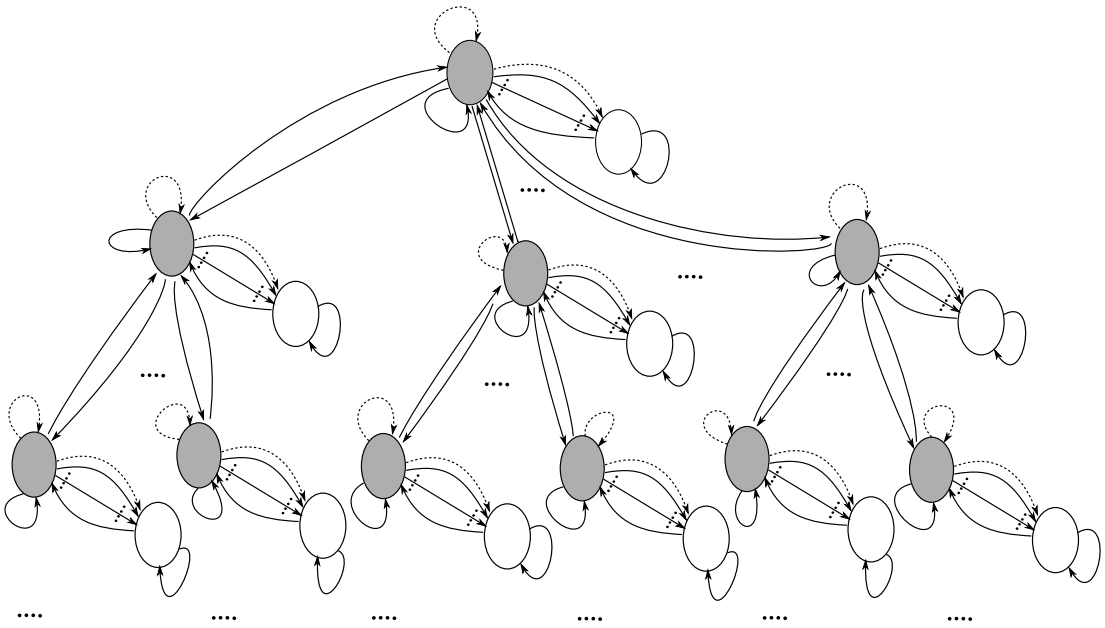

Figure 3: The composite MDP M (Jaksch et al., 2010)

bound in M, too). Note that by construction of M, it holds that $\mathbf{V}_{\max}$ in M equals $\mathbf{V}_{\max}$ in M'. Denote by $(s_0^\star, a^\star)$ the *good copy*, i.e., the one containing the good action $a^\star$. We assume that $a^\star$ is chosen uniformly at random among all actions $\{1,\dots,k\} \times \{1,\dots,A'\}$. Let $\mathbb{E}_\star[\cdot]$ and $\mathbb{E}_{\text{unif}}[\cdot]$ respectively denote the expectation with respect to the random choice of $(s_0^\star, a^\star)$ and the expectation when there is no good action. Furthermore, let $\mathbb{E}_a[\cdot]$ denote the expectation conditioned on $a = a^\star$, and introduce N_1, N_0, and $N_0^\star$ as the respective number of visits to s_1, s_0, and $(s_0, a^\star)$.

The proof proceeds in the same steps as in the proof of (Jaksch et al., 2010, Theorem 5) up to Equation (36) there, where it is shown that assuming that the initial state is s_0,

$$\begin{aligned}\mathbb{E}_a[N_1] &\leqslant \mathbb{E}_a[N_0 - N_0^\star] + (\delta+\varepsilon)D'\mathbb{E}_a[N_0^\star] \leqslant T - \mathbb{E}_{\text{unif}}[N_1] + \varepsilon D'\mathbb{E}_a[N_0^\star]\,,\\ \mathbb{E}_{\text{unif}}[N_1] &\geqslant \frac{T-D'}{2}\,,\end{aligned}$$

so that the accumulated reward $R_{\mathbb{A},T}$ by the algorithm $\mathbb{A}$ in M' up to time step T satisfies

$$\mathbb{E}_a[R_{\mathbb{A},T}] \leqslant \mathbb{E}_a[N_1] \leqslant \frac{T+D'}{2} + \varepsilon D'\mathbb{E}_a[N_0^\star]\,.$$

The following lemma, which is a straightforward modification to (Jaksch et al., 2010, Lemma 13), enables us to control $\mathbb{E}_a[N_0^\star]$:

Lemma 18 *Let $f : \{s_0, s_1\}^{T+1} \mapsto [0, B]$ be any function defined on any trajectory $\boldsymbol{s}_{T+1} = (s_t)_{1\leqslant t\leqslant T+1}$ in M'. Then, for any $\delta \in [0, \frac{1}{3}]$, $\varepsilon \in (0, 1-2\delta)$, and $a \in \{1,\dots,kA'\}$,*

$$\mathbb{E}_a[f(\boldsymbol{s})] \leqslant \mathbb{E}_{\text{unif}}[f(\boldsymbol{s})] + \varepsilon B\sqrt{\frac{\log(2)\mathbb{E}_{\text{unif}}[N_0^\star]}{2(\delta+\varepsilon)(1-\delta-\varepsilon)}}\,.$$

Noting that $N_0^\star$ is a function of $\boldsymbol{s}_{T+1}$ satisfying $N_0^\star \in [0,T]$, by Lemma 18 we deduce

$$\begin{aligned}\mathbb{E}_a[N_0^\star] &\leqslant \mathbb{E}_{\text{unif}}[N_0^\star] + \varepsilon T\sqrt{\frac{\log(2)\mathbb{E}_{\text{unif}}[N_0^\star]}{2(\delta+\varepsilon)(1-\delta-\varepsilon)}} \\ &= \mathbb{E}_{\text{unif}}[N_0^\star] + \varepsilon\Psi T\sqrt{\frac{\log(2)\mathbb{E}_{\text{unif}}[N_0^\star]}{2\mathbf{V}_{\max}}}\,,\end{aligned}$$

where we used $\sqrt{\mathbf{V}_{\max}} = \Psi\sqrt{(\delta+\varepsilon)(1-\varepsilon-\delta)}$. As shown in the proof of (Jaksch et al., 2010, Theorem 5), $\sum_{a=1}^{kA'}\mathbb{E}_{\text{unif}}[N_0^\star] \leqslant (T+D')/2$ and $\sum_{a=1}^{kA'}\sqrt{\mathbb{E}_{\text{unif}}[N_0^\star]} \leqslant \sqrt{kA'(T+D')/2}$, so that we finally get, using the relation $\mathbb{E}_\star[R_{\mathbb{A},T}] = \frac{1}{kA'}\sum_{a=1}^{kA'}\mathbb{E}_a[R_{\mathbb{A},T}]$,

$$\begin{aligned}\mathbb{E}_\star[\mathfrak{R}_{\mathbb{A},T,M'}] &= \frac{\delta+\varepsilon}{2\delta+\varepsilon}T - \mathbb{E}_\star[R_{\mathbb{A},T}] \\ &\geqslant \frac{\delta+\varepsilon}{2\delta+\varepsilon}T - \frac{T}{2} - \frac{\varepsilon D'T}{2kA'} - \frac{\varepsilon D'^2}{2kA'} \\ &\quad - \frac{\varepsilon^2\Psi D'T}{kA'}\sqrt{\frac{\log(2)kA'T}{4\mathbf{V}_{\max}}} - \frac{\varepsilon^2\Psi D'T}{kA'}\sqrt{\frac{\log(2)kA'D'}{4\mathbf{V}_{\max}}} - \frac{D'}{2} \\ &\geqslant \frac{\varepsilon T}{4\delta+2\varepsilon} - \frac{\varepsilon D'}{2kA'}(T+D') - \frac{0.42\varepsilon^2\Psi D'T}{\sqrt{kA'\mathbf{V}_{\max}}}(\sqrt{T}+\sqrt{D'}) - \frac{D'}{2}\,.\end{aligned}$$

Noting that the assumption $T \geqslant DSA$ implies $T \geqslant 16D'kA'$, we deduce that

$$\mathbb{E}_\star[\mathfrak{R}_{\mathbb{A},T,M'}] \geqslant \frac{\varepsilon T}{4\delta+2\varepsilon} - \frac{\varepsilon D'T}{2kA'}\Big(1+\frac{1}{16kA'}\Big) - \frac{0.42\varepsilon^2\Psi D'T\sqrt{T}}{\sqrt{kA'\mathbf{V}_{\max}}}\Big(1+\frac{1}{4\sqrt{kA'}}\Big) - \frac{D'}{2}\,.$$

The first term in the right-hand side of the above satisfies

$$\frac{\varepsilon T}{4\delta+2\varepsilon} = \frac{\varepsilon T\Psi}{2} \geqslant \frac{5\varepsilon\mathbf{V}_{\max}T}{6}\,,$$

since

$$\frac{\Psi}{\mathbf{V}_{\max}} = \frac{2\delta+\varepsilon}{(\delta+\varepsilon)(1-\delta-\varepsilon)} \geqslant 1 + \frac{\delta}{\delta+\varepsilon} > \frac{5}{3}\,,$$

where we used $\varepsilon \leqslant \frac{\delta}{2}$ in the last step. Hence, we get

$$\mathbb{E}_\star[\mathfrak{R}_{\mathbb{A},T,M'}] \geqslant \frac{5}{6}\varepsilon\mathbf{V}_{\max}T - \frac{\varepsilon D'T}{2kA'}\Big(1+\frac{1}{16kA'}\Big) - \frac{0.42\varepsilon^2\Psi D'T\sqrt{T}}{\sqrt{kA'\mathbf{V}_{\max}}}\Big(1+\frac{1}{4\sqrt{kA'}}\Big) - \frac{D'}{2}\,.$$

In particular, setting $\varepsilon = c\sqrt{\frac{kA'}{\mathbf{V}_{\max}T}}$ for some c (which will be determined later) yields

$$\begin{aligned}\mathbb{E}_\star[\mathfrak{R}_{\mathbb{A},T,M'}] &\geqslant \frac{5}{6}c\sqrt{kA'\mathbf{V}_{\max}T} - \sqrt{kA'\mathbf{V}_{\max}T}\left(\frac{cD'}{2kA'\mathbf{V}_{\max}}\Big(1+\frac{1}{16kA'}\Big)\right) \\ &\quad - \sqrt{kA'\mathbf{V}_{\max}T}\left(\frac{0.42c^2}{kA'}\frac{D'\Psi}{\mathbf{V}_{\max}^2}\Big(1+\frac{1}{4\sqrt{kA'}}\Big)\right) - \frac{D'}{2}\,.\end{aligned}$$

To simplify the above bound, note that

$$\frac{D'}{\mathbf{V}_{\max}} \leqslant \frac{(2\delta+\varepsilon)^2}{\delta(\delta+\varepsilon)(1-\delta-\varepsilon)} \leqslant 2\Big(\frac{2\delta+\varepsilon}{\delta}\Big)^2 \leqslant 12.5\,, \tag{7}$$

where we used $1-\varepsilon-\delta \geqslant \frac{1}{2}$ since $\varepsilon \leqslant \frac{\delta}{2}$. Moreover,

$$\begin{aligned}\frac{D'\Psi}{\mathbf{V}_{\max}^2} &= \frac{D'\Psi}{\Psi^4(\delta+\varepsilon)^2(1-\delta-\varepsilon)^2} \\ &= \frac{(2\delta+\varepsilon)^3}{\delta(\delta+\varepsilon)^2(1-\delta-\varepsilon)^2} \leqslant 4\Big(\frac{2\delta+\varepsilon}{\delta}\Big)^3 \leqslant 62.5\,.\end{aligned}$$

Putting these together with the fact that

$$\frac{D'}{2} \leqslant \frac{\sqrt{D'}}{2}\sqrt{\frac{T}{16kA'}} \leqslant \frac{\sqrt{12.5/16}}{2}\sqrt{\frac{\mathbf{V}_{\max}T}{kA'}} \leqslant 0.45\sqrt{\frac{\mathbf{V}_{\max}T}{kA'}}\,,$$

which follows from (7), we deduce that

$$\mathbb{E}_\star[\mathfrak{R}_{\mathbb{A},T,M'}] \geqslant \sqrt{kA'\mathbf{V}_{\max}T}\Big(\frac{5c}{6} - \frac{12.5c}{2kA'} - \frac{12.5c}{32(kA')^2} - \frac{26.25c^2}{kA'} - \frac{6.6c^2}{(kA')^{3/2}} - \frac{0.45}{kA'}\Big),$$

Taking $c = 0.132$ and using the facts $k = \lfloor \frac{S}{2} \rfloor \geqslant 5$ and $A' = \lfloor \frac{A-1}{2} \rfloor \geqslant 4$ yield the announced result. This completes the proof provided that we show that this choice of c satisfies $\varepsilon \leqslant \frac{\delta}{2}$. To this end, observe that by the assumption $T \geqslant DSA \geqslant \frac{16kA'}{\delta}$, it follows that

$$\varepsilon = 0.132\sqrt{\frac{kA'}{\mathbf{V}_{\max}T}} \leqslant \frac{0.132}{4}\sqrt{\frac{\delta}{\mathbf{V}_{\max}}} \leqslant \frac{0.132}{4}\sqrt{\frac{\delta(2\delta+\varepsilon)^2}{(\delta+\varepsilon)(1-\delta-\varepsilon)}} \leqslant 0.047(2\delta+\varepsilon)\,,$$

so that $\varepsilon \leqslant 0.1\delta$. This concludes the proof. □

A.2 Proof of Lemma 18

The lemma follows by a slight modification of the proof of (Jaksch et al., 2010, Lemma 13). We recall that according to Equations (49)-(51) in (Jaksch et al., 2010),

$$\mathbb{E}_a[f(\boldsymbol{s})] - \mathbb{E}_{\text{unif}}[f(\boldsymbol{s})] \leqslant \frac{B}{2}\sqrt{2\log(2)\mathtt{KL}(\mathbb{P}_{\text{unif}}, \mathbb{P}_a)}\,, \tag{8}$$

where

$$\begin{aligned}\mathtt{KL}(\mathbb{P}_{\text{unif}}, \mathbb{P}_a) &= \sum_{t=1}^{T} \mathtt{KL}(\mathbb{P}_{\text{unif}}(s_{t+1}|\boldsymbol{s}^t), \mathbb{P}_a(s_{t+1}|\boldsymbol{s}^t)) \\ &= \sum_{t=1}^{T} \mathbb{P}_{\text{unif}}(s_t = s_0, a_t = a)\left(\delta\log\Big(\frac{\delta}{\delta+\varepsilon}\Big) + (1-\delta)\log\Big(\frac{1-\delta}{1-\delta-\varepsilon}\Big)\right).\end{aligned}$$

Now using the inequality $\mathtt{kl}(a,b) \leqslant \frac{(a-b)^2}{b(1-b)}$ valid for all $a, b \in (0,1)$ (instead of (Jaksch et al., 2010, Lemma 20)) and noting that $\mathbb{E}_{\text{unif}}[N_0^\star] = \sum_{t=1}^{T} \mathbb{P}_{\text{unif}}(s_t = s_0, a_t = a)$, we obtain

$$\mathtt{KL}(\mathbb{P}_{\text{unif}}, \mathbb{P}_a) = \mathtt{kl}(\delta, \delta+\varepsilon)\mathbb{E}_{\text{unif}}[N_0^\star] \leqslant \frac{\varepsilon^2}{(1-\delta)(1-\delta-\varepsilon)}\mathbb{E}_{\text{unif}}[N_0^\star]\,.$$

Plugging this into (8) completes the proof. □

Appendix B. Concentration Inequalities

B.1 Proof of Lemma 10

Let us recall the fundamental equality

$$\forall \lambda \in \mathbb{R}, \quad \log \mathbb{E}_P[\exp(\lambda(X - \mathbb{E}_P[X])] = \sup_{Q \ll P} \left[\lambda\Big(\mathbb{E}_Q[X] - \mathbb{E}_P[X]\Big) - \mathtt{KL}(Q,P) \right].$$

In particular, we obtain on the one hand that (see also (Boucheron et al., 2013, Lemma 2.4))

$$\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \;\leqslant\; \min_{\lambda \in \mathbb{R}^+} \frac{\varphi_f(\lambda) + \mathtt{KL}(Q,P)}{\lambda}\,.$$

Since $\varphi_f(0) = 0$, then the right-hand side of the above is non-negative. Let us call it u. Now, we note that for any t such that $u \geqslant t \geqslant 0$, by construction of u, it holds that $\mathtt{KL}(Q,P) \geqslant \varphi_{\star,f}(t)$. Thus, $\{x \geqslant 0 : \varphi_{f,\star}(x) > \mathtt{KL}(Q,P)\} = (u, \infty)$ and hence, $u = \varphi_{+,f}^{-1}(\mathtt{KL}(Q,P))$.

On the other hand, it holds

$$\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \;\geqslant\; \max_{\lambda \in \mathbb{R}^-} \frac{\varphi_f(\lambda) + \mathtt{KL}(Q,P)}{\lambda}\,.$$

Since $\varphi(0) = 0$, then the right-hand side quantity is non-positive. Let us call it v. Now, we note that for any t such that $v \leqslant t \leqslant 0$, by construction of v, it holds that $\mathtt{KL}(Q,P) \geqslant \varphi_{\star,f}(t)$. Thus, $\{x \leqslant 0 : \varphi_{\star,f}(x) > \mathtt{KL}(Q,P)\} = (-\infty, v)$ and hence, $v = \varphi_{-,f}^{-1}(\mathtt{KL}(Q,P))$. □

B.2 Proof of Corollary 11

By a standard Bernstein argument (see for instance (Boucheron et al., 2013, Section 2.8)), it holds

$$\begin{aligned}
\forall \lambda \in [0, 3/\mathbb{S}(f)), \quad \varphi_f(\lambda) &\leqslant \frac{\mathbb{V}_P[f]}{2} \frac{\lambda^2}{1 - \frac{\mathbb{S}(f)\lambda}{3}}\,, \\
\forall x \geqslant 0, \quad \varphi_{\star,f}(x) &\geqslant \frac{x^2}{2(\mathbb{V}_P[f] + \frac{\mathbb{S}(f)}{3}x)}\,.
\end{aligned}$$

Then, a direct computation (solving for x in $\varphi_{\star,f}(x) = t$) shows that

$$\begin{aligned}
\varphi_{+,f}^{-1}(t) &\leqslant \frac{\mathbb{S}(f)}{3}t + \sqrt{2t\mathbb{V}_P[f] + \Big(\frac{\mathbb{S}(f)}{3}t\Big)^2} \leqslant \sqrt{2t\mathbb{V}_P[f]} + \frac{2}{3}t\mathbb{S}(f)\,, \\
\varphi_{-,f}^{-1}(t) &\geqslant \frac{\mathbb{S}(f)}{3}t - \sqrt{2t\mathbb{V}_P[f] + \Big(\frac{\mathbb{S}(f)}{3}t\Big)^2} \geqslant -\sqrt{2t\mathbb{V}_P[f]}\,,
\end{aligned}$$

where we used that $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ for $a, b \geqslant 0$. Combining these bounds, we get

$$\begin{aligned}
\mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \sqrt{2\mathbb{V}_P[f]\mathtt{KL}(Q,P)} + \frac{2}{3}\mathbb{S}(f)\mathtt{KL}(Q,P)\,, \\
\mathbb{E}_P[f] - \mathbb{E}_Q[f] &\leqslant \sqrt{2\mathbb{V}_P[f]\mathtt{KL}(Q,P)}\,.
\end{aligned}$$

□

B.3 Proof of Lemma 12

If $\mathbb{E}_Q[f] \leqslant \mathbb{E}_P[f]$, then the result holds trivially. We thus assume that $\mathbb{E}_Q[f] > \mathbb{E}_P[f]$. It is straightforward to verify that

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] = \sum_{x:Q(x)\geqslant P(x)} (f(x) - \mathbb{E}_Q[f])(Q(x) - P(x)) + \sum_{x:Q(x)<P(x)} (f(x) - \mathbb{E}_P[f])(Q(x) - P(x)) + \sum_{x:P(x)>Q(x)} (\mathbb{E}_P[f] - \mathbb{E}_Q[f])(Q(x) - P(x)) \,. \tag{9}$$

The first term in the right-hand side of (9) is upper bounded as

$$\begin{aligned}\sum_{x:Q(x)\geqslant P(x)} (f(x) - \mathbb{E}_Q[f])(Q(x) - P(x)) &= \sum_{x:Q(x)\geqslant P(x)} \sqrt{Q(x)}(f(x) - \mathbb{E}_Q[f])\frac{Q(x) - P(x)}{\sqrt{Q(x)}} \\ &\overset{(a)}{\leqslant} \sqrt{\sum_{x:Q(x)\geqslant P(x)} Q(x)(f(x) - \mathbb{E}_Q[f])^2}\sqrt{\sum_{x:Q(x)\geqslant P(x)} \frac{(Q(x) - P(x))^2}{Q(x)}} \\ &\overset{(b)}{\leqslant} \sqrt{\mathcal{V}_{Q,P}(f)}\sqrt{2\mathtt{KL}(P, Q)} \,, \end{aligned} \tag{10}$$

where (a) follows from Cauchy-Schwarz inequality and (b) follows from Lemma 22.

Similarly, the second term in (9) satisfies

$$\begin{aligned}\sum_{x:Q(x)<P(x)} (f(x) - \mathbb{E}_P[f])(Q(x) - P(x)) &= \sum_{x:Q(x)<P(x)} \sqrt{P(x)}(f(x) - \mathbb{E}_P[f])\frac{Q(x) - P(x)}{\sqrt{P(x)}} \\ &\leqslant \sqrt{\mathcal{V}_{P,Q}(f)}\sqrt{2\mathtt{KL}(P, Q)} \,. \end{aligned} \tag{11}$$

Finally, we bound the last term in (9):

$$\begin{aligned}(\mathbb{E}_P[f] - \mathbb{E}_Q[f]) \sum_{x:P(x)>Q(x)} (Q(x) - P(x)) &\overset{(a)}{=} \frac{1}{2}(\mathbb{E}_Q[f] - \mathbb{E}_P[f])\|P - Q\|_1 \\ &\leqslant \frac{1}{2}\mathbb{S}(f)\|P - Q\|_1^2 \overset{(b)}{\leqslant} \mathbb{S}(f)\mathtt{KL}(P, Q) \,, \end{aligned} \tag{12}$$

where (a) follows from the fact that for any pair of distributions $U, V \in \mathcal{P}(\mathcal{X})$, it holds that $\sum_{x\in\mathcal{X}} |U(x) - V(x)| = 2\sum_{x:U(x)\geqslant V(x)}(U(x) - V(x))$, and where (b) follows from Pinsker's inequality. The proof is concluded by combining (10), (11), and (12). □

B.4 Proof of Lemma 13

Statement (i) is a direct consequence of the definition of $\mathcal{V}_{P,Q}$. We next prove statement (ii). Observe that Lemma 22 implies that for all $x \in \mathcal{X}$

$$|P(x) - Q(x)| \leqslant \sqrt{2\max(P(x), Q(x))\mathtt{KL}(Q, P)} \,.$$

Hence,

$$\begin{aligned}\mathcal{V}_{P,Q}(f) &= \sum_{x:P(x)\geqslant Q(x)} P(x)(f(x)-\mathbb{E}_P[f])^2 \\ &\leqslant \sum_{x:P(x)\geqslant Q(x)} Q(x)(f(x)-\mathbb{E}_P[f])^2 + \sqrt{2\mathtt{KL}(Q,P)} \sum_{x:P(x)\geqslant Q(x)} \sqrt{P(x)}(f(x)-\mathbb{E}_P[f])^2 \,. \end{aligned} \tag{13}$$

The first term in the right-hand side of (13) is bounded as follows:

$$\begin{aligned}\sum_{x:P(x)\geqslant Q(x)} Q(x)(f(x)-\mathbb{E}_P[f])^2 &\leqslant 2 \sum_{x:P(x)\geqslant Q(x)} Q(x)(f(x)-\mathbb{E}_Q[f])^2 + 2(\mathbb{E}_Q[f]-\mathbb{E}_P[f])^2 \\ &\leqslant 2\mathbb{V}_Q(f) + 2(\mathbb{E}_Q[f]-\mathbb{E}_P[f])^2 \,.\end{aligned}$$

Note that

$$(\mathbb{E}_Q[f]-\mathbb{E}_P[f])^2 \leqslant \mathbb{S}(f)^2\|P-Q\|_1^2 \leqslant 2\mathbb{S}(f)^2\mathtt{KL}(Q,P) \,,$$

which further gives

$$\sum_{x:P(x)\geqslant Q(x)} Q(x)(f(x)-\mathbb{E}_P[f])^2 \leqslant 2\mathbb{V}_Q(f) + 4\mathbb{S}(f)^2\mathtt{KL}(Q,P) \,.$$

Now we consider the second term in (13). First observe that

$$\begin{aligned}\sum_{x:P(x)\geqslant Q(x)} \sqrt{P(x)}(f(x)-\mathbb{E}_P[f])^2 &\leqslant \sqrt{\sum_{x:P(x)\geqslant Q(x)} P(x)(f(x)-\mathbb{E}_P[f])^2}\sqrt{\sum_x (f(x)-\mathbb{E}_P[f])^2} \\ &\leqslant \sqrt{\mathcal{V}_{P,Q}(f)}\mathbb{S}(f)\sqrt{|\mathcal{X}|} \,,\end{aligned}$$

thanks to Cauchy-Schwarz inequality. Hence, the second term in (13) is upper bounded by

$$\mathbb{S}(f)\sqrt{2|\mathcal{X}|\mathcal{V}_{P,Q}(f)\mathtt{KL}(Q,P)} \,.$$

Combining the previous bounds together, we get

$$\mathcal{V}_{P,Q}(f) \leqslant 2\mathbb{V}_Q(f) + 4\mathbb{S}(f)^2\mathtt{KL}(Q,P) + \mathbb{S}(f)\sqrt{2|\mathcal{X}|\mathcal{V}_{P,Q}(f)\mathtt{KL}(Q,P)} \,,$$

which leads to

$$\left(\sqrt{\mathcal{V}_{P,Q}(f)} - \mathbb{S}(f)\sqrt{|\mathcal{X}|\mathtt{KL}(Q,P)/2}\right)^2 \leqslant 2\mathbb{V}_Q(f) + \mathbb{S}(f)^2(|\mathcal{X}|/2+4)\mathtt{KL}(Q,P) \,,$$

so that using the inequality $\sqrt{a+b} \leqslant \sqrt{a}+\sqrt{b}$, we finally obtain

$$\begin{aligned}\sqrt{\mathcal{V}_{Q,P}(f)} &\leqslant \sqrt{2\mathbb{V}_Q(f) + \mathbb{S}(f)^2(|\mathcal{X}|/2+4)\mathtt{KL}(Q,P)} + \mathbb{S}(f)\sqrt{|\mathcal{X}|\mathtt{KL}(Q,P)/2} \\ &\leqslant \sqrt{2\mathbb{V}_Q(f)} + \mathbb{S}(f)(\sqrt{2|\mathcal{X}|}+2)\sqrt{\mathtt{KL}(Q,P)} \,.\end{aligned}$$

The proof is completed by observing that $\sqrt{2|\mathcal{X}|}+2 \leqslant 3\sqrt{|\mathcal{X}|}$ for $|\mathcal{X}| \geqslant 2$. □

B.5 Proof of Lemma 16

Let $\delta \in (0,1)$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Consider an episode $k \geqslant 1$ such that $M \in \mathcal{M}_k$, and define $\hat{p}_k = \hat{p}_k(\cdot|s,a)$, $p = p(\cdot|s,a)$, and $N_k = N_k(s,a)$. Observe that by a Bernstein-like inequality (Dann et al., 2017, Lemma F.2), we have: for all $s' \in \mathcal{S}$, with probability at least $1-\delta$,

$$\hat{p}_k(s') - p(s') \leqslant \sqrt{\frac{2p(s')C_b}{N_k}} + \frac{2C_b}{N_k} \,,$$

with $C_b = C_b(t,\delta) := \log(3\log(\max(e,t))/\delta)$. It then follows that with probability at least $1-\delta$,

$$\begin{aligned}
\mathbb{V}_{\hat{p}_k}(f) &= \sum_{s'} \hat{p}_k(s')(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2 \\
&\leqslant \sum_{s'} p(s')(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2 + \sqrt{\frac{2C_b}{N_k}} \sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2 + \frac{2C_b}{N_k} \sum_{s'} (f(s') - \mathbb{E}_{\hat{p}_k}[f])^2 \\
&\leqslant \underbrace{\sum_{s'} p(s')(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2}_{Z_1} + \sqrt{\frac{2C_b}{N_k}} \underbrace{\sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2}_{Z_2} + \frac{2C_b S \mathbb{S}(f)^2}{N_k} \,. \qquad (14)
\end{aligned}$$

Next we bound Z_1 and Z_2. Observe that

$$\begin{aligned}
Z_1 &\leqslant 2\sum_{s'} p(s')(f(s') - \mathbb{E}_p[f])^2 + 2(\mathbb{E}_p[f] - \mathbb{E}_{\hat{p}_k}[f])^2 \\
&\leqslant 2\mathbb{V}_p(f) + 4\mathbb{S}(f)^2 \mathtt{KL}(\hat{p}_k, p) \,,
\end{aligned}$$

where the last inequality follows from

$$(\mathbb{E}_p[f] - \mathbb{E}_{\hat{p}_k}[f])^2 \leqslant \mathbb{S}(f)^2 \|p - \hat{p}_k\|_1^2 \leqslant 2\mathbb{S}(f)^2 \mathtt{KL}(\hat{p}_k, p) \,. \qquad (15)$$

For Z_2 we have

$$Z_2 \leqslant 2\sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_p[f])^2 + 2(\mathbb{E}_p[f] - \mathbb{E}_{\hat{p}_k}[f])^2 \sum_{s'} \sqrt{p(s')} \,.$$

Now, using Cauchy-Schwarz inequality

$$\begin{aligned}
\sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_p[f])^2 &\leqslant \sqrt{\sum_{s'} p(s')(f(s') - \mathbb{E}_p[f])^2 \sum_{s'} (f(s') - \mathbb{E}_p[f])^2} \\
&\leqslant \sqrt{S\mathbb{V}_p(f)}\mathbb{S}(f) \,,
\end{aligned}$$

so that using (15), we deduce that

$$\begin{aligned}
Z_2 &\leqslant 2\mathbb{S}(f)\sqrt{S\mathbb{V}_p(f)} + 4\mathbb{S}(f)^2 \mathtt{KL}(\hat{p}_k, p) \sum_{s'} \sqrt{p(s')} \\
&\leqslant 2\mathbb{S}(f)\sqrt{S\mathbb{V}_p(f)} + 4\mathbb{S}(f)^2 \mathtt{KL}(\hat{p}_k, p)\sqrt{S} \,,
\end{aligned}$$

where the last inequality follows from Jensen's inequality:

$$\sum_{s'} \sqrt{p(s')} = \sum_{s'} p(s') \sqrt{\frac{1}{p(s')}} \leqslant \sum_{s'} \sqrt{\frac{p(s')}{p(s')}} = \sqrt{S} \,.$$

Putting together, we deduce that with probability at least $1-\delta$,

$$\mathbb{V}_{\hat{p}_k}(f) \leqslant 2\mathbb{V}_p(f) + 2\mathbb{S}(f)\sqrt{\frac{2SC_b}{N_k}}\Big(\sqrt{\mathbb{V}_p(f)} + 2\mathbb{S}(f)\mathtt{KL}(\hat{p}_k, p)\Big) + \mathbb{S}(f)^2\Big(4\mathtt{KL}(\hat{p}_k, p) + \frac{2SC_b}{N_k}\Big) \,.$$

Noting that $M \in \mathcal{M}_k$, we obtain

$$\begin{aligned}
\mathbb{V}_{\hat{p}_k}(f) &\leqslant 2\mathbb{V}_p(f) + \mathbb{S}(f)\sqrt{\frac{8S\mathbb{V}_p(f)C_b}{N_k}} + 4\mathbb{S}(f)^2\frac{\sqrt{2SC_b}C_p}{N_k^{3/2}} + \frac{(4C_p + 2SC_b)\mathbb{S}(f)^2}{N_k} \\
&\leqslant 2\mathbb{V}_p(f) + \mathbb{S}(f)\sqrt{\frac{8S\mathbb{V}_p(f)C_b}{N_k}} + \frac{S\mathbb{S}(f)^2}{N_k}(16B\sqrt{2SC_b} + 16B + 2C_b) \\
&\leqslant 2\mathbb{V}_p(f) + \mathbb{S}(f)\sqrt{\frac{8S\mathbb{V}_p(f)B}{N_k}} + \frac{36S^{3/2}B^{3/2}\mathbb{S}(f)^2}{N_k} \,,
\end{aligned}$$

with probability at least $1-\delta$, where we used $C_p = 4SB$, $C_b \leqslant B$, and $S \geqslant 2$. The proof is concluded by observing that

$$\begin{aligned}
\sqrt{\mathbb{V}_{\hat{p}_k}(f)} &\leqslant \sqrt{2\mathbb{V}_p(f)} + \mathbb{S}(f)\sqrt{\frac{SB}{N_k}} + 6\mathbb{S}(f)B\sqrt{\frac{S^{3/2}}{N_k}} \\
&\leqslant \sqrt{2\mathbb{V}_p(f)} + \frac{6S\mathbb{S}(f)B}{\sqrt{N_k}} \,,
\end{aligned}$$

with probability at least $1-\delta$. □

Appendix C. Regret Upper Bound for `KL-Ucrl`

In this section, we provide the proof of the main result (Theorem 14). We will try to closely follow the notations used in the proof of (Jaksch et al., 2010, Theorem 2).

We first recall the following result indicating that the true model belongs to the set of plausible MDPs with high probability. Recall that for $\delta \in (0,1]$ and $t \in \mathbb{N}$,

$$\begin{aligned}
C_\mu &:= C_\mu(T,\delta) = \log(4SA\log(T)/\delta)/1.99 \,, \\
C_p &:= C_p(T,\delta) = S\left(B + \log(G)(1+1/G)\right) \,,
\end{aligned}$$

where

$$\begin{aligned}
B &:= B(T,\delta) = \log(2eS^2A\log(T)/\delta) \,, \\
G &:= G(T,\delta) = B + 1/\log(T) \,.
\end{aligned} \tag{16}$$

Moreover, observe that $C_p \leqslant 4SB$.

Lemma 19 ((Filippi et al., 2010, Proposition 1)) *For all $T \geqslant 1$ and $\delta > 0$, and for any pair (s,a), it holds that*

$$\mathbb{P}\Big(\forall t \leqslant T,\ |\hat{\mu}_t(s,a) - \mu(s,a)| \leqslant \sqrt{C_\mu / N_t(s,a)}\Big) \geqslant 1 - \frac{\delta}{SA}\,,$$
$$\mathbb{P}\Big(\forall t \leqslant T,\ N_t(s,a)\mathtt{KL}(\hat{p}_t(s,a), p(\cdot|s,a)) \leqslant C_p\Big) \geqslant 1 - \frac{\delta}{SA}\,.$$

In particular, $\mathbb{P}(\forall t \leqslant T,\ M \in \mathcal{M}_t) \geqslant 1 - 2\delta$.

Next we prove the theorem.

Proof (of Theorem 14). Let $T \geqslant 1$ and $\delta \in (0,1)$. Fix algorithm $\mathbb{A} = \texttt{KL-UCRL}$. Denote by $m(T)$ the number of episodes started by KL-UCRL up to time step T (hence, $1 \leqslant k \leqslant m(T)$).

By applying Azuma-Hoeffding inequality, as in the proof of (Jaksch et al., 2010, Theorem 2), we deduce that

$$\mathfrak{R}_{\mathbb{A},T} = Tg^\star - \sum_{t=1}^{T} r(s_t,a_t) \leqslant \sum_{s,a} N_T(s,a)(g^\star - \mu(s,a)) + \sqrt{\tfrac{1}{2}T\log(1/\delta)}\,,$$

with probability at least $1-\delta$. The regret up to time T can be decomposed as the sum of the regret incurred in various episodes. Let Δ_k denote the regret in episode k:

$$\Delta_k := \sum_{s,a} v_k(s,a)(g^\star - \mu(s,a))\,.$$

Therefore, Lemma 19 implies that with probability at least $1-3\delta$,

$$\mathfrak{R}_{\mathbb{A},T} \leqslant \sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \in \mathcal{M}_k\} + \sqrt{\tfrac{1}{2}T\log(1/\delta)}\,.$$

Next we derive an upper bound on the first term in the right-hand side of the above inequality. Consider an episode $k \geqslant 1$ such that $M \in \mathcal{M}_k$. The state-action pair (s,a) is considered as *sufficiently sampled* in episode k if its number of observations satisfies $N_k(s,a) \geqslant \ell_{s,a}$, with

$$\ell_{s,a} = \ell_{s,a}(T,\delta) := \max\Big\{\frac{128SB\max(\Psi^2,1)}{\varphi(s,a)^2},\, 32SB\Big(\frac{\log(D)}{\log(1/\gamma)}\Big)^2\Big\}, \quad \forall s,a\,,$$

where B is given in (16), and where γ denotes the contraction factor of the mapping induced by the transition probability matrix $P_\star$ of the optimal policy (γ can be determined as a function of elements of $P_\star$).

Now consider the case where all state-action pairs are sufficiently sampled in episode k (we analyse the case where some pairs are under-sampled (i.e., not sufficiently sampled) at the end of the proof). We have

$$|\tilde{\mu}_k(s,a) - \mu(s,a)| \leqslant |\tilde{\mu}_k(s,a) - \hat{\mu}_k(s,a)| + |\hat{\mu}_k(s,a) - \mu(s,a)| \leqslant 2\sqrt{\frac{C_\mu}{N_k(s,a)^+}}\,.$$

Hence,

$$\Delta_k = \sum_{s,a} v_k(s,a)(g^\star - \tilde{\mu}_k(s,a)) + \sum_{s,a} v_k(s,a)(\tilde{\mu}_k(s,a) - \mu(s,a))$$
$$\leqslant \sum_{s,a} v_k(s,a)(g^\star - \tilde{\mu}_k(s,a)) + 2\sqrt{C_\mu}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}}\,.$$

Let $\tilde{\mu}_k$ and $\widetilde{P}_k$ respectively denote the reward vector and transition probability matrix induced by the policy $\tilde{\pi}_k$ on $\tilde{M}_k$, i.e., $\tilde{\mu}_k := (\tilde{\mu}_k(s,\tilde{\pi}_k(s)))_s$, $\widetilde{P}_k := \big(\tilde{p}_k(s'|s,\tilde{\pi}_k(s))\big)_{s,s'}$. By Bellman optimality equation, $\tilde{g}_k - \tilde{\mu}_k(s,a) = (\widetilde{P}_k - \boldsymbol{I})\tilde{b}_k$. Hence, defining $v_k = (v_k(s,\tilde{\pi}_k(s))_s$ yields

$$\Delta_k \leqslant v_k(\widetilde{P}_k - I)\tilde{b}_k + (g^\star - \tilde{g}_k)v_k\mathbf{1} + 2\sqrt{C_\mu}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}}\,.$$

Now we use the following decomposition:

$$v_k(\widetilde{P}_k - I)\tilde{b}_k = \underbrace{v_k(\widetilde{P}_k - P_k)b^\star}_{F_1(k)} + \underbrace{v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star)}_{F_2(k)} + \underbrace{v_k(P_k - I)\tilde{b}_k}_{F_3(k)}\,.$$

Let $c = 1 + \sqrt{2}$. The following two lemmas provide upper bounds for $F_1(k)$ and $F_2(k)$:

Lemma 20 *For all $k \in \mathbb{N}$ such that $M \in \mathcal{M}_k$, with probability at least $1-\delta$, it holds that*

$$F_1(k) \leqslant (4+6\sqrt{2})\sqrt{SB}\sum_{s,a} v_k(s,a)\sqrt{\frac{\mathbf{V}^\star_{s,a}}{N_k(s,a)^+}} + 63\Psi S^{3/2}B^{3/2}\sum_{s,a}\frac{v_k(s,a)}{N_k(s,a)^+}\,.$$

Lemma 21 *Let $k \in \mathbb{N}$ be the index of an episode such that $M \in \mathcal{M}_k$. Assuming that $N_k(s,a) \geqslant \ell_{s,a}$ for all s,a, it holds that*

$$F_2(k) + (g^\star - \tilde{g}_k)v_k\mathbf{1} \leqslant \big(2\sqrt{32SB}+1\big)\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}}\,.$$

Analysis of Term F_3. Now we bound the term $\sum_{k=1}^{m(T)} F_3(k)$. To this end, similarly to the proof of (Jaksch et al., 2010, Theorem 2) and (Filippi et al., 2010, Theorem 1), we define the martingale difference sequence $(Z_t)_{t\geqslant 1}$, where $Z_t = (p(\cdot|s_t,a_t) - \mathbf{e}_{s_{t+1}})\tilde{b}_{k(t)}\mathbb{I}\{M \in \mathcal{M}_{k(t)}\}$ for $t \in \{t_k, t_{k+1}-1\}$, where $k(t)$ denotes the episode containing t. Note that for all t, $|Z_t| \leqslant 2D$. Now applying Azuma-Hoeffding inequality, we deduce that with probability at least $1-\delta$

$$\sum_{k=1}^{m(T)} F_3(k) \leqslant \sum_{t=1}^{T} Z_t + 2m(T)D$$
$$\leqslant D\sqrt{2T\log(1/\delta)} + 2DSA\log_2\big(\tfrac{8T}{SA}\big)\,.$$

The regret due to under-sampled state-action pairs. To analyze the under-sampled regime, where some state-action pair is not sufficiently sampled, we borrow some techniques from (Auer and Ortner, 2007). For any state-action pair (s,a), let $L_{s,a}$ denote the set of indexes of episodes in which (s,a) is chosen and yet (s,a) is under-sampled; namely $k \in L_{s,a}$ if $\tilde{\pi}_k(s) = a$ and $N_k(s,a) \leqslant \ell_{s,a}$. Furthermore, let $\tau_k(s,a)$ denote the length of such an episode.

Consider an episode $k \in L_{s,a}$. By Markov's inequality, with probability at least $\frac{1}{2}$, it takes at most $2T_M$ to reach state s from any state s' in k, where T_M is the mixing time of M. Let us divide episode k into $\lfloor \frac{\tau_k(s,a)}{2T_M} \rfloor$ sub-episodes, each with length greater than $2T_M$. It then follows that in each sub-episode, (s,a) is visited with probability at least $\frac{1}{2}$.

Using Hoeffding's inequality, if we consider n such sub-episodes, with probability at least $1 - \frac{\delta}{SA}$,

$$N(s,a) > n/2 - \sqrt{n \log(SA/\delta)}.$$

Now we find n that implies $N(s,a) < \ell_{s,a}$. Noting that $x \mapsto \frac{x}{2} - \sqrt{\alpha x}$ is increasing for $x \geqslant \alpha$, we have that for $n > 10 \max(\ell_{s,a}, \log(SA/\delta))$,

$$\begin{aligned} n/2 - \sqrt{n \log(SA/\delta)} &> 5 \max(\ell_{s,a}, \log(SA/\delta)) - \sqrt{10 \max(\ell_{s,a}, \log(SA/\delta)) \log(SA/\delta)} \\ &> \max(\ell_{s,a}, \log(SA/\delta)) \,. \end{aligned}$$

Hence, with probability at least $1 - \frac{\delta}{SA}$, it holds that

$$\sum_{k \in L_{s,a}} \Big\lfloor \frac{\tau_k(s,a)}{2T_M} \Big\rfloor \leqslant 10 \max(\ell_{s,a}, \log(SA/\delta)) \,.$$

Hence, the regret due to under-sampled state-action pairs can be upper bounded by

$$\begin{aligned} \sum_{s,a} \sum_{k \in L_{s,a}} \tau_k(s,a) &\leqslant 20 T_M \sum_{s,a} \max(\ell_{s,a}, \log(SA/\delta)) + 2T_M \sum_{s,a} |L_{s,a}| \\ &\leqslant 20 T_M \sum_{s,a} \max(\ell_{s,a}, \log(SA/\delta)) + 2T_M S^2 A^2 \log_2\big(\tfrac{8T}{SA}\big) \,, \end{aligned}$$

with probability at least $1 - \delta$. Here we used that $|L_{s,a}| \leqslant m(T)$.

Now applying Lemmas 20 and 21 together with the above bounds, and using the fact $C_\mu \leqslant B/1.99$, we deduce that with probability at least $1 - 3\delta$

$$\begin{aligned} \sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \in \mathcal{M}_k\} &\leqslant (4 + 6\sqrt{2}) \sqrt{SB} \sum_{s,a} \frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} \sqrt{\mathbf{V}^\star_{s,a}} \\ &\quad + (2\sqrt{32SB} + 3\sqrt{B} + 1) \sum_{s,a} \frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} \\ &\quad + 63 \Psi S^{3/2} B^{3/2} \sum_{s,a} \frac{v_k(s,a)}{N_k(s,a)^+} \\ &\quad + D\sqrt{2T \log(1/\delta)} + 2DSA \log_2\big(\tfrac{8T}{SA}\big) \\ &\quad + 20 T_M \sum_{s,a} \max(\ell_{s,a}, \log(SA/\delta)) + 2T_M S^2 A^2 \log_2\big(\tfrac{8T}{SA}\big) \,. \end{aligned}$$

To simplify the above bound, we will use Lemmas 23, 24, and 25 together with Jensen's inequality:

$$\sum_{k=1}^{m(T)} \sum_{s,a} \frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} \leqslant c \sum_{s,a} \sqrt{N_T(s,a)} \leqslant c\sqrt{SAT}\,,$$

$$\sum_{k=1}^{m(T)} \sum_{s,a} \frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} \sqrt{\mathbf{V}^\star_{s,a}} \leqslant c \sum_{s,a} \sqrt{\mathbf{V}^\star_{s,a} N_T(s,a)} \leqslant c\sqrt{T \textstyle\sum_{s,a} \mathbf{V}^\star_{s,a}}\,,$$

$$\sum_{k=1}^{m(T)} \sum_{s,a} \frac{v_k(s,a)}{N_k(s,a)^+} \leqslant 2 \sum_{s,a} \log(N_T(s,a)) + SA \leqslant 2SA \log\big(\tfrac{T}{SA}\big) + SA\,.$$

Putting everything together, we deduce that with probability at least $1-6\delta$,

$$\begin{aligned}
\mathfrak{R}_{\mathbb{A},T} &\leqslant \sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \in \mathcal{M}_k\} + \sqrt{\tfrac{1}{2} T \log(1/\delta)} \\
&\leqslant 31 \sqrt{S \sum_{s,a} \mathbf{V}^\star_{s,a} TB} + 35 S \sqrt{ATB} + (\sqrt{2}D+1)\sqrt{T\log(1/\delta)} \\
&\quad + 126 S^{5/2} A B^{5/2} \log\big(\tfrac{T}{SA}\big) + 2DSA \log_2\big(\tfrac{8T}{SA}\big) \\
&\quad + 20 T_M \sum_{s,a} \max(\ell_{s,a}, \log(SA/\delta)) + 2T_M S^2 A^2 \log_2\big(\tfrac{8T}{SA}\big) + 63 S^{5/2} A\,.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathfrak{R}_{\mathbb{A},T} &\leqslant 31 \sqrt{S \sum_{s,a} \mathbf{V}^\star_{s,a} TB} + 35 S \sqrt{ATB} + (\sqrt{2}D+1)\sqrt{T\log(1/\delta)} \\
&\quad + \widetilde{\mathcal{O}}\Big(SA(T_M SA + D + S^{3/2}) \log(T) \Big)\,.
\end{aligned}$$

Noting that $B = \mathcal{O}(\log(\log(T)/\delta))$ gives the desired scaling and completes the proof. □

Next we prove Lemmas 20 and 21.

C.1 Proof of Lemma 20

We have

$$F_1(k) = \underbrace{v_k(\widehat{P}_k - P_k)b^\star}_{G_1} + \underbrace{v_k(\widetilde{P}_k - \widehat{P}_k)b^\star}_{G_2}$$

Next we provide upper bounds for G_1 and G_2.

Term G_1. We have

$$\begin{aligned}
G_1 &= \sum_s v_k(s,\pi_k(s)) \sum_{s'} b^\star(s') \big(\hat{p}_k(s'|s,\pi_k(s)) - p(s'|s,\pi_k(s)) \big) \\
&\leqslant \sum_{s,a} v_k(s,a) \sum_{s'} b^\star(s') \big(\hat{p}_k(s'|s,a) - p(s'|s,a) \big)\,.
\end{aligned}$$

Fix $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Define the short-hands $p = p(\cdot|s,a)$, $\hat{p}_k = \hat{p}_k(\cdot|s,a)$, and $N_k^+ = N_k(s,a)^+$. Applying Corollary 11 (the first statement) and using the fact that $M \in \mathcal{M}_k$ give:

$$\begin{aligned}\sum_{s'} b^\star(s')(\hat{p}_k(s') - p(s')) &\leqslant \sqrt{2\mathbf{V}^\star_{s,a}\mathtt{KL}(\hat{p}_k, p)} + \frac{2}{3}\Psi\mathtt{KL}(\hat{p}_k, p) \\ &\leqslant \sqrt{8S\mathbf{V}^\star_{s,a}B/N_k^+} + \frac{8\Psi SB}{3N_k^+} \,.\end{aligned}$$

Therefore,

$$G_1 \leqslant \sqrt{8SB}\sum_{s,a} v_k(s,a)\sqrt{\mathbf{V}^\star_{s,a}/N_k(s,a)^+} + \frac{8}{3}\Psi SB\sum_{s,a} v_k(s,a)/N_k(s,a)^+ \,.$$

Term G_2. We have

$$G_2 \leqslant \sum_{s,a} v_k(s,a)\sum_{s'} b^\star(s')\big(\tilde{p}_k(s'|s,a) - \hat{p}_k(s'|s,a)\big) \,.$$

Fix $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Define the short-hands $\hat{p}_k = \hat{p}_k(\cdot|s,a)$, $\tilde{p}_k = \tilde{p}_k(\cdot|s,a)$, and $N_k^+ = N_k(s,a)^+$. An application of Lemma 12 and Lemma 13 gives

$$\begin{aligned}\sum_{s'} b^\star(s')(\tilde{p}_k(s') - \hat{p}_k(s')) &\leqslant \Big(\sqrt{\mathcal{V}_{\tilde{p}_k,\hat{p}_k}(b^\star)} + \sqrt{\mathcal{V}_{\hat{p}_k,\tilde{p}_k}(b^\star)}\Big)\sqrt{2\mathtt{KL}(\hat{p}_k,\tilde{p}_k)} + \Psi\mathtt{KL}(\hat{p}_k,\tilde{p}_k) \\ &\leqslant c\sqrt{2\mathbb{V}_{\hat{p}_k}(b^\star)\mathtt{KL}(\hat{p}_k,\tilde{p}_k)} + \Psi(1+3\sqrt{2S})\mathtt{KL}(\hat{p}_k,\tilde{p}_k) \,,\end{aligned}$$

where $c = 1+\sqrt{2}$. Note that when $M \in \mathcal{M}_k$, an application of Lemma 16 implies that, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{s'} b^\star(s')(\tilde{p}_k(s') - \hat{p}_k(s')) &\leqslant 4c\sqrt{S\mathbf{V}^\star_{s,a}B/N_k^+} + \frac{\Psi S^{3/2}B^{3/2}}{N_k^+}(12c\sqrt{2} + 12\sqrt{2} + 4/\sqrt{S}) \\ &\leqslant 4c\sqrt{S\mathbf{V}^\star_{s,a}B/N_k^+} + \frac{61\Psi S^{3/2}B^{3/2}}{N_k^+} \,,\end{aligned}$$

where we used that $S \geqslant 2$. Multiplying by $v_k(s,a)$ and summing over s,a yields

$$G_2 \leqslant 4c\sqrt{SB}\sum_{s,a} v_k(s,a)\sqrt{\mathbf{V}^\star_{s,a}/N_k(s,a)^+} + 61\Psi S^{3/2}B^{3/2}\sum_{s,a} v_k(s,a)/N_k(s,a)^+ \,.$$

The lemma follows by combing bounds on G_1 and G_2.

□

C.2 Proof of Lemma 21

Let $k \geqslant 1$ be the index of an episode such that $M \in \mathcal{M}_k$. Let $\tilde{\star} := \tilde{\star}_k$ denote the optimal policy in $\mathcal{M}_k$. The proof proceeds in three steps.

Step 1. We remark that by definition of the bias functions, it holds that

$$\begin{aligned}\tilde{b}_k - b^\star &= (g^\star - \tilde{g}_k)\mathbf{1} + \tilde{\mu}_k + \widetilde{P}_k b^\star - \mu_\star - P_\star b^\star + \widetilde{P}_k(\tilde{b}_k - b^\star) \\ &\leqslant (\tilde{g}_{\tilde{\star}} - \tilde{g}_k)\mathbf{1} + \tilde{\mu}_k - \mu_k + (\widetilde{P}_k - P_k)b^\star + \widetilde{P}_k(\tilde{b}_k - b^\star) - \varphi_k\,,\end{aligned}$$

where we define $\varphi_k(s) := \varphi(s, \tilde{\pi}_k(s))$ for all s. Defining

$$\xi_k(s) = 2\sqrt{C_\mu / N_k(s,\tilde{\pi}_k(s))^+}, \qquad \zeta_k(s) = \Psi\sqrt{32SB/N_k(s,\tilde{\pi}_k(s))^+}\,,$$

we obtain the following bound:

$$\tilde{b}_k - b^\star \leqslant \frac{1}{\sqrt{t_k}}\mathbf{1} + \xi_k + \zeta_k - \varphi_k + \widetilde{P}_k(\tilde{b}_k - b^\star)\,.$$

It is straightforward to check that the assumption $N_k(s,\tilde{\pi}_k(s)) \geqslant \ell_{s,\tilde{\pi}_k(s)}$ for all s implies

$$\tilde{b}_k - b^\star \leqslant \widetilde{P}_k(\tilde{b}_k - b^\star)\,. \tag{17}$$

Note also that $\varphi(s,\tilde{\pi}_k(s)) \geqslant 0$ since $\star$ is $b^\star$-improving.

On the other hand, it holds that

$$\begin{aligned}b^\star - \tilde{b}_{\tilde{\star}} &= (\tilde{g}_{\tilde{\star}} - g^\star)\mathbf{1} + \mu_\star + P_\star b^\star - \tilde{\mu}_{\tilde{\star}} - \widetilde{P}_{\tilde{\star}}\tilde{b}_{\tilde{\star}} \\ &\leqslant (\tilde{g}_{\tilde{\star}} - g^\star)\mathbf{1} + \mu_\star + P_\star b^\star - \mu_\star - P_\star \tilde{b}_{\tilde{\star}} \\ &= (\tilde{g}_{\tilde{\star}} - g^\star)\mathbf{1} + P_\star(b^\star - \tilde{b}_{\tilde{\star}})\,.\end{aligned}$$

Noting $P_\star\mathbf{1} = \mathbf{1}$, and since all entries of $P_\star$ are non-negative, we thus get for all $J \in \mathbb{N}$,

$$b^\star - \tilde{b}_{\tilde{\star}} \leqslant J(\tilde{g}_{\tilde{\star}} - g^\star)\mathbf{1} + P_\star^J(b^\star - \tilde{b}_{\tilde{\star}})\,.$$

Step 2. Let us now introduce $\mathcal{S}_s^+ = \{x \in \mathcal{S} : \widetilde{P}_k(s,x) > P_k(s,x)\}$ as well as its complementary set $\mathcal{S}_s^- = \mathcal{S} \setminus \mathcal{S}_s^+$. Using (17), $\tilde{b}_k - b^\star \leqslant 0$ so that

$$\begin{aligned}v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star) &= \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}} (\widetilde{P}_k(s,x) - P_k(s,x))(\tilde{b}_k(x) - b^\star(x)) \\ &\leqslant \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} \underbrace{(P_k(s,x) - \widetilde{P}_k(s,x))}_{\geqslant 0}(b^\star(x) - \tilde{b}_k(x))\,.\end{aligned}$$

We thus obtain

$$\begin{aligned}v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star) &\leqslant \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(b^\star(x) - \tilde{b}_{\tilde{\star}}(x)) \\ &\quad + \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(\tilde{b}_{\tilde{\star}}(x) - \tilde{b}_k(x)) \\ &\leqslant \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))[P_\star^J(b^\star - \tilde{b}_{\tilde{\star}})](x) \\ &\quad + \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(\tilde{b}_{\tilde{\star}}(x) - \tilde{b}_k(x)) \\ &\quad - J\sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(g^\star - \tilde{g}_{\tilde{\star}})\,.\end{aligned} \tag{18}$$

We thus get

$$\sum_s v_k(s,\tilde{\pi}_k(s))\Big((\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star)(s) + g^\star - \tilde{g}_{\tilde{\star}}\Big)$$
$$\leqslant \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))[P_\star^J(b^\star - \tilde{b}_{\tilde{\star}})](x) + \eta_k$$
$$+ \sum_s v_k(s,\tilde{\pi}_k(s))\Big[1 - J\sum_{x\in\mathcal{S}_s^-}(P_k(s,x) - \widetilde{P}_k(s,x))\Big](g^\star - \tilde{g}_{\tilde{\star}})\,, \qquad (19)$$

where $\eta_k := \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-}(P_k(s,x) - \widetilde{P}_k(s,x))(\tilde{b}_{\tilde{\star}}(x) - \tilde{b}_k(x))$ is controlled by the error of computing $\tilde{b}_k$ in episode k. In particular, for the considered variant of the algorithm,

$$\eta_k \leqslant \sum_s v_k(s,\tilde{\pi}_k(s))\|p(\cdot|s,\tilde{\pi}_k(s)) - \tilde{p}_k(\cdot|s,\tilde{\pi}_k(s))\|_1 \frac{1}{\sqrt{t_k}}$$
$$\leqslant \sqrt{32SB}\sum_s \frac{v_k(s,\tilde{\pi}_k(s))}{N_k(s,\tilde{\pi}_k(s))^+}\,,$$

where we used $t_k \geqslant N_k(s,\tilde{\pi}_k(s))$ for all s.

Step 3. It remains to choose J. To this end, we remark that the mapping induced by $P_\star$ is a contractive mapping, namely there exists some $\gamma < 1$ such that for any function f,

$$\mathbb{S}(P_\star f) \leqslant \gamma \mathbb{S}(f)\,.$$

Let us choose $J \geqslant \frac{\log(D)}{\log(1/\gamma)}$, so that with a simple upper bound, it comes

$$\sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))[P_\star^J(b^\star - \tilde{b}_{\tilde{\star}})](x)$$
$$\leqslant \sum_s v_k(s,\tilde{\pi}_k(s))\|p(\cdot|s,\tilde{\pi}_k(s)) - \tilde{p}_k(\cdot|s,\tilde{\pi}_k(s))\|_1 \frac{\mathbb{S}(P_\star^J(b^\star - \tilde{b}_{\tilde{\star}}^\star))}{2}$$
$$\leqslant \sum_s v_k(s,\tilde{\pi}_k(s))\|p(\cdot|s,\tilde{\pi}_k(s)) - \tilde{p}_k(\cdot|s,\tilde{\pi}_k(s))\|_1 D e^{-\log(D)}$$
$$\leqslant \sum_s v_k(s,\tilde{\pi}_k(s))\sqrt{\frac{32SB}{N_k(s,\tilde{\pi}_k(s))^+}}\,.$$

In the sequel, we take $J = \frac{\log(D)}{\log(1/\gamma)}$. This enables us to control the first two terms in (19) and it remains to control the term

$$\sum_s v_k(s,\tilde{\pi}_k(s))\Big[1 - J\sum_{x\in\mathcal{S}_s^-}(P_k(s,x) - \widetilde{P}_k(s,x))\Big](g^\star - \tilde{g}_{\tilde{\star}})\,.$$

In particular we would like to ensure that the term in brackets is non-negative, since in that case, it is multiplied by a term that is negative. To this end, we note that the term in brackets is lower bounded by

$$1 - J\|p(\cdot|s,\tilde{\pi}_k(s)) - \tilde{p}_k(\cdot|s,\tilde{\pi}_k(s))\|_1 \geqslant 1 - \frac{\log(D)}{\log(1/\gamma)}\sqrt{\frac{32SB}{N_k(s,\tilde{\pi}_k(s))^+}}\,,$$

and is thus guaranteed to be non-negative since

$$N_k(s,\tilde{\pi}_k(s)) \geqslant \ell_{s,\tilde{\pi}_k(s)} \geqslant 32SB\Big(\frac{\log(D)}{\log(1/\gamma)}\Big)^2 .$$

Putting together, we finally have shown that

$$\begin{aligned} v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star) + v_k(g^\star - \tilde{g}_k)\mathbf{1} &\leqslant v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star) + v_k(g^\star - \tilde{g}_{\tilde{\star}})\mathbf{1} + \frac{1}{\sqrt{t_k}} v_k \mathbf{1} \\ &\leqslant \big(2\sqrt{32SB}+1\big)\sum_s \frac{v_k(s,\tilde{\pi}_k(s))}{\sqrt{N_k(s,\tilde{\pi}_k(s))^+}} \\ &\leqslant \big(2\sqrt{32SB}+1\big)\sum_{s,a} \frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} , \end{aligned}$$

which completes the proof. □

Appendix D. Technical Lemmas

In this section we provide supporting lemmas for the regret analysis. The following lemma provides a local version of Pinsker's inequality for two probability distributions, which can be seen as the extension of (Garivier et al., 2016, Lemma 2) for the case of discrete probability measures.

Lemma 22 *Let P and Q be two probability distributions on a finite alphabet $\mathcal{X}$. Then,*

$$\mathtt{KL}(P,Q) \geqslant \frac{1}{2}\sum_{x:P(x)\neq Q(x)} \frac{(P(x)-Q(x))^2}{\max(P(x),Q(x))} .$$

Proof. The first and second derivatives of $\mathtt{KL}$ satisfy:

$$\begin{aligned} \frac{\partial}{\partial P(x)}\mathtt{KL}(P,Q) &= 1 + \log\frac{P(x)}{Q(x)}, \quad \forall x \in \mathcal{X}, \\ \frac{\partial^2}{\partial P(x)\partial P(y)}\mathtt{KL}(P,Q) &= \frac{\mathbb{I}\{x=y\}}{P(x)}, \quad \forall x,y \in \mathcal{X}. \end{aligned}$$

By Taylor's Theorem, there exists a probability vector Ξ, where $\Xi = tP + (1-t)Q$ for some $t \in (0,1)$, so that

$$\begin{aligned} \mathtt{KL}(P,Q) &= \mathtt{KL}(Q,Q) + \sum_x (P(x)-Q(x))\frac{\partial}{\partial P}\mathtt{KL}(Q,Q) \\ &\quad + \frac{1}{2}\sum_{x,y}(P(x)-Q(x))(P(y)-Q(y))\frac{\partial^2}{\partial P(x)\partial P(y)}\mathtt{KL}(\Xi,Q) \\ &= \sum_x (P(x)-Q(x)) + \sum_x \frac{(P(x)-Q(x))^2}{2\Xi(x)} \\ &\geqslant \sum_{x:P(x)\neq Q(x)} \frac{(P(x)-Q(x))^2}{2\max(P(x),Q(x))} , \end{aligned}$$

thus concluding the proof. □

Lemma 23 ((Jaksch et al., 2010, Lemma 19)) *Consider the sequence $(z_k)_{1\leqslant k\leqslant n}$ with $0 \leqslant z_k \leqslant Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$ for $k \geqslant 1$ and $Z_0 \geqslant 1$. Then,*

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \leqslant (\sqrt{2}+1)\sqrt{Z_n}\,.$$

Lemma 24 *Consider a sequence $(z_k)_{1\leqslant k\leqslant n}$ with $0 \leqslant z_k \leqslant Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$ for $k \geqslant 1$ and $Z_0 = z_1$. Then,*

$$\sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} \leqslant 2\log(Z_n) + 1\,.$$

Proof. We prove the lemma by induction over n. For $n = 1$, we have $z_1/Z_0 = 1$. Since $Z_1 = \max\{1, z_1\}$, it holds that $z_1/Z_0 \leqslant 2\log(Z_1) + 1$.

Now consider $n > 1$. By the induction hypothesis, it holds that $\sum_{k=1}^{n-1} z_k/Z_{k-1} \leqslant 2\log(Z_{n-1}) + 1$. Now it follows from the facts $z_n = Z_n - Z_{n-1}$ and $Z_{n-1} \leqslant Z_n \leqslant 2Z_{n-1}$ for $n \geqslant 2$, that

$$\begin{aligned}
\sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} &\leqslant 2\log(Z_{n-1}) + \frac{z_n}{Z_{n-1}} + 1 \\
&\leqslant 2\log(Z_{n-1}) + 2\frac{Z_n - Z_{n-1}}{Z_n} + 1 \\
&= 2\log(Z_{n-1}) + 2\Big(1 - \frac{1}{Z_n/Z_{n-1}}\Big) + 1 \leqslant 2\log(Z_n) + 1\,,
\end{aligned}$$

where the last inequality follows from $\log(x) \geqslant 1 - \frac{1}{x}$ valid for all $x \geqslant 1$ (see, e.g., (Topsøe, 2006)). This concludes the proof. □

Lemma 25 *Let $\alpha_i, \dots, \alpha_d$ be non-negative numbers and $T \geqslant 1$, and denote by V the optimal value of the following problem:*

$$\begin{aligned}
\max_x \quad & \sum_{i=1}^{d} \sqrt{\alpha_i x_i} \\
\text{s.t.} \quad & \sum_{i=1}^{d} x_i = T\,.
\end{aligned}$$

Then, $V = \sqrt{T\sum_{i=1}^{d} \alpha_i}$.

Proof. Introduce the Lagrangian

$$L(x, \lambda) = \sum_{i=1}^{d} \sqrt{\alpha_i x_i} + \lambda\Big(T - \sum_{i=1}^{d} x_i\Big)\,.$$

Writing KKT conditions, we observe that the optimal point $x_i^\star, i = 1, \dots, d$ satisfies

$$\frac{\alpha_i}{2\sqrt{x_i^\star}} - \lambda = 0, \ \forall i, \quad \text{and} \quad \sum_{i=1}^d x_i^\star - T = 0\,.$$

Hence, we obtain $x_i^\star = \alpha_i/(4\lambda^2)$. Plugging this into the equality constraint, it follows that $\lambda = \sqrt{\frac{1}{4T}\sum_{j=1}^d \alpha_j}$, thus giving $x_i^\star = \alpha_i T / \sum_{j=1}^d \alpha_j$. Therefore,

$$V = \sum_{i=1}^d \sqrt{\alpha_i x_i^\star} = \sum_{i=1}^d \frac{\alpha_i}{\sum_{j=1}^d \alpha_j} \sqrt{T \textstyle\sum_{j=1}^d \alpha_j} = \sqrt{T \textstyle\sum_{j=1}^d \alpha_j},$$

which completes the proof. □

Appendix E. Background Material on Undiscounted MDPs

In this section, we provide the proof of a number of standard results for the sake of self-containedness, and as we believe it helps get intuition on learning in MDPs.

E.1 Proof of Lemma 4

We provide below a short proof of this standard result for the sake of self-containedness.

The fundamental matrix. We first prove the relation involving the fundamental matrix. We note that by direct application of the relation $\overline{P}_\pi P_\pi = P_\pi \overline{P}_\pi = \overline{P}_\pi \overline{P}_\pi = \overline{P}_\pi$, it comes

$$\begin{aligned}(I - P_\pi + \overline{P}_\pi) b_\pi &= \sum_{t=1}^\infty (I - P_\pi)(P_\pi^{t-1} - \overline{P}_\pi)\mu_\pi + \underbrace{\overline{P}_\pi (P_\pi^{t-1} - \overline{P}_\pi)}_{0} \mu_\pi \\ &= \sum_{t=1}^\infty (I - P_\pi) P_\pi^{t-1} \mu_\pi - \underbrace{(I - P_\pi)\overline{P}_\pi}_{0} \mu_\pi = \sum_{t=1}^\infty \big(P_\pi^{t-1} - P_\pi^t\big)\mu_\pi\,.\end{aligned}$$

Thus, it remains to show that the latter sum equals $I - \overline{P}_\pi$. When P_π is aperiodic, then the limit $\lim_t P_\pi^t$ exists and is equal to $\overline{P}_\pi$. Thus, we easily get

$$\sum_{t=1}^\infty P_\pi^{t-1} - P_\pi^t = \lim_{T\to\infty} (I - P_\pi^T) = I - \lim_{T\to\infty} P_\pi^T = I - \overline{P}_\pi\,.$$

The general case is more intricate, and we refer to (Puterman, 2014).

Bellman equation. Now in order to obtain the Bellman equation, we simply note that

$$\begin{aligned}P_\pi b_\pi &= \sum_{t=1}^\infty (P_\pi^t - \overline{P}_\pi)\mu_\pi = \sum_{t=2}^\infty (P_\pi^{t-1} - \overline{P}_\pi)\mu_\pi \\ &= b_\pi - (I - \overline{P}_\pi)\mu_\pi = b_\pi - \mu_\pi + g_\pi\,.\end{aligned}$$

E.2 Value Iteration and Stopping Criterion

Definition 26 (Value iteration) *The value iteration procedure defines a sequence of functions $(u_n)_{n\in\mathbb{N}}$ and policies $(\pi_n)_{n\in\mathbb{N}}$ according to the following equations*

$$\forall n \in \mathbb{N} \begin{cases} u_{n+1}(s) = \max_{a\in\mathcal{A}} \mu(s,a) + (P_a u_n)(s)\,, \quad \text{where } u_0 = 0 \\ \pi_{n+1}(s) = \mathcal{U}\Big(\operatorname{Argmax}_{a\in\mathcal{A}} \mu(s,a) + (P_a u_n)(s) \Big)\,, \end{cases}$$

where $\mathcal{U}(\mathcal{B})$ denotes the uniform distribution over a set $\mathcal{B}$.

The following result is useful in order to better understand the effect of the classical stopping criterion used for the value iteration procedure.

Lemma 27 (Value and gain) *Let us assume that n is such that $\mathbb{S}(u_{n+1}-u_n) \leqslant \varepsilon$. Then it holds that*

$$g_\star - g_{\pi_{n+1}} \leqslant \varepsilon, \quad |u_{n+1} - u_n - g_\star| \leqslant \varepsilon, \quad \text{and} \quad |u_{n+1} - u_n - g_{\pi_{n+1}}| \leqslant \varepsilon\,.$$

Proof. We first show that the average gain satisfies $P_\pi g_\pi = g_\pi$ and

$$\forall n \in \mathbb{N}, \quad \overline{P}_{\pi_{n+1}}[u_{n+1}-u_n] \leqslant g_{\pi_{n+1}} \leqslant g_\star \leqslant \overline{P}_\star[u_{n+1}-u_n]\,.$$

Indeed, we note that since $\overline{P}_\star P_\star = \overline{P}_\star$, then for any function f, $g_\star = \overline{P}_\star[\mu_\star + P_\star f - f]$. Applying to the function u_n, it comes

$$\begin{aligned} g_\star &= \overline{P}_\star[\mu_\star + P_\star u_n - u_n] \\ &\leqslant \overline{P}_\star[\mu_{\pi_{n+1}} + P_{\pi_{n+1}} u_n - u_n] \\ &= \overline{P}_\star(u_{n+1} - u_n)\,, \end{aligned}$$

where in the second line, we used the maximal property of π_{n+1}. On the other hand, we use the equality

$$g_{\pi_{n+1}} = \overline{P}_{\pi_{n+1}}[\mu_{\pi_{n+1}} + P_{\pi_{n+1}} u_n - u_n] = \overline{P}_{\pi_{n+1}}(u_{n+1} - u_n)\,,$$

together with the fact that by optimality of $\star$, $g_\star \geqslant g_{\pi_{n+1}}$.

Thus, all in all it holds on the one hand

$$\begin{aligned} g_\star - g_{\pi_{n+1}} &\leqslant \overline{P}_\star[u_{n+1}-u_n] - \overline{P}_{\pi_{n+1}}[u_{n+1}-u_n] \\ &\leqslant \max_{s\in\mathcal{S}}(u_{n+1}-u_n)(s) - \min_{s\in\mathcal{S}}[u_{n+1}-u_n] = \mathbb{S}(u_{n+1}-u_n)\,. \end{aligned}$$

On the other hand, using similar steps,

$$\begin{aligned} 0 &\leqslant \overline{P}_\star[u_{n+1}-u_n] - g_\star \leqslant \max_{s\in\mathcal{S}}[u_{n+1}-u_n] - g_\star \\ &\leqslant \max_{s\in\mathcal{S}}[u_{n+1}-u_n] - \overline{P}_{\pi_{n+1}}[u_{n+1}-u_n] \leqslant \mathbb{S}(u_{n+1}-u_n)\,. \end{aligned}$$

Thus, for all $s \in \mathcal{S}$, $(u_{n+1}-u_n)(s) - g_\star \leqslant \varepsilon$. Likewise, we get the reverse inequality $0 \leqslant g_\star - \min_{s\in\mathcal{S}}(u_{n+1}-u_n)(s) \leqslant \mathbb{S}(u_{n+1}-u_n) \leqslant \varepsilon$. The last bound is immediate from the relation $g_{\pi_{n+1}} = \overline{P}_{\pi_{n+1}}(u_{n+1}-u_n)$. □

E.3 Pseudo-Regret

The following result relates the effective regret to the pseudo-regret

Lemma 28 (Effective regret to pseudo-regret reduction) *Let π be any stationary policy. Then it comes for all T,*

$$\begin{aligned}\mathbb{E}[\mathfrak{R}_{\pi,T}(s_1)] &= \big([P_\pi^{T-1} - I]b_\star\big)(s_1) + \sum_{s,a}\mathbb{E}[N_T(s,a)]\varphi(s,a) \\ &\leqslant D + \sum_{s,a}\mathbb{E}[N_T(s,a)]\varphi(s,a)\,.\end{aligned}$$

Proof. Since $g_\star$ is a constant function, it first comes

$$\mathbb{E}[\mathfrak{R}_{\pi,T}] = \sum_{t=1}^T \Big(g_\star - P_\pi^{t-1}\mu_\pi\Big) = \sum_{t=1}^T P_\pi^{t-1}\Big(g_\star - \mu_\pi\Big)\,.$$

Then, we note that by construction, it holds that $g_\star - \mu_\star = (P_\star - I)b_\star$. Introducing the sub-optimality gap $\varphi_\pi(s) := \varphi(s,\pi(s)) = \mu_\star(s) + (P_\star b_\star)(s) - \mu_\pi(s) - (P_\pi b_\star)(s)$, it then comes

$$g_\star - \mu_\pi = \varphi_\pi + g_\star - \mu_\star - P_\star b_\star + P_{\tilde\pi} b_\star = (P_\pi - I)b_\star + \varphi_\pi\,.$$

Thus far, we have we obtained that

$$\mathbb{E}[\mathfrak{R}_{\pi,T}] = \sum_{t=1}^T P_\pi^{t-1}\varphi_\pi + \sum_{t=1}^T P_\pi^{t-1}(P_\pi - I)b_\star = \sum_{t=1}^T P_\pi^{t-1}\varphi_\pi + (P_\pi^{T-1} - I)b_\star\,.$$

In order to conclude, we note that

$$\begin{aligned}(\sum_{t=1}^T P_{\pi_k}^{t-1}\varphi_{\pi_k})(s_1) &= \sum_{t=1}^T \mathbb{E}_{s_{t-1}}[\varphi_{\pi_k}(s_{t-1})] \\ &= \sum_{s,a}\varphi(s,a)\sum_{t=1}^T \mathbb{E}_{s_{t-1}}[\mathbb{I}\{s_{t-1}=s, \pi_k(s)=a\}] = \sum_{s,a}\varphi(s,a)\mathbb{E}[N_T(s,a)]\,.\end{aligned}$$

For the inequality, we use the simple bound $[P_\pi^{T-1} - I]b_\star \leqslant \|P_\pi^{T-1} - I\|_1 \frac{1}{2}\mathbb{S}(b_\star) \leqslant D$. Putting these together concludes the proof. □