KL-UCRL Revisited: Variance-Aware Regret Bound*

Mohammad Sadegh Talebi† SADEGH.TALEBI@INRIA.FR
INRIA Lille – Nord Europe, Villeneuve d'Ascq, France

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

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 provide a novel analysis of the KL-UCRL algorithm establishing a high-probability regret bound scaling as $\widetilde{\mathcal{O}}\Big(\sqrt{S\sum_{s,a}\mathbb{V}_{p(\cdot|s,a)}(b^\star)T}\Big)$ for this algorithm for ergodic MDPs, where S denotes the number of states and where $\mathbb{V}_{p(\cdot|s,a)}(b^\star)$ is the variance of the bias function of an optimal policy, 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, KL-UCRL

1. Introduction

In this paper, we consider Reinforcement Learning (RL) in an unknown environment modeled by a discrete Markov Decision Process (MDP). The learner interacts with the system in a single stream of observations, starting from an initial state without any reset, and wishes to maximize the long term average-reward. More formally, let $M = (\mathcal{S}, \mathcal{A}, \nu, P)$ denote an MDP where $\mathcal{S}$ and $\mathcal{A}$ respectively denote state-space and action-space (at any state), with respective cardinalities S and A. Furthermore, ν and P denote the reward function and the transition kernel, respectively. At time $t = 1$, the learner starts in some state $s_1 \in \mathcal{S}$. At each time step $t \in \mathbb{N}$, she 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, she 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 decision step begins. As the transition probabilities and reward functions are unknown, the learner needs to learn them by trying different actions and recording the realized rewards

*. This paper is based on (Talebi and Maillard, 2018).
†. The authors contributed equally.
‡. The authors contributed equally.

and state transitions. We refer to (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 $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping from states to actions. Following (Jaksch et al., 2010), we consider the following definition of regret:

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

where $a_t = \mathbb{A}(s_t, (\{s_{t'}, a_{t'}, r_{t'}\}_{t'<t})$, and where $g^\star$ and $b^\star$ respectively denote the maximal average-reward gain and the bias function of MDP M, which satisfy the following fixed-point equation, referred to as the *Bellman optimality equation* (see, e.g., (Puterman, 2014)):

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

To date, several algorithms implementing *optimism in the face of uncertainty* principle have been proposed for regret minimization in RL. These algorithms 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. (Burnetas and Katehakis, 1997) propose one of the first algorithms of this kind, which 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 norm. Due to the simplicity of the polytopic uncertainty model resulting from the confidence bounds defined by the L_1 norm, such a model is known to provide poor representations of underlying uncertainties; see (Nilim and El Ghaoui, 2005) and (Filippi et al., 2010). More precisely, 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 illustrations on pages 120–121 in (Filippi et al., 2010)).

Defining confidence bounds using the KL divergence can avoid these shortcomings. (Filippi et al., 2010) introduce the `KL-UCRL` algorithm that modifies the `UCRL2` algorithm of (Jaksch et al., 2010) by replacing L_1 norms with the KL divergences. 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 the KL divergence to define confidence bounds for transition function has remained largely unexplored. The goal of this paper is to investigate this gap.

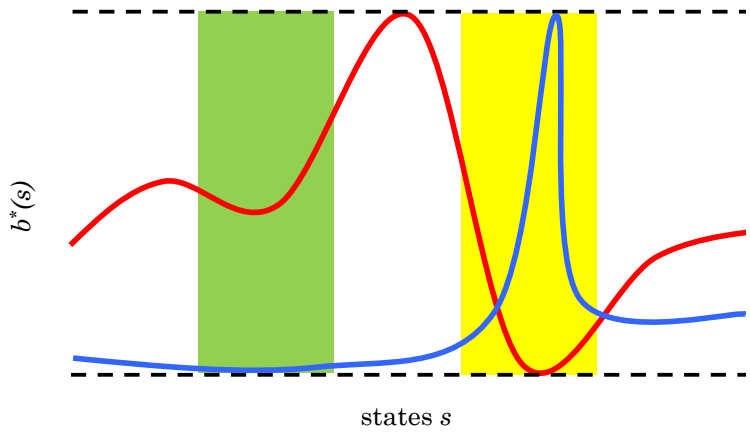

Figure 1: Two bias functions with the same span but different variances: MDP M (blue) vs. MDP M' (red)

In this paper we revisit the KL-UCRL algorithm and provide a new regret bound scaling as $\widetilde{\mathcal{O}}\Big(\sqrt{S\sum_{s,a}\mathbb{V}_{p(\cdot|s,a)}(b^\star)T} + D\sqrt{T}\Big)$ for ergodic MDPs with S states, A actions, and diameter D. Here, $\mathbb{V}_{p(\cdot|s,a)}(b^\star)$ 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 $\sqrt{\mathbb{V}_{p(\cdot|s,a)}(b^\star)} \leqslant D$. Interestingly, in several examples $\sqrt{\mathbb{V}_{p(\cdot|s,a)}(b^\star)} \ll D$. Our numerical experiments on typical MDPs further confirm that $\sqrt{S\sum_{s,a}\mathbb{V}_{p(\cdot|s,a)}(b^\star)}$ 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.

In order to further motivate the reported regret bound for KL-UCRL, we provide the following illustration. Consider two MDPs M and M' defined over the same state-space, whose corresponding bias functions have the same span. These bias functions, which are shown in Figure 1, exhibit different variation profiles over state-space: the bias function of M' (in red) has variation over all states whereas that of M (in blue) does not vary much except for a small part of the state-space. In order to see how such variation profiles may be captured by the variance of the bias function of the optimal policy, we consider two state-action pairs: $z^H := (s^H, a^H)$ whose support is shown in yellow, and $z^L := (s^L, a^L)$ whose support is shown in green. For the pair z^H, both quantities $\mathbb{V}_{p(\cdot|z^H)}(b^\star_M)$ and $\mathbb{V}_{p(\cdot|z^H)}(b^\star_{M'})$ may be high. In contrast, we expect $\mathbb{V}_{p(\cdot|z^L)}(b^\star_M)$ to be much smaller than $\mathbb{V}_{p(\cdot|z^L)}(b^\star_{M'})$.

We remark that the navigation cost (namely the amount of regret incurred due to the need to travel in the state-space) in M' could be much higher than that in M. Existing regret bounds only bound such a cost by the span of bias function or the diameter of MDP. Hence, existing bounds would conservatively provide the same bound on the navigation cost

for both MDPs M and M'. In contrast, our presented regret bound is relates the regret to the variance of the optimal bias function, which can capture such navigation costs more finely.

1.1 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 regret lower bound 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), (Maillard et al., 2014), and (Fruit et al., 2018).

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. The Recently, (Fruit et al., 2018) propose a method to efficiently implement a variant of the REGAL algorithm.

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 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:

$$\mathtt{KL}(\hat{p}_t(\cdot|s,a), p'(\cdot|s,a)) \leqslant C_p/N_t(s,a)\,, \tag{1}$$

$$|\hat{\mu}_t(s,a) - \mu'(s,a)| \leqslant \sqrt{C_\mu/N_t(s,a)}\,, \tag{2}$$

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 $C_\mu := C_\mu(T,\delta) = \log(4SA\log(T)/\delta)/1.99$. 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 $M_{\texttt{ext},k} = (\mathcal{S}, \mathcal{A} \times \mathcal{M}_k, \mu_{\texttt{ext}}, P_{\texttt{ext}})$, where for an extended action $a_{\texttt{ext}} = (a, M')$, it defines $\mu_{\texttt{ext}}(s, a_{\texttt{ext}}) = \mu'(s, a)$ and $p_{\texttt{ext}}(s'|s, a_{\texttt{ext}}) = p'(s'|s, a)$. Then, a $\frac{1}{\sqrt{t_k}}$-optimal extended policy $\pi_{\texttt{ext},k}$ is computed in the form $\pi_{\texttt{ext},k}(s) = (\tilde{M}_k, \tilde{\pi}_k(s))$, 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 $v_k = v_{k,t_{k+1}-1}$.

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 $v_k(s_t, a_t) < N_k(s_t, a_t)^+$ **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. Variance-Aware Regret Bounds

In this section, we present a regret upper bound for KL-UCRL that leverages the results presented in the previous section. In the following theorem, we provide our improved regret bounds for KL-UCRL:

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

$$\mathfrak{R}_{\texttt{KL-UCRL},T} = \mathcal{O}\Big(\big[\sqrt{S \sum_{s,a} \mathbb{V}_{p(\cdot|s,a)}(b^\star)} + D \big] \sqrt{T \log(\log(T)/\delta)} \Big)\,.$$

Remark 2 *Most steps in the proof of Theorem 1 carries over for 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.*

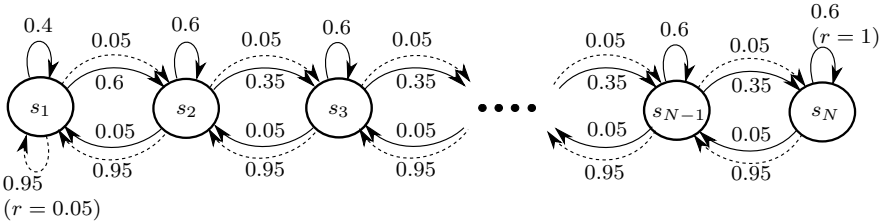

Figure 2: The N-state Ergodic *RiverSwim* MDP

S	$\mathbb{S}(b^\star)$	$\max_{s,a} \mathbb{V}_{p(\cdot\|s,a)}(b^\star)$	$\mathbb{S}(b^\star)\sqrt{SA}$	$\sqrt{\sum_{s,a} \mathbb{V}_{p(\cdot\|s,a)}(b^\star)}$
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*.

Illustrative numerical experiments. For the sake of illustration, we consider the *RiverSwim* MDP (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 2). 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 $\mathbb{S}(b^\star)\sqrt{SA} \leqslant D\sqrt{SA}$ grows very large as S increases, $\mathbb{V}_{p(\cdot|s,a)}(b^\star)$ is very small, on all tested environments, and does not change as S increases. Further, even on this simple environment, we see that $\sqrt{\sum_{s,a} \mathbb{V}_{p(\cdot|s,a)}(b^\star)}$ is an order or magnitude smaller than $\mathbb{S}(b^\star)\sqrt{SA}$. We believe that these computations highlight the fact that the regret bound of Theorem 1 captures a massive improvement over the initial analysis of `KL-UCRL` in (Filippi et al., 2010), and over alternative algorithms such as `UCRL2`.

4. Concentration Inequalities and The Kullback-Leibler Divergence

In this section we review our main technical tool for regret analysis, which we believe to be of independent interest beyond RL. To gently start, we first provide the *transportation lemma*; see, e.g., (Boucheron et al., 2013, Lemma 4.18):

Lemma 3 (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*

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

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 $b_\star$). Although the quantities $\varphi_{+,f}^{-1}$, $\varphi_{-,f}^{-1}$ may look complicated, 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 4 (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 4 cannot be handled, and that seems to be new (up to our knowledge):

Lemma 5 (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*

$$\forall P \ll Q, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leqslant \left(\sqrt{\mathcal{V}_{P,Q}(f)} + \sqrt{\mathcal{V}_{Q,P}(f)}\right)\sqrt{2\mathtt{KL}(P,Q)} + \mathbb{S}(f)\mathtt{KL}(P,Q)\,,$$

where $\mathcal{V}_{P,Q}(f) := \sum_{x\in\mathcal{X}:P(x)\geqslant Q(x)} P(x)(f(x)-\mathbb{E}_P[f])^2$.

In the following lemma, we related the operator $\mathcal{V}_{P,Q}$ to the variances of underlying distributions P and Q:

Lemma 6 *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*

$$\sqrt{\mathcal{V}_{P,Q}(f)} \leqslant \min\left\{\sqrt{\mathbb{V}_P(f)},\, \sqrt{2\mathbb{V}_Q(f)} + 3\mathbb{S}(f)\sqrt{|\mathcal{X}|\mathtt{KL}(Q,P)}\right\}.$$

As shown in the proof of Theorem 1, Corollary 4 and Lemmas 5 and 6 allow us to derive a regret bound scaling as $\widetilde{\mathcal{O}}\left(\sqrt{S\sum_{s,a}\mathbb{V}_{p(\cdot|s,a)}(b^\star)T}\right)$ for `KL-UCRL`.

5. Conclusion

In this paper, we revisited the analysis of `KL-UCRL` for ergodic MDPs, in order to make appear the local variance of the bias function of an optimal policy in the true MDP. Our findings show that, owing to properties of the Kullback-Leibler divergence, the leading term $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ obtained for the regret of `KL-UCRL` and `UCRL2` can be reduced to $\widetilde{\mathcal{O}}\left(\sqrt{S\sum_{s,a}\mathbb{V}_{p(\cdot|s,a)}(b^\star)T}\right)$. 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 optimal bias function may result in huge improvements.

References

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.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating markov decision processes. *arXiv preprint arXiv:1807.02373*, 2018.

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.

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.

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

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.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 770–805, 2018.

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. Concentration Inequalities

A.1 Proof of Lemma 3

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, $u = \varphi^{-1}_{+,f}(\mathtt{KL}(Q, P))$.

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, $v = \varphi^{-1}_{-,f}(\mathtt{KL}(Q, P))$. □

A.2 Proof of Corollary 4

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

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

□

A.3 Proof of Lemma 5

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

$$\begin{aligned}\mathbb{E}_Q[f] - \mathbb{E}_P[f] = &\sum_{x:Q(x)\geqslant P(x)} (f(x) - \mathbb{E}_Q[f])(Q(x) - P(x)) + \sum_{x:Q(x)<P(x)} (f(x) - \mathbb{E}_P[f])(Q(x) - P(x)) \\ &+ \sum_{x:P(x)>Q(x)} (\mathbb{E}_P[f] - \mathbb{E}_Q[f])(Q(x) - P(x))\,. \end{aligned} \tag{3}$$

The first term in the right-hand side of (3) is upper bounded as

$$\begin{aligned}\sum_{x:Q(x)\geqslant P(x)} (f(x) - \mathbb{E}_Q[f])(Q(x) - P(x)) &= \sum_{x:Q(x)\geqslant P(x)} \sqrt{Q(x)}(f(x) - \mathbb{E}_Q[f])\frac{Q(x) - P(x)}{\sqrt{Q(x)}} \\ &\overset{(a)}{\leqslant} \sqrt{\sum_{x:Q(x)\geqslant P(x)} Q(x)(f(x) - \mathbb{E}_Q[f])^2} \sqrt{\sum_{x:Q(x)\geqslant P(x)} \frac{(Q(x) - P(x))^2}{Q(x)}} \\ &\overset{(b)}{\leqslant} \sqrt{\mathcal{V}_{Q,P}(f)}\sqrt{2\mathtt{KL}(P,Q)}\,, \end{aligned} \tag{4}$$

where (a) follows from Cauchy-Schwarz inequality and (b) follows from Lemma 11.

Similarly, the second term in (3) satisfies

$$\sum_{x:Q(x)<P(x)} (f(x)-\mathbb{E}_P[f])(Q(x)-P(x)) = \sum_{x:Q(x)<P(x)} \sqrt{P(x)}(f(x)-\mathbb{E}_P[f])\frac{Q(x)-P(x)}{\sqrt{P(x)}} \leqslant \sqrt{\mathcal{V}_{P,Q}(f)}\sqrt{2\mathtt{KL}(P,Q)}\,. \tag{5}$$

Finally, we bound the last term in (3):

$$(\mathbb{E}_P[f]-\mathbb{E}_Q[f]) \sum_{x:P(x)>Q(x)} (Q(x)-P(x)) \overset{(a)}{=} \frac{1}{2}(\mathbb{E}_Q[f]-\mathbb{E}_P[f])\|P-Q\|_1 \leqslant \frac{1}{2}\mathbb{S}(f)\|P-Q\|_1^2 \overset{(b)}{\leqslant} \mathbb{S}(f)\mathtt{KL}(P,Q)\,, \tag{6}$$

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 (4), (5), and (6). □

A.4 Proof of Lemma 6

Statement (i) is a direct consequence of the definition of $\mathcal{V}_{P,Q}$. We next prove statement (ii). Observe that Lemma 11 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,

$$\begin{aligned}\mathcal{V}_{P,Q}(f) &= \sum_{x:P(x)\geqslant Q(x)} P(x)(f(x)-\mathbb{E}_P[f])^2 \\ &\leqslant \sum_{x:P(x)\geqslant Q(x)} Q(x)(f(x)-\mathbb{E}_P[f])^2 + \sqrt{2\mathtt{KL}(Q,P)} \sum_{x:P(x)\geqslant Q(x)} \sqrt{P(x)}(f(x)-\mathbb{E}_P[f])^2\,.\end{aligned} \tag{7}$$

The first term in the right-hand side of (7) is bounded as follows:

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

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 (7). First observe that

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

thanks to Cauchy-Schwarz inequality. Hence, the second term in (7) 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

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

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

Appendix B. Regret Upper Bound for KL-Ucrl

In this section, we provide the proof of the main result (Theorem 1). We will try to closely follow the notations used in the proof of (Jaksch et al., 2010, Theorem 2).

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 $\tilde{g}_k \geqslant \tilde{g}_{\tilde{\star}_k} - \frac{1}{\sqrt{t_k}}$.[1]

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 $t\in\mathbb{N}$,

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

where

$$B := B(T,\delta) = \log(2eS^2A\log(T)/\delta)\,, \\ G := G(T,\delta) = B + 1/\log(T)\,. \tag{8}$$

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

1. 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 $SA\log(T)$), with sufficient computational power such an algorithm could be practical.

Lemma 7 ((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, $\mathbb{P}(\forall t \leqslant T,\ M \in \mathcal{M}_t) \geqslant 1 - 2\delta$.

Let $\Psi := \mathbb{S}(b^\star)$ denote the span of the bias function. Next we prove the theorem.

Proof (of Theorem 1). 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 7 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 (8), 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,

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

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, $\tilde{g}_k - \tilde{\mu}_k(s,a) = (\widetilde{P}_k - I)\tilde{b}_k$. 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 8 *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 9 *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 $1-\delta$

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

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 8 and 9 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 12, 13, and 14 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 8 and 9.

B.1 Proof of Lemma 8

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 4 (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 5 and Lemma 6 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 10, stated below, 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.

Lemma 10 *For any episode $k \geqslant 1$ such that $M \in \mathcal{M}_k$, it holds that for any pair (s,a),*

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

B.2 Proof of Lemma 10

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 $1-\delta$,

$$\begin{aligned}
\mathbb{V}_{\hat{p}_k}(f) &= \sum_{s'} \hat{p}_k(s')(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2 \\
&\leqslant \sum_{s'} p(s')(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2 + \sqrt{\frac{2C_b}{N_k}} \sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2 + \frac{2C_b}{N_k} \sum_{s'} (f(s') - \mathbb{E}_{\hat{p}_k}[f])^2 \\
&\leqslant \underbrace{\sum_{s'} p(s')(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2}_{Z_1} + \sqrt{\frac{2C_b}{N_k}} \underbrace{\sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2}_{Z_2} + \frac{2C_b S \mathbb{S}(f)^2}{N_k} \,. \qquad (9)
\end{aligned}$$

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

$$(\mathbb{E}_p[f] - \mathbb{E}_{\hat{p}_k}[f])^2 \leqslant \mathbb{S}(f)^2 \|p - \hat{p}_k\|_1^2 \leqslant 2\mathbb{S}(f)^2 \mathtt{KL}(\hat{p}_k, p) \,. \qquad (10)$$

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 (10), 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 $1-\delta$. □

□

B.3 Proof of Lemma 9

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

$$\tilde{b}_k - b^\star \leqslant \widetilde{P}_k(\tilde{b}_k - b^\star)\,. \tag{11}$$

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 (11), $\tilde{b}_k - b^\star \leqslant 0$ 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

$$\begin{aligned} v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star) &\leqslant \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(b^\star(x) - \tilde{b}_{\tilde{\star}}(x)) \\ &\quad + \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(\tilde{b}_{\tilde{\star}}(x) - \tilde{b}_k(x)) \\ &\leqslant \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))[P_\star^J(b^\star - \tilde{b}_{\tilde{\star}})](x) \\ &\quad + \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(\tilde{b}_{\tilde{\star}}(x) - \tilde{b}_k(x)) \\ &\quad - J\sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(g^\star - \tilde{g}_{\tilde{\star}})\,. \end{aligned} \tag{12}$$

We thus get

$$\begin{aligned} &\sum_s v_k(s,\tilde{\pi}_k(s))\Big((\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star)(s) + g^\star - \tilde{g}_{\tilde{\star}}\Big) \\ &\qquad \leqslant \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))[P_\star^J(b^\star - \tilde{b}_{\tilde{\star}})](x) + \eta_k \\ &\qquad\quad + \sum_s v_k(s,\tilde{\pi}_k(s))\Big[1 - J\sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))\Big](g^\star - \tilde{g}_{\tilde{\star}})\,, \end{aligned} \tag{13}$$

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

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

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

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

In the sequel, we take $J = \frac{\log(D)}{\log(1/\gamma)}$. This enables us to control the first two terms in (13) and it remains to control the term

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

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

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

which completes the proof. □

Appendix C. 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 11 *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 12 ((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 13 *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 14 *Let $\alpha_i, \ldots, \alpha_d$ 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 $x_i^\star, i = 1, \ldots, d$ 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,

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

which completes the proof. □