Tightening Exploration in Upper Confidence Reinforcement Learning

Hippolyte Bourel, Odalric-Ambrym Maillard, Mohammad Sadegh Talebi

Tightening Exploration in Upper Confidence Reinforcement Learning

Hippolyte Bourel [1] **Odalric-Ambrym Maillard** [1] **Mohammad Sadegh Talebi** [2]

Abstract

The upper confidence reinforcement learning (UCRL2) algorithm introduced in (Jaksch et al., 2010) is a popular method to perform regret minimization in unknown discrete Markov Decision Processes under the average-reward criterion. Despite its nice and generic theoretical regret guarantees, this algorithm and its variants have remained until now mostly theoretical as numerical experiments in simple environments exhibit long burn-in phases before the learning takes place. In pursuit of practical efficiency, we present UCRL3, following the lines of UCRL2, but with two key modifications: First, it uses state-of-the-art time-uniform concentration inequalities to compute confidence sets on the reward and (component-wise) transition distributions for each state-action pair. Furthermore, to tighten exploration, it uses an adaptive computation of the support of each transition distribution, which in turn enables us to revisit the extended value iteration procedure of UCRL2 to optimize over distributions with reduced support by disregarding low probability transitions, while still ensuring near-optimism. We demonstrate, through numerical experiments in standard environments, that reducing exploration this way yields a substantial numerical improvement compared to UCRL2 and its variants. On the theoretical side, these key modifications enable us to derive a regret bound for UCRL3 improving on UCRL2, that for the first time makes appear notions of local diameter and local effective support, thanks to variance-aware concentration bounds.

[1]SequeL, Inria Lille – Nord Europe, Villeneuve d'Ascq, France [2]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark. Correspondence to: Hippolyte Bourel <hippolyte.bourel@ens-rennes.fr>, Odalric-Ambrym Maillard <odalric.maillard@inria.fr>, Mohammad Sadegh Talebi <sadegh.talebi@di.ku.dk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

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, infinite* stream of observations, starting from an initial state without any reset. More formally, let $M = (\mathcal{S}, \mathcal{A}, p, \nu)$ be an undiscounted MDP, where $\mathcal{S}$ denotes the discrete state-space with cardinality S, and $\mathcal{A}$ denotes the discrete action-space with cardinality A. p is the transition kernel such that $p(s'|s,a)$ denotes the probability of transiting to state s', starting from state s and executing action a. We denote by $\mathcal{K}_{s,a}$ the set of successor states of the state-action pair (s,a), that is $\mathcal{K}_{s,a} := \{x \in \mathcal{S} : p(x|s,a) > 0\}$, and further define $K_{s,a} := |\mathcal{K}_{s,a}|$. Finally, ν is a reward distribution function supported on $[0,1]$ with mean function denoted by μ. The interaction between the learner and the environment proceeds as follows. The learner starts in some state $s_1 \in \mathcal{S}$ at time $t = 1$. At each time step $t \in \mathbb{N}$, where the learner is in state s_t, she chooses an action $a_t \in \mathcal{A}$ based on s_t as well as her past decisions and observations. When executing action a_t in state s_t, the learner receives a random reward $r_t := r_t(s_t, a_t)$ drawn (conditionally) independently from distribution $\nu(s_t, a_t)$, and whose mean is $\mu(s_t, a_t)$. The state then transits to a next state $s_{t+1} \sim p(\cdot|s_t, a_t)$, and a new decision step begins. For background material on MDPs and RL, we refer to standard textbooks (Sutton & Barto, 1998; Puterman, 2014).

The goal of the learner is to maximize the *cumulative reward* gathered in the course of her interaction with the environment. The transition kernel p and the reward function ν are initially *unknown*, and so the learner has to learn them by trying different actions and recording the realized rewards and state transitions. The performance of the learner can be assessed through the notion of *regret*, which compares the cumulative reward gathered by an oracle, being aware of p and ν, to that gathered by the learner. Following (Jaksch et al., 2010), we define the regret of a learning algorithm $\mathbb{A}$ after T steps as $\mathfrak{R}(\mathbb{A}, T) := Tg^\star - \sum_{t=1}^{T} r_t$, where $g^\star$ denotes the *average-reward (or gain)* attained by an optimal policy. Alternatively, the objective of the learner is to minimize the regret, which entails balancing exploration and exploitation.

To date, several algorithms have been proposed in order to

minimize the regret based on the *optimism in the face of uncertainty* principle, a.k.a. the optimistic principle, originated from the seminal work (Lai & Robbins, 1985) on stochastic multi-armed bandits. 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-reward. A popular algorithm implementing the optimistic principle for the presented RL setup is UCRL2, which was introduced in the seminal work (Jaksch et al., 2010). UCRL2 achieves a non-asymptotic regret upper bound scaling as $\widetilde{\mathcal{O}}(DS\sqrt{AT})$[1] with high probability, in any communicating MDP with S states, A actions, and diameter D.[2] Jaksch et al. (2010) also report a regret lower bound scaling as $\Omega(\sqrt{DSAT})$, indicating that the above regret bound for UCRL2 is rate-optimal (up to logarithmic factors), i.e., it has a tight dependence on T, and can only be improved by a factor of, at most, $\sqrt{DS}$.

Since the advent of UCRL2, several of its variants have been presented in the literature; see, e.g., (Filippi et al., 2010; Bartlett & Tewari, 2009; Maillard et al., 2014; Fruit et al., 2018b; Talebi & Maillard, 2018). These variants mainly strive to improve the regret guarantee and/or empirical performance of UCRL2 by using improved confidence bounds or planning procedures. Although these algorithms enjoy delicate and strong theoretical regret guarantees, their numerical assessments have shown that they typically achieve a bad performance even for state-spaces of moderate size. In particular, they suffer from a long burn-in phase before the learning takes place, rendering them impractical for state-spaces of moderate size. It is natural to ask whether such a bad empirical performance is due to the main principle of UCRL2-style strategies, such as the optimistic principle, or to a not careful enough application of this principle. For instance, in a different, episodic and Bayesian framework, PSRL (Osband et al., 2013) has been reported to significantly outperform UCRL2 in numerical experiments. In this paper, we answer this question by showing, perhaps surprisingly, that a simple but crucial modification of UCRL2 that we call UCRL3 significantly outperforms other variants, while preserving (an improving on) their theoretical guarantees. Though our results do not imply that optimistic strategies are the best, they show that they can be much stronger competitors than vanilla UCRL2.

Contributions. We introduce UCRL3, a refined variant of UCRL2, whose design combines the following key elements: First, it uses tighter confidence bounds on components of the transition kernel (similarly to (Dann et al., 2017)) that are *uniform in time*, a property of independent interest for algorithm design in other RL setups; we refer to Section 3.1 for a detailed presentation. More specifically, for each component of a next-state transition distribution, it uses one time-uniform concentration inequality for $[0,1]$-bounded observations and one for Bernoulli distributions with a Bernstein flavor.

The second key design of the algorithm is a novel procedure, which we call NOSS[3], that adaptively computes an estimate of the support of transition probabilities of various state-action pairs. Such estimates are in turn used to compute a near-optimistic value and policy (Section 3.2). Combining NOSS with the Extended Value Iteration (EVI) procedure, used for planning in UCRL2, allows us to devise EVI-NOSS, which is a refined variant of EVI. This step is non-trivial as it requires to find a near-optimistic, as opposed to *fully optimistic*, policy. Furthermore, this enables us to make appear in the regret analysis notions of *local diameter* (Definition 1) as well as *local effective support* (Section 3.3), which in turn leads to a more problem-dependent regret bound. We define the local diameter below.

Definition 1 (Local diameter of state s) *Consider state $s \in \mathcal{S}$. For $s_1, s_2 \in \cup_{a\in\mathcal{A}}\mathcal{K}_{s,a}$ with $s_1 \neq s_2$, let $T^\pi(s_1, s_2)$ denote the number of steps it takes to get to s_2 starting from s_1 and following policy π. Then, the local diameter of MDP M for s, denoted by $D_s := D_s(M)$, is defined as*

$$D_s := \max_{s_1,s_2\in\cup_a\mathcal{K}_{s,a}} \min_\pi \mathbb{E}[T^\pi(s_1, s_2)].$$

On the theoretical side, we show in Theorem 1 that UCRL3 enjoys a regret bound scaling similarly to that established for the best variant of UCRL2 in the literature as in, e.g., (Fruit et al., 2018b). For better comparison with other works, we make sure to have an explicit bound including small constants for the leading terms. Thanks to a refined and careful analysis that we detail in the appendix, we also improve on the lower-order terms of the regret that we show should not be overlooked in practice. We provide in Section 4 a detailed comparison of the leading terms involved in several state-of-the-art algorithms to help better understand the behavior of these bounds. We also demonstrate through numerical experiments on standard environments that combining these refined, state-of-the-art confidence intervals together with EVI-NOSS yield a substantial improvement over UCRL2 and its variants. In particular, UCRL3 admits a burn-in phase, which is smaller than that of UCRL2 by an order of magnitude.

Related work. The study of RL under the average-reward criterion dates back to the seminal papers (Graves & Lai, 1997) and (Burnetas & Katehakis, 1997), followed by (Tewari & Bartlett, 2008). Among these studies, for the case

[1] The notation $\widetilde{\mathcal{O}}(\cdot)$ hides terms that are poly-logarithmic in T.

[2] Given an MDP M, the diameter $D := D(M)$ is defined as $D(M) := \max_{s\neq s'} \min_\pi \mathbb{E}[T^\pi(s, s')]$, where $T^\pi(s, s')$ denotes the number of steps it takes to get to s' starting from s and following policy π (Jaksch et al., 2010).

[3] Near-Optimistic Support Optimization

Algorithm	Regret bound	Comment
UCRL2 (Jaksch et al., 2010)	$\mathcal{O}\Big(DS\sqrt{AT\log(T/\delta)}\Big)$	
KL-UCRL (Filippi et al., 2010)	$\mathcal{O}\Big(DS\sqrt{AT\log(\log(T)/\delta)}\Big)$	Valid for fixed T provided as input.
KL-UCRL (Talebi & Maillard, 2018)	$\mathcal{O}\Big(\Big[D+\sqrt{S\sum_{s,a}(\mathbb{V}_{s,a}\vee 1)}\Big]\sqrt{T\log(\log(T)/\delta)}\Big)$	Restricted to ergodic MDPs.
SCAL$^+$ (QIAN et al., 2019)	$\mathcal{O}\Big(D\sqrt{\sum_{s,a}K_{s,a}T\log(T/\delta)}\Big)$	Without knowledge of the span.
UCRL2B (Fruit et al., 2019)	$\mathcal{O}\Big(\sqrt{D\sum_{s,a}K_{s,a}T\log(T)\log(T/\delta)}\Big)$	Note the extra $\sqrt{\log(T)}$ term.
UCRL3 **(This Paper)**	$\mathcal{O}\Big(\big(D+\sqrt{\sum_{s,a}(D_s^2L_{s,a}\vee 1)}\big)\sqrt{T\log(T/\delta)}\Big)$	
Lower Bound (Jaksch et al., 2010)	$\Omega\big(\sqrt{DSAT}\big)$	

Figure 1. Regret bounds of state-of-the-art algorithms for average-reward reinforcement learning. Here, $x\vee y$ denotes the maximum between x and y. For KL-UCRL, $\mathbb{V}_{s,a}$ denotes the variance of the optimal bias function of the true MDP, when the state is distributed according to $p(\cdot|s,a)$. For UCRL3, $L_{s,a}:=\big(\sum_{x\in\mathcal{S}}\sqrt{p(x|s,a)(1-p(x|s,a))}\big)^2$ denotes the local effective support of $p(\cdot|s,a)$.

of ergodic MDPs, Burnetas & Katehakis (1997) derive an asymptotic MDP-dependent lower bound on the regret, and provides an asymptotically optimal algorithm. Algorithms with finite-time regret guarantees and for wider classes of MDPs are presented in (Auer & Ortner, 2007; Jaksch et al., 2010; Bartlett & Tewari, 2009; Filippi et al., 2010; Maillard et al., 2014; Talebi & Maillard, 2018; Fruit et al., 2018a;b; Zhang & Ji, 2019; QIAN et al., 2019). Among these works, Filippi et al. (2010) introduce KL-UCRL, which is a variant of UCRL2 that uses the KL divergence to define confidence bounds. Similarly to UCRL2, KL-UCRL achieves a regret of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ in communicating MDPs. A more refined regret bound for KL-UCRL in ergodic MDPs is presented in (Talebi & Maillard, 2018). Bartlett & Tewari (2009) present REGAL and report a $\widetilde{\mathcal{O}}(D'S\sqrt{AT})$ regret with high probability in the larger class of weakly communicating MDPs, provided that the learner knows an upper bound D' on the span of the optimal bias function of the true MDP. Fruit et al. (2018b) present SCAL, which similarly to REGAL works in weakly communicating MDPs, but admits an efficient implementation. A similar algorithm called SCAL$^+$ is presented in (QIAN et al., 2019). Both SCAL and SCAL$^+$ admit a regret bound scaling as $\widetilde{\mathcal{O}}\Big(D\sqrt{\sum_{s,a}K_{s,a}T}\Big)$. In a recent work, Zhang & Ji (2019) present EBF achieving a regret of $\widetilde{\mathcal{O}}\big(\sqrt{HSAT}\big)$ assuming that the learner knows an upper bound H on the span of the optimal bias function of the true MDP.[4] However, EBF does not admit a computationally efficient implementation.

Another related line of works considers posterior sampling methods such as (Osband et al., 2013) inspired by Thompson sampling (Thompson, 1933). For average-reward RL, existing works on these methods report Bayesian regret bounds, with the exception of (Agrawal & Jia, 2017a), whose corrected regret bound, reported in (Agrawal & Jia, 2017b), scales as $O(DS\sqrt{AT}\log^3(T))$ and is valid for $T\geq S^4A^3$.

We finally mention that some studies consider regret minimization in MDPs in the *episodic* setting, with a fixed and known horizon; see, e.g., (Osband et al., 2013; Gheshlaghi Azar et al., 2017; Dann et al., 2017; Efroni et al., 2019; Zanette & Brunskill, 2019). Despite some similarity between the episodic and average-reward settings, the techniques developed for the episodic setting in these papers strongly rely on the fixed length of the episode. Hence, the tools in these papers do not directly carry over to the case of average-reward RL considered here (in particular, when closing the gap between lower and upper bounds is concerned).

In Figure 1, we report regret upper bounds of state-of-the-art algorithms for average-reward RL. We do not report REGAL and EBF in this table, as no corresponding efficient implementation is currently known. Furthermore, we stress that the presented regret bound for UCRL3 does not contradict the worst-case lower bound of $\Omega(\sqrt{DSAT})$ presented in (Jaksch et al., 2010). Indeed, for the worst-case MDP used to establish this lower bound in (Jaksch et al., 2010), both the local and global diameters coincide.

Notations. We introduce some notations that will be used throughout. For $x,y\in\mathbb{R}$, $x\vee y$ denotes the maximum between x and y. $\Delta_{\mathcal{S}}$ represents the set of all probability distributions defined on $\mathcal{S}$. For a distribution $p\in\Delta_{\mathcal{S}}$ and a vector-function $f=(f(s))_{s\in\mathcal{S}}$, we let Pf denote its application on f, defined by $Pf=\mathbb{E}_{X\sim p}[f(X)]$. We introduce $\Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}:=\{q:q(\cdot|s,a)\in\Delta_{\mathcal{S}},(s,a)\in\mathcal{S}\times\mathcal{A}\}$, and for $p\in\Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$, we define the corresponding operator P such that $Pf:s,a\mapsto\mathbb{E}_{X\sim p(\cdot|s,a)}[f(X)]$. We also introduce $\mathbb{S}(f)=\max_{s\in\mathcal{S}}f(s)-\min_{s\in\mathcal{S}}f(s)$.

Under a given algorithm and for a pair (s,a), we denote by $N_t(s,a)$ the total number of observations of (s,a) up to time t, and if (s,a) is not sampled yet by t, we set $N_t(s,a)=1$. Namely, $N_t(s,a):=1\vee\sum_{t'=1}^{t-1}\mathbb{I}\{(s_{t'},a_{t'})=(s,a)\}$. Let us define $\widehat{\mu}_t(s,a)$ as the empirical mean reward built using $N_t(s,a)$ i.i.d. samples from $\nu(s,a)$ (and whose mean is $\mu(s,a)$), and $\widehat{p}_t(\cdot|s,a)$ as the empirical distribution built using $N_t(s,a)$ i.i.d. observations from $p(\cdot|s,a)$.

[4] We remark that the universal constants of the leading term here are fairly large.

2. Background: The UCRL2 Algorithm

Before presenting UCRL3 in Section 3, we briefly present UCRL2 (Jaksch et al., 2010). To this end, let us introduce the following two sets: For each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$c_{t,\delta}^{\text{UCRL2}}(s,a) = \left\{ \mu' \in [0,1] : |\widehat{\mu}_t(s,a) - \mu'| \leq \sqrt{\frac{3.5 \log(\frac{2SAt}{\delta})}{N_t(s,a)}} \right\},$$

$$\mathcal{C}_{t,\delta}^{\text{UCRL2}}(s,a) = \left\{ p' \in \Delta_{\mathcal{S}} : \|\widehat{p}_t(\cdot|s,a) - p'\|_1 \leq \sqrt{\frac{14S \log(\frac{2At}{\delta})}{N_t(s,a)}} \right\}.$$

At a high level, UCRL2 maintains the set of MDPs $\mathcal{M}_{t,\delta} = \{\widetilde{M} = (\mathcal{S}, \mathcal{A}, \widetilde{p}, \widetilde{\nu})\}$, where for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\widetilde{p}(\cdot|s,a) \in \mathcal{C}_{t,\delta}^{\text{UCRL2}}(s,a)$ and $\widetilde{\mu}(s,a) \in c_{t,\delta}^{\text{UCRL2}}(s,a)$ (with $\widetilde{\mu}$ denoting the mean of $\widetilde{\nu}$). It then implements the optimistic principle by trying to compute $\overline{\pi}_t^+ = \operatorname{argmax}_{\pi:\mathcal{S}\to\mathcal{A}} \max\{g_\pi^M : M \in \mathcal{M}_{t,\delta}\}$, where g_π^M is the average-reward (or gain) of policy π in MDP M. This is carried out approximately by EVI that builds a near-optimal policy π_t^+ and an MDP $\widetilde{M}_t$ such that $g_{\pi_t^+}^{\widetilde{M}_t} \geq \max_{\pi, M \in \mathcal{M}_{t,\delta}} g_\pi^M - \frac{1}{\sqrt{t}}$. Finally, UCRL2 does not recompute π_t^+ at each time step. Instead, it proceeds in internal episodes, indexed by $k \in \mathbb{N}$, where a near-optimistic policy π_t^+ is computed only at the starting time of each episode. Letting t_k denote the starting time of episode k, the algorithm computes $\pi_k^+ := \pi_{t_k}^+$ and applies it until $t = t_{k+1} - 1$, where the sequence $(t_k)_{k\geq 1}$ is defined as follows: $t_1 = 1$, and for all $k > 1$,

$$t_k = \min \left\{ t > t_{k-1} : \max_{s,a} \frac{v_{t_{k-1}:t}(s,a)}{N_{t_{k-1}}(s,a)} \geq 1 \right\},$$

where $v_{t_1:t_2}(s,a)$ denotes the number of observations of pair (s,a) between time t_1 and t_2. The EVI algorithm writes as presented in Algorithm 1.

Algorithm 1 Extended Value Iteration (EVI)

Input: ε_t
Let $u_0 \equiv 0, u_{-1} \equiv -\infty, n = 0$
while $\mathbb{S}(u_n - u_{n-1}) > \varepsilon_t$ **do**
Compute $\begin{cases} \mu^+ : s,a \mapsto \max\{\mu' : \mu' \in c_{t,\delta}^{\text{UCRL2}}(s,a)\} \\ p_n^+ : s,a \mapsto \operatorname{argmax}\{P' u_n : p' \in \mathcal{C}_{t,\delta}^{\text{UCRL2}}(s,a)\} \end{cases}$
Update $\begin{cases} u_{n+1}(s) = \max\{\mu^+(s,a) + (P_n^+ u_n)(s,a) : a \in \mathcal{A}\} \\ \pi_{n+1}^+(s) \in \operatorname{Argmax}\{\mu^+(s,a) + (P_n^+ u_n)(s,a) : a \in \mathcal{A}\} \end{cases}$
$n = n + 1$
end while

3. The UCRL3 Algorithm

In this section, we introduce the UCRL3 algorithm, a variant of UCRL2 that relies on two main ideas motivated as follows:

(i) While being a theoretically appealing strategy, UCRL2 suffers from conservative confidence intervals, yielding an unacceptable empirical performance. Indeed, in the design of UCRL2, the random stopping times $N_t(s,a)$ are handled using simple union bounds, resulting in loose confidence bounds. The first modification we introduce has thus the same design as UCRL2, but replaces these confidence bounds with those derived from tighter time-uniform concentration inequalities. Furthermore, unlike UCRL2, UCRL3 does not use the L_1 norm to define the confidence bound of transition probabilities p. Rather it defines confidence bounds for each transition probability $p(s'|s,a)$, for each pair (s,a), similarly to SCAL or UCRL2B. Indeed, one drawback of L_1-type confidence bounds is that they require an upper bound on the size of the support of the distribution. Without further knowledge, only the conservative bound of S on the support can be applied. In UCRL2, this causes a factor S to appear inside the square-root, due to a union bound over 2^S terms. Deriving L_1-type confidence bounds adaptive to the support size seems challenging. In stark contrast, entry-wise confidence bounds can be used without knowing the support: when $p(\cdot|s,a)$ has a support much smaller than S, this may lead to a substantial improvement. Hence, UCRL3 relies on time-uniform Bernoulli concentration bounds (presented in Section 3.1 below).

(ii) In order to further tighten exploration, the second idea behind UCRL3 is to revisit EVI to compute a near-optimistic policy. Indeed, the optimization procedure used in EVI considers all plausible transition probabilities without support restriction, causing unwanted exploration. We introduce a novel value iteration procedure, called EVI-NOSS, which uses a restricted support optimization, where the considered support is chosen adaptively in order to retain near-optimistic guarantees.

We discuss these two modifications below in greater detail.

3.1. Confidence Bounds

We introduce the following high probability confidence sets for the mean rewards: For each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$c_{t,\delta_0}(s,a) = \left\{ \mu' \in [0,1] : |\widehat{\mu}_t(s,a) - \mu'| \leq b_{t,\delta_0/(SA)}^r(s,a) \right\},$$

where we introduced the notation

$$b_{t,\delta_0/(SA)}^r(s,a) := \max \left\{ \tfrac{1}{2}\beta_{N_t(s,a)}\left(\tfrac{\delta_0}{SA}\right), \sqrt{\frac{2\widehat{\sigma}_t^2(s,a)}{N_t(s,a)} \ell_{N_t(s,a)}\left(\tfrac{\delta_0}{SA}\right)} + \frac{7\ell_{N_t(s,a)}\left(\tfrac{\delta_0}{SA}\right)}{3N_t(s,a)} \right\},$$

with $\widehat{\sigma}_t^2(s,a)$ denoting the empirical variance of the reward function of (s,a) built using the observations gathered up to time t, and where $\ell_n(\delta) = \eta \log\left(\frac{\log(n)\log(\eta n)}{\log^2(\eta)\delta}\right)$ with

$\eta = 1.12$,[5] and $\beta_n(\delta) := \sqrt{\frac{2(1+\frac{1}{n})\log(\sqrt{n+1}/\delta)}{n}}$.

The definition of this confidence set is motivated by Hoeffding-type concentration inequalities for $1/2$-sub-Gaussian distributions[6], modified to hold for an arbitrary random stopping time, using the method of mixtures (a.k.a. the Laplace method) from (Peña et al., 2008). This satisfies by construction that

$$\mathbb{P}\Big(\exists t \in \mathbb{N}, (s,a) \in \mathcal{S}\times\mathcal{A},\ \mu(s,a) \notin c_{t,\delta_0}(s,a)\Big) \le 3\delta_0.$$

We recall the proof of this powerful result for completeness in Appendix A. Regarding the transition probabilities, we introduce the two following sets: For each $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$,

$$C_{t,\delta_0}(s,a,s') = \Bigg\{ q \in [0,1] : |\widehat{p}_t(s'|s,a) - q| \le \sqrt{\frac{2q(1-q)}{N_t(s,a)}\ell_{N_t(s,a)}\big(\tfrac{\delta_0}{SA}\big)} + \frac{\ell_{N_t(s,a)}\big(\frac{\delta_0}{SA}\big)}{3N_t(s,a)}, \text{ and } -\sqrt{\underline{g}(q)} \le \frac{\widehat{p}_t(s'|s,a)-q}{\beta_{N_t(s,a)}\big(\frac{\delta_0}{SA}\big)} \le \sqrt{g(q)} \Bigg\},$$

where $\underline{g}(p) = \begin{cases} g(p) & \text{if } p < 0.5 \\ p(1-p) & \text{else} \end{cases}$, with $g(p) = \frac{1/2-p}{\log(1/p-1)}$.

The first inequality comes from the Bernstein concentration inequality, modified using a peeling technique in order to handle arbitrary random stopping times. We refer the interested reader to (Maillard, 2019) for the generic proof technique behind this result. Dann et al. (2017) use similar proof techniques for Bernstein's concentration, however the resulting bounds are looser; we discuss this more in Appendix A.3. The last two inequalities are obtained by applying again the method of mixture for sub-Gaussian random variables, with a modification: Indeed, Bernoulli random variables are not only $1/2$-sub-Gaussian, but satisfy a stronger sub-Gaussian tail property, already observed in (Berend & Kontorovich, 2013; Raginsky & Sason, 2013). We discuss this in great detail in Appendix A.2.

UCRL3 finally considers the set of plausible MDPs $\mathcal{M}_{t,\delta} = \{\widetilde{M} = (\mathcal{S},\mathcal{A},\widetilde{p},\widetilde{\nu})\}$, where for each $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\widetilde{\mu}(s,a) \in c_{t,\delta_0}(s,a), \tag{1}$$
$$\widetilde{p}(\cdot|s,a) \in \mathcal{C}_{t,\delta_0}(s,a) = \Big\{ p' \in \Delta_{\mathcal{S}} : \forall s', p'(s') \in C_{t,\delta_0}(s,a,s') \Big\}.$$

Finally, the confidence level is chosen as[7] $\delta_0 = \delta/(3+3S)$.

[5] Any $\eta > 1$ is valid, and $\eta = 1.12$ yields a small bound.

[6] We recall that random variables bounded in $[0,1]$ are $\frac{1}{2}$-sub-Gaussian.

[7] When an upper bound $\overline{K}$ on $\max_{s,a} K_{s,a}$ is known, one could choose the confidence level $\delta_0 = \delta/(3+3\overline{K})$.

Lemma 1 (Time-uniform confidence bounds) *For any MDP with rewards bounded in $[0,1]$, mean reward function μ, and transition function p, for all $\delta \in (0,1)$, it holds*

$$\mathbb{P}\Big(\exists t \in \mathbb{N}, (s,a) \in \mathcal{S}\times\mathcal{A}, \ \mu(s,a) \notin c_{t,\delta_0}(s,a) \ \text{ or } \ p(\cdot|s,a) \notin \mathcal{C}_{t,\delta_0}(s,a) \Big) \le \delta\,.$$

3.2. Near-Optimistic Support-Adaptive Optimization

Last, we revisit the EVI procedure of UCRL2. When computing an optimistic MDP, EVI uses for each pair (s,a) an optimization over the set of all plausible transition probabilities (that is, over all distributions $q \in \mathcal{C}_{t,\delta}(s,a)$). This procedure comes with no restriction on the support of the considered distributions. In the case where $p(\cdot|s,a)$ is supported on a sparse subset of $\mathcal{S}$, this may however lead to computing an optimistic distribution with a large support, which in turn results in unnecessary exploration, and thereby degrades the performance. The motivation to revisit EVI is to provide a more adaptive way of handling sparse supports.

Let $\widetilde{\mathcal{S}} \subset \mathcal{S}$ and f be a given function (intuitively, the value function u_i at the current iterate i of EVI), and consider the following optimization problem for a specific state-action pair (s,a):

$$\overline{f}_{s,a}(\widetilde{\mathcal{S}}) = \max_{\widetilde{p}\in\mathcal{X}} \sum_{s'\in\widetilde{\mathcal{S}}} f(s')\widetilde{p}(s')\,, \quad \text{where} \tag{2}$$
$$\mathcal{X} = \Big\{ \widetilde{p} : \forall s' \in \widetilde{\mathcal{S}},\ \widetilde{p}(s') \in C_{t,\delta}(s,a,s') \text{ and } \sum_{s'\in\widetilde{\mathcal{S}}} \widetilde{p}(s') \le 1 \Big\}.$$

Remark 1 (Optimistic value) *The quantity $\overline{f}_{s,a}(\widetilde{\mathcal{S}})$ is conveniently defined by an optimization over positive measures whose mass may be less than one. The reason is that $p(\widetilde{\mathcal{S}}|s,a) \le 1$ in general. This ensures that $p(\cdot|s,a) \in \mathcal{X}$ indeed holds with high probability, and thus $\overline{f}_{s,a}(\widetilde{\mathcal{S}}) \ge \sum_{s'\in\widetilde{\mathcal{S}}} f(s')p(s'|s,a)$ as well.*

The original EVI procedure (Algorithm 1) computes $\overline{f}_{s,a}(\mathcal{S})$ for the function $f = u_i$ at each iteration i. When $p = p(\cdot|s,a)$ has a sparse support included in $\widetilde{\mathcal{S}}$, $C_{t,\delta}(s,a,s')$ often does not reduce to $\{0\}$ for $s' \notin \widetilde{\mathcal{S}}$, while one may prefer to force a solution with a sparse support. A naive way to proceed is to define $\widetilde{\mathcal{S}}$ as the empirical support (i.e., the support of $\widehat{p}_t(\cdot|s,a)$). Doing so, one however solves a *different* optimization problem than the one using the full set $\mathcal{S}$, which means we may lose the optimistic property (i.e., $\overline{f}_{s,a}(\widetilde{\mathcal{S}}) \ge \mathbb{E}_{X\sim p(\cdot|s,a)}[f(X)]$ may not hold) and get an uncontrolled error. Indeed, the following decomposition

$$\mathbb{E}_{X\sim p}[f(X)] \;=\; \sum_{s'\in\widetilde{\mathcal{S}}} f(s')p(s') + \underbrace{\sum_{s'\notin\widetilde{\mathcal{S}}} f(s')p(s')}_{\text{error}}\,,$$

shows that computing an optimistic value restricted on $\widetilde{\mathcal{S}}$ only upper bounds the first term in the right-hand side. The second term (the error term) needs to be upper bounded as well. Consider a pair (s,a), $t \geq 1$, and let $n := N_t(s,a)$. Provided that $\widetilde{S}$ contains the support of $\widehat{p}_t$, thanks to Bernstein's confidence bounds, it is easy to see[8] that the first term in the above decomposition contains terms scaling as $\widetilde{\mathcal{O}}(n^{-1/2})$, while the error term contains only terms scaling as $\widetilde{\mathcal{O}}(n^{-1})$. On the other hand, the error term sums $|\mathcal{S}\backslash\widetilde{\mathcal{S}}|$ many elements, which can be large in case p is sparse, and thus may even exceed $\overline{f}_{s,a}(\widetilde{\mathcal{S}})$ for small n. To ensure the error term does not dominate the first term, we introduce the Near-Optimistic Support-adaptive Optimization (NOSS) procedure, whose generic pseudo-code is presented in Algorithm 2. For instance, for a given pair (s,a) and time t, NOSS takes as input a target function $f = u_i$ (i.e., the value function at iterate i), the support $\widehat{\mathcal{S}}$ of the empirical distribution $\widehat{p}_t(\cdot|s,a)$, high-probability confidence sets $\mathcal{C} := \{C_{t,\delta}(s,a,s'), s' \in \mathcal{S}\}$, and a parameter $\kappa \in (0,1)$. It then adaptively augments $\widehat{\mathcal{S}}$ in order to find a set $\widetilde{S}$, whose corresponding value function $\overline{f}_{s,a}(\widetilde{\mathcal{S}})$ is near-optimistic, as formalized in the following lemma:

Algorithm 2 NOSS$(f, \widehat{\mathcal{S}}, \mathcal{C}, \kappa)$

Let $\widetilde{\mathcal{S}} = \widehat{\mathcal{S}} \cup \operatorname{argmax}_{s\in\mathcal{S}} f(s)$, and define $\overline{f}$ using f and confidence sets $\mathcal{C}$ (see (2)).
while $\overline{f}(\mathcal{S}\setminus\widetilde{\mathcal{S}}) \geq \min(\kappa, \overline{f}(\widetilde{\mathcal{S}}))$ **do**
  Let $\tilde{s} \in \operatorname{Argmax}_{s\notin\widetilde{\mathcal{S}}} f(s)$
  $\widetilde{\mathcal{S}} = \widetilde{\mathcal{S}} \cup \{\tilde{s}\}$
end while
return $\widetilde{\mathcal{S}}$

Algorithm 3 EVI-NOSS$(p, c, \mathcal{C}, N_{\max}, \varepsilon)$

Let $u_0 \equiv 0, u_{-1} \equiv -\infty, n = 0$
while $\mathbb{S}(u_n - u_{n-1}) > \varepsilon$ **do**
  Compute for all (s,a):
  $\widetilde{\mathcal{S}}_{s,a} = \text{NOSS}(u_n - \min_s u_n, \operatorname{supp}(p(\cdot|s,a)), \mathcal{C}, \kappa)$, with $\kappa = 10\mathbb{S}(u_n)|\operatorname{supp}(p(\cdot|s,a))|/N_{\max}^{3/2}$
  $\widetilde{\mathcal{C}}(s,a) = \{p' \in \mathcal{C}(s,a) : p'(x) = 0, \forall x \in \mathcal{S}\backslash\widetilde{\mathcal{S}}_{s,a}\}$
  Compute $\begin{cases}\mu^+ : s,a \mapsto \max\{\mu' : \mu' \in c(s,a)\}\\ p_n^+ : s,a \mapsto \operatorname{argmax}\{P'u_n : p' \in \widetilde{\mathcal{C}}(s,a)\}\end{cases}$
  Update $\begin{cases}u_{n+1}(s) = \max\{\mu^+(s,a) + (P_n^+ u_n)(s,a) : a \in \mathcal{A}\}\\ \pi_{n+1}^+(s) \in \operatorname{Argmax}\{\mu^+(s,a) + (P_n^+ u_n)(s,a) : a \in \mathcal{A}\}\end{cases}$
  $n = n + 1$
end while

Lemma 2 (Near-optimistic support selection) *Let $\widetilde{S}$ be a set output by NOSS. Then, with probability higher than $1-\delta$,*

$$\overline{f}_{s,a}(\widetilde{\mathcal{S}}) \geq \mathbb{E}_{X\sim p(\cdot|s,a)}[f(X)] - \min\{\kappa, \overline{f}_{s,a}(\widetilde{\mathcal{S}}), \overline{f}_{s,a}(\mathcal{S}\backslash\widetilde{\mathcal{S}})\}.$$

In other words, the value function $\overline{f}_{s,a}(\widetilde{\mathcal{S}})$ is near-optimistic.

Near-optimistic value iteration: The EVI-NOSS algorithm. In UCRL3, we thus naturally revisit the EVI procedure and combine the following step at each iterate n of EVI

$$p_n^+ : s,a \mapsto \operatorname{argmax}\{P'u_n, p' \in \mathcal{C}_{t,\delta}(s,a)\},$$

with NOSS: For a state-action pair (s,a), UCRL3 applies NOSS (Algorithm 2) to the function $u_n - \min_s u_n(s)$ (i.e., the relative optimistic value function) and empirical distribution $\widehat{p}_t(\cdot|s,a)$. We refer to the resulting algorithm as EVI-NOSS, as it combines EVI with NOSS, and present its pseudo-code in Algorithm 3. Finally, for iterate n in EVI-NOSS, we set the value of κ to

$$\kappa = \kappa_{t,n}(s,a) = \frac{\gamma\mathbb{S}(u_n)|\operatorname{supp}(\widehat{p}_t(\cdot|s,a))|}{\max_{s,a} N_t(s,a)^{2/3}}, \text{ where } \gamma = 10. \tag{3}$$

The scaling with the size of support and the span of the considered function is intuitive. The reason to further normalize by $\max_{s',a'} N_t(s',a')^{2/3}$ is to deal with the case when $N_t(s,a)$ is small: First, in the case of Bernstein's bounds, and since $\widetilde{\mathcal{S}}$ contains at least the empirical support, $\min\{\overline{f}_{s,a}(\widetilde{\mathcal{S}}), \overline{f}_{s,a}(\mathcal{S}\backslash\widetilde{\mathcal{S}})\}$ should essentially scale as $\widetilde{\mathcal{O}}(N_t(s,a)^{-1})$. Hence for pairs such that $N_t(s,a)$ is large, κ is redundant. Now for pairs that are not sampled a lot, $N_t(s,a)^{-1}$ may still be large even for large t, resulting in a possibly uncontrolled error. Forcing a $\max_{s,a} N_t(s,a)^{2/3}$ scaling ensures the near-optimality of the solution is preserved with enough accuracy to keep the cumulative regret controlled. This is summarized in the following lemma, whose proof is deferred to Appendix B.

Lemma 3 (Near-optimistic value iteration) *Using the stopping criterion $\mathbb{S}(u_{n+1} - u_n) \leq \varepsilon$, the EVI-NOSS algorithm satisfies that the average-reward (gain) g_{n+1}^+ of the policy π_{n+1}^+ and the MDP $\widetilde{M} = (\mathcal{S}, \mathcal{A}, \mu_{n+1}^+, p_{n+1}^+)$ computed at the last iteration $n+1$ is near-optimistic, in the sense that with probability higher than $1-\delta$, uniformly over all t, $g_{n+1}^+ \geq g^\star - \varepsilon - \overline{\kappa}$, where $\overline{\kappa} = \overline{\kappa}_{t,n} = \frac{\gamma\mathbb{S}(u_n)K}{\max_{s,a} N_t(s,a)^{2/3}}$.*

The pseudo-code of UCRL3 is provided in Algorithm 4.

3.3. Regret Bound of UCRL3

We are now ready to present a finite-time regret bound for UCRL3. Before presenting the regret bound in Theorem 1 below, we introduce the notion of *local effective support*. Given a pair (s,a), we define the *local effective support* $L_{s,a}$ of (s,a) as:

$$L_{s,a} := \Big(\sum_{x\in\mathcal{S}} \sqrt{p(x|s,a)\big(1-p(x|s,a)\big)}\Big)^2.$$

[8] They are of the form $p' - \widehat{p}_n(s') \leq a\sqrt{p'} + b$ where $a = \widetilde{\Theta}(n^{-1/2})$ and $b = \widetilde{\Theta}(n^{-1})$. This implies that for s' outside of the support of $\widehat{p}_n$, $p' \leq a\sqrt{p'} + b$, that is $p' \leq (\sqrt{a/4} + \sqrt{a/4+b})^2$.

Algorithm 4 UCRL3 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 $\delta_0 = \delta/(3+3S)$. Set $t_0 = 0$, $t = 1$, $k = 1$.
for episodes $k = 1, 2, \ldots$ **do**
  Set $t_k = t$
  Set $N_{t_k}(s,a) = N_{t_{k-1}}(s,a) + v_{k-1}(s,a)$ for all (s,a)
  Compute empirical estimates $\widehat{\mu}_{t_k}(s,a)$ and $\widehat{p}_{t_k}(\cdot|s,a)$ for all (s,a)
  Using Algorithm 3, compute

$$\pi^+_{t_k} = \texttt{EVI-NOSS}\Big(\widehat{p}_{t_k}, c_{t_k,\delta_0}, \mathcal{C}_{t_k,\delta_0}, \max_{s,a} N_{t_k}(s,a), \tfrac{1}{\sqrt{t_k}}\Big)$$

  Set $v_k(s,a) = 0$ for all (s,a)
  while $v_k(s_t, \pi^+_{t_k}(s_t)) < N_{t_k}(s_t, \pi^+_{t_k}(s_t))$ **do**
    Observe the current state s_t, play action $a_t = \pi^+_{t_k}(s_t)$, and receive reward r_t
    Set $v_k(s_t,a_t) = v_k(s_t,a_t) + 1$
    Set $t = t+1$
  end while
end for

In Lemma 4 below we show that $L_{s,a}$ is always controlled by the number $K_{s,a}$ of successor states of (s,a).[9] The lemma also relates $L_{s,a}$ to the Gini index of the transition distribution of (s,a), defined as $G_{s,a} := \sum_{x\in\mathcal{S}} p(x|s,a)(1-p(x|s,a))$.

Lemma 4 (Local effective support) *For any (s,a):*

$$L_{s,a} \le K_{s,a} G_{s,a} \le K_{s,a} - 1.$$

Theorem 1 (Regret of UCRL3) *With probability higher than $1-4\delta$, uniformly over all $T \ge 3$,*

$$\mathfrak{R}(\textit{UCRL3}, T) \le c\sqrt{T \log\left(\tfrac{6S^2A\sqrt{T+1}}{\delta}\right)} + 60DKS^{2/3}A^{2/3}T^{1/3} + \mathcal{O}\Big(DS^2A\log^2\left(\tfrac{T}{\delta}\right)\Big),$$

with $c = 5\sum_{s,a} D_s^2 L_{s,a} + 10\sqrt{SA} + 2D$. Therefore, the regret of UCRL3 asymptotically grows as

$$\mathcal{O}\left(\left[\sqrt{\sum_{s,a}\left(D_s^2 L_{s,a} \vee 1\right)} + D\right]\sqrt{T\log(\sqrt{T}/\delta)}\right).$$

We now compare the regret bound of UCRL3 against that of UCRL2B. As shown in Table 1, the latter algorithm attains a regret bound of $\mathcal{O}(\sqrt{D\sum_{s,a} K_{s,a} T \log(T)\log(T/\delta)})$. The two regret bounds are not directly comparable: The regret bound of UCRL2B depends on $\sqrt{D}$ whereas that of UCRL3 has a term scaling as D. However, the regret bound of UCRL2B suffers from an additional $\sqrt{\log(T)}$ term. Let us compare the two bounds for MDPs where quantities such as $K_{s,a}$, $L_{s,a}$, and D_s are *local* parameters in the sense that they do not scale with S, but where D could grow with S (one example is RiverSwim) — In other words, $K_{s,a}$, $L_{s,a}$, and D_s scale as $o(S)$. In such a case, comparing the two bounds boils down to comparing $(\sqrt{SA}+D)\sqrt{T\log(T)}$ against $\sqrt{DSAT\log^2(T)}$. When $T \ge \exp\left(\frac{(D+\sqrt{SA})^2}{DSA}\right)$ the effect of $\sqrt{\log(T)}$ is not small, and the regret bound of UCRL3 dominates that of UCRL2B. For instance, in 100-state RiverSwim, this happens for all $T \ge 71$. It has been left open whether this latter extra factor can be removed.

[9] We recall that for a pair (s,a), we define $\mathcal{K}_{s,a} := \mathrm{supp}(p(\cdot|s,a))$, and denote its cardinality by $K_{s,a}$.

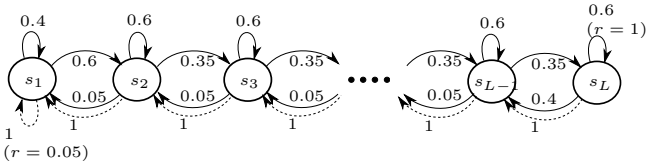

Figure 2. The L-state *RiverSwim* MDP

4. Numerical Experiments

In this section we provide illustrative numerical experiments that show the benefit of UCRL3 over UCRL2 and some of its popular variants. Specifically, we compare the empirical performance of UCRL3 against that of state-of-the-art algorithms including UCRL2, KL-UCRL, and UCRL2B — We also present further results in Appendix E, where we empirically compare UCRL3 against PSRL. For all algorithms, we set $\delta = 0.05$ and use the same tie-breaking rule. The full code and implementation details are made available to the community (see Appendix D for details).

In the first set of experiments, we consider the S-state RiverSwim environment (corresponding to the MDP shown in Figure 4). To better understand Theorem 1 in this environment, we report in Table 1 a computation of some of the key quantities appearing in the regret bounds, as well as the diameter D, for several values of S. We further provide in Table 2 a computation of the leading terms of several regret analyses. More precisely, for a given algorithm $\mathbb{A}$, we introduce $\overline{\mathfrak{R}}(\mathbb{A})$ to denote the regret bound normalized by $\sqrt{T\log(T/\delta)}$ *ignoring* universal constants. For instance, $\overline{\mathfrak{R}}(\texttt{UCRL2}) = D\sqrt{SA}$.[10] In Table 2, we compare $\overline{\mathfrak{R}}$ for various algorithms, for S-state RiverSwim for several values of S. We stress that $\overline{\mathfrak{R}}(\texttt{UCRL2B})$ grows with T unlike $\overline{\mathfrak{R}}$ for UCRL2, SCAL$^+$, and UCRL3. Note that even choosing a small value of $T = 100$, and ignoring universal constants (which disadvantage UCRL3), we get smaller regret bounds with UCRL3.

In Figure 3, we plot the regret under UCRL2, KL-UCRL, UCRL2B, and UCRL3 examined in the 6-state RiverSwim

[10] Ignoring universal constants here provides a more fair comparison; for example the final regret bound of UCRL2 has no second-order term at the expense of a rather large universal constant. Another reason in doing so is that for UCRL2B and SCAL$^+$, universal constants in their corresponding papers are not reported.

S	D	$\min_s D_s$	$\max_s D_s$	$\min_{s,a} L_{s,a}$	$\max_{s,a} L_{s,a}$
6	14.72	1.67	6.66	0	1.40
12	34.72	1.67	6.67	0	1.40
20	61.39	1.67	6.67	0	1.40
40	128.06	1.67	6.67	0	1.40
70	228.06	1.67	6.67	0	1.40
100	328.06	1.67	6.67	0	1.40

Table 1. Problem-dependent quantities for S-state *RiverSwim*

S	$\overline{\mathfrak{R}}$(UCRL2)	$\overline{\mathfrak{R}}$(SCAL$^+$)	$\overline{\mathfrak{R}}$(UCRL2B)	$\overline{\mathfrak{R}}$(UCRL3)
6	124.9	69.1	38.6	30.0
12	589.3	235.5	85.8	59.5
20	1736.3	542.2	148.5	94.9
40	7243.9	1609.6	305.3	176.9
70	22576	3802.4	540.0	293.6
100	46394	6544.7	775.3	407.6.2

Table 2. Comparison of the quantity $\overline{\mathfrak{R}}$ of various algorithms for S-state *RiverSwim*: $\overline{\mathfrak{R}}(\texttt{UCRL2}) = DS\sqrt{A}$, $\overline{\mathfrak{R}}(\texttt{SCAL}^+) = D\sqrt{\sum_{s,a} K_{s,a}}$, $\overline{\mathfrak{R}}(\texttt{UCRL2B}) = \sqrt{D\sum_{s,a} K_{s,a}\log(T)}$ for $T = 100$, and $\overline{\mathfrak{R}}(\texttt{UCRL3}) = \sqrt{\sum_{s,a}(D_s^2 L_{s,a} \vee 1)} + D$

environment. The curves show the results averaged over 50 independent runs along with the first and the third quantiles. We observe that UCRL3 achieves the smallest regret amongst these algorithms and significantly outperforms UCRL2, KL-UCRL, and UCRL2B (note the logarithmic scale). Figure 4 shows similar results on the larger 25-state RiverSwim environment.

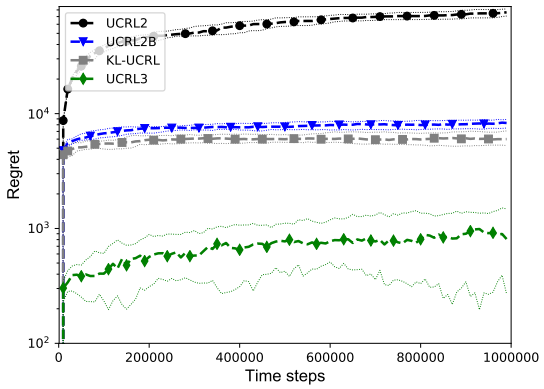

Figure 3. Regret for the 6-state *RiverSwim* environment

We further provide results in larger MDPs. We consider two frozen lake environments of respective sizes of 7×7 and 9×11 as shown in Figure 5, thus yielding MDPs with, respectively, $S = 20$ and $S = 55$ states (after removing walls). In such grid-worlds, the learner starts in the upper-left corner. A reward of 1 is placed in the lower-right corner, and the rest of states give no reward. Upon reaching the rewarding state, the learner is sent back to the initial state. The learner can perform 4 actions (when away from walls): Going up, left, down, or right. Under each, the learner moves in the chosen direction (with probability 0.7), stays in the same state (with probability 0.1), or goes in each of the two perpendicular directions (each with probability 0.1) – Walls act as reflectors moving back the leaner to the current state.

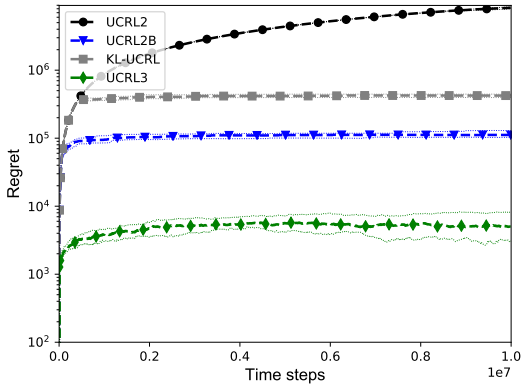

Figure 4. Regret for the 25-state *RiverSwim* environment

Remark 2 *Importantly,* UCRL2 *and its variants are generic purpose algorithms, and as such, are not aware of the specific structure of the MDP, such as being a grid-world. In particular, no prior knowledge is assumed on the support of the transition distributions by any of the algorithms, which makes it a highly non-trivial learning task, since the number of unknowns (i.e., problem dimension) is then* S^2A *(*$SA(S-1)$ *for the transition function, and* SA *for the rewards). For instance, a 4-room MDP is really seen as a problem of dimension* 1600 *by these algorithms, and a 2-room MDP as a problem of dimension* 12100*.*

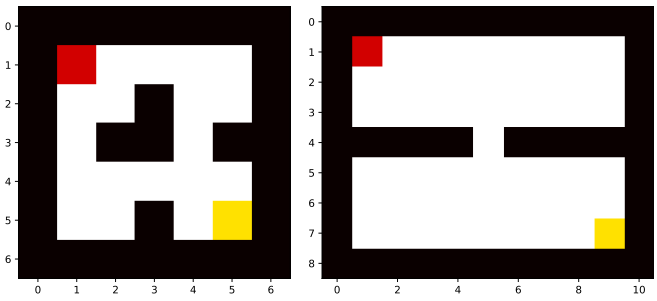

Figure 5. A 4-room (left) and a 2-room (right) grid-world environment, with 20 and 55 states: the starting state is shown in red, and the rewarding state is shown in yellow. From the yellow state, all actions bring the learner to the red state. Other transitions are noisy as in a *frozen-lake* environment.

Figures 6 (respectively, Figure 7) shows the regret performance of UCRL2, KL-UCRL, UCRL2B, and UCRL3 in the 2-room (respectively, 4-room) grid-world MDP. Finally, since all these algorithms are generic-purpose MDP learners, we provide numerical experiments in a large randomly-generated MDP consisting of 100 states and 3 actions, hence seen as being of dimension 3×10^4. UCRL3 still outperforms other state-of-the-art algorithms by a large margin consistently in all these environments. We provide below, an illustration of a randomly-generated MDP, with 15 states and 3 actions (blue, red, green). Such an MDP is a type of Garnet (Generalized Average Reward Non-stationary Envi-

ronment Test-bench) introduced in (Bhatnagar et al., 2009), in which we can specify the numbers of states and actions, the average size of the support of transition distributions, the sparsity of the reward function, as well as the minimal non-zero probability mass and minimal non-zero mean-reward.

Comparing UCRL3 against UCRL2B in experiments reveals that the gain achieved here is not only due to Bernstein's confidence intervals. Let us recall that on top of using Berstein's confidence intervals, UCRL3 also uses a refinement using sub-Gaussianity of Bernoulli distributions as well as the EVI-NOSS instead of EVI for planning. Experimental results verify that both tight confidence sets (see also Figure 11 in the appendix) and EVI-NOSS play an essential role in achieving small empirical regret.

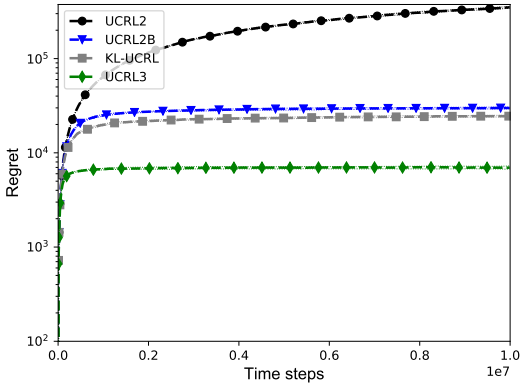

Figure 6. Regret for the 4-room environment

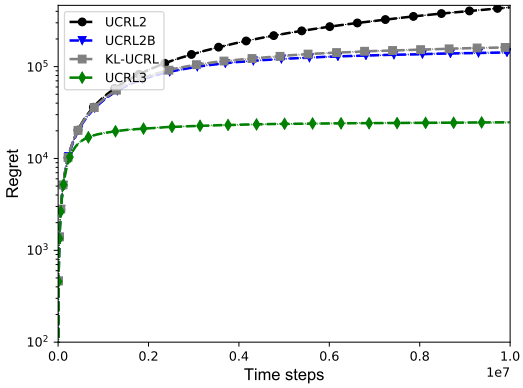

Figure 7. Regret for the 2-room environment

5. Conclusion

We studied reinforcement learning in finite Markov decision processes (MDPs) under the average-reward criterion, and introduced UCRL3, a refined variant of UCRL2 (Jaksch et al., 2010), that efficiently balances exploration and exploitation in communicating MDPs. The design of UCRL3 combines two main ingredients: (i) Tight time-uniform confidence bounds on individual elements of transition and reward functions, and (ii) a refined Extended Value Iteration procedure being adaptive to the support of transition function. We provided a non-asymptotic

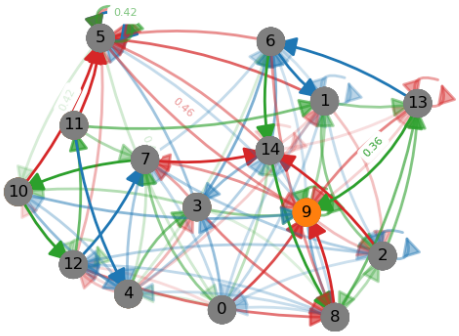

Figure 8. A randomly-generated MDP with 15 states: One color per action, shaded according to the corresponding probability mass, labels indicate mean reward, and the current state is highlighted in orange.

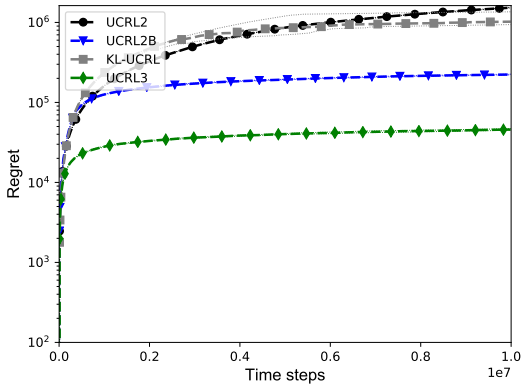

Figure 9. Regret in one 100-state randomly generated MDP

and high-probability regret bound for UCRL3 scaling as $\widetilde{\mathcal{O}}\big(\big(D + \sqrt{\sum_{s,a}(D_s^2 L_{s,a} \vee 1)}\big)\sqrt{T}\big)$, where D denotes the (global) diameter of the MDP, D_s denotes the *local* diameter of state s, and $L_{s,a}$ represents the local effective support of transition distribution for state-action pair (s,a). We further showed that $D_s \leq D$ and that $L_{s,a}$ is upper bounded by the number of successor states of (s,a), and therefore, the above regret bound improves on that of UCRL2. Through numerical experiments we showed that UCRL3 significantly outperforms existing variants of UCRL2 in standard environments. An interesting yet challenging research direction is to derive problem-dependent logarithmic regret bounds for UCRL3.

Acknowledgement

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 (the BADASS project). Part of this work was done while M. S. Talebi was a postdoctoral researcher in Inria Lille – Nord Europe.

References

Agrawal, S. and Jia, R. Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. In *Advances in Neural Information Processing Systems 30*, pp. 1184–1194, 2017a.

Agrawal, S. and Jia, R. Posterior sampling for reinforcement learning: Worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017b.

Auer, P. and Ortner, R. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems 19*, pp. 49–56, 2007.

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

Berend, D. and Kontorovich, A. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, 2013.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor–critic algorithms. *Automatica*, 45(11): 2471–2482, 2009.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

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

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30*, pp. 5711–5721, 2017.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pp. 12203–12213, 2019.

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

Fruit, R., Pirotta, M., and Lazaric, A. Near optimal exploration-exploitation in non-communicating Markov decision processes. In *Advances in Neural Information Processing Systems 31*, pp. 2994–3004, 2018a.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1578–1586, 2018b.

Fruit, R., Pirotta, M., and Lazaric, A. Improved analysis of UCRL2 with empirical Bernstein inequality. *Available at rlgammazero.github.io/docs/ucrl2b_improved.pdf*, 2019.

Gheshlaghi Azar, M., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 263–272, 2017.

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

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

Kearns, M. and Saul, L. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 311–319. Morgan Kaufmann Publishers Inc., 1998.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.

Maillard, O.-A. Mathematics of statistical sequential decision making. *Habilitation à Diriger des Recherches*, 2019.

Maillard, O.-A., Mann, T. A., and Mannor, S. How hard is my MDP? "the distribution-norm to the rescue". In *Advances in Neural Information Processing Systems 27*, pp. 1835–1843, 2014.

Osband, I., Russo, D., and Van Roy, B. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26*, pp. 3003–3011, 2013.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown Markov decision processes: A Thompson Sampling approach. In *Advances in Neural Information Processing Systems 30*, pp. 1333–1342, 2017.

Peña, V. H., Lai, T. L., and Shao, Q.-M. *Self-normalized processes: Limit theory and statistical applications*. Springer Science & Business Media, 2008.

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

QIAN, J., Fruit, R., Pirotta, M., and Lazaric, A. Exploration bonus for regret minimization in discrete and continuous average reward MDPs. In *Advances in Neural Information Processing Systems 32*, pp. 4891–4900, 2019.

Raginsky, M. and Sason, I. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.

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

Talebi, M. S. and Maillard, O.-A. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 770–805, 2018.

Tewari, A. and Bartlett, P. L. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Advances in Neural Information Processing Systems 20*, pp. 1505–1512, 2008.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pp. 285–294, 1933.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Technical Report*, 2003.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7304–7312, 2019.

Zhang, Z. and Ji, X. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, pp. 2823–2832, 2019.

A. Concentration Inequalities

A.1. Time-Uniform Laplace Concentration for Sub-Gaussian Distributions

Definition 2 (Sub-Gaussian observation noise) *A sequence $(Y_t)_t$ has conditionally σ-sub-Gaussian noise if*

$$\forall t, \forall \lambda \in \mathbb{R}, \quad \log \mathbb{E}[\exp\left(\lambda(Y_t - \mathbb{E}[Y_t|\mathcal{F}_{t-1}])\right)|\mathcal{F}_{t-1}] \leq \frac{\lambda^2\sigma^2}{2}\,,$$

where $\mathcal{F}_{t-1}$ denotes the σ-algebra generated by $Y_1, \dots, Y_{t-1}$.

Lemma 5 (Uniform confidence intervals) *Let $Y_1, \dots, Y_t$ be a sequence of t i.i.d. real-valued random variables with mean μ, such that $Y_t - \mu$ is σ-sub-Gaussian. Let $\mu_t = \frac{1}{t}\sum_{s=1}^{t} Y_s$ be the empirical mean estimate. Then, for all $\delta \in (0,1)$, it holds*

$$\mathbb{P}\left(\exists t \in \mathbb{N}, \quad |\mu_t - \mu| \geq \sigma\sqrt{\left(1+\frac{1}{t}\right)\frac{2\log\left(\sqrt{t+1}/\delta\right)}{t}}\right) \quad \leq \quad \delta\,.$$

The "Laplace" method refers to using the Laplace method of integration for optimization.

Proof of Lemma 5:

We introduce for a fixed $\delta \in (0,1)$ the random variable

$$\tau = \min\left\{t \in \mathbb{N} : \mu_t - \mu \geq \sigma\sqrt{\left(1+\frac{1}{t}\right)\frac{2\log\left(\sqrt{1+t}/\delta\right)}{t}}\right\}.$$

This quantity is a random stopping time for the filtration $\mathcal{F} = (\mathcal{F}_t)_t$, where $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$, since $\{\tau \leq m\}$ is $\mathcal{F}_m$-measurable for all m. We want to show that $\mathbb{P}(\tau < \infty) \leq \delta$. To this end, for any λ and t, we introduce the following quantity:

$$M_t^\lambda = \exp\left(\sum_{s=1}^{t}\left(\lambda(Y_s - \mu) - \frac{\lambda^2\sigma^2}{2}\right)\right).$$

By assumption, the centered random variables are σ-sub-Gaussian and it is immediate to show that $(M_t^\lambda)_{t\in\mathbb{N}}$ is a non-negative super-martingale that satisfies $\log \mathbb{E}[M_t^\lambda] \leq 0$ for all t. It then follows that $M_\infty^\lambda = \lim_{t\to\infty} M_t^\lambda$ is almost surely well-defined and so is M_τ^λ. Furthermore, using the fact that M_t^λ and $\{\tau > t\}$ are $\mathcal{F}_t$-measurable, it comes

$$\begin{aligned}
\mathbb{E}[M_\tau^\lambda] &= \mathbb{E}[M_1^\lambda] + \mathbb{E}[\sum_{t=1}^{\tau-1} M_{t+1}^\lambda - M_t^\lambda] \\
&= 1 + \sum_{t=1}^{\infty}\mathbb{E}[(M_{t+1}^\lambda - M_t^\lambda)\mathbb{I}\{\tau > t\}] \\
&= 1 + \sum_{t=1}^{\infty}\mathbb{E}[(\mathbb{E}[M_{t+1}^\lambda|\mathcal{F}_t] - M_t^\lambda)\mathbb{I}\{\tau > t\}] \\
&\leq 1\,.
\end{aligned}$$

The next step is to introduce the auxiliary variable $\Lambda \sim \mathcal{N}(0, \sigma^{-2})$, independent of all other variables, and study the quantity $M_t = \mathbb{E}[M_t^\Lambda|\mathcal{F}_\infty]$. Note that the standard deviation of Λ is σ^{-1} due to the fact we consider σ-sub-Gaussian random variables. We immediately get $\mathbb{E}[M_\tau] = \mathbb{E}[\mathbb{E}[M_\tau^\Lambda|\Lambda]] \leq 1$. For convenience, let $S_t = t(\mu_t - \mu)$. By construction

of M_t, we have

$$\begin{aligned} M_t &= \frac{1}{\sqrt{2\pi\sigma^{-2}}}\int_{\mathbb{R}} \exp\left(\lambda S_t - \frac{\lambda^2\sigma^2 t}{2} - \frac{\lambda^2\sigma^2}{2}\right)\mathrm{d}\lambda \\ &= \frac{1}{\sqrt{2\pi\sigma^{-2}}}\int_{\mathbb{R}} \exp\Big(- \left[\lambda\sigma\sqrt{\frac{t+1}{2}} - \frac{S_t}{\sigma\sqrt{2(t+1)}}\right]^2 + \frac{S_t^2}{2\sigma^2(t+1)}\Big)\mathrm{d}\lambda \\ &= \exp\left(\frac{S_t^2}{2\sigma^2(t+1)}\right)\frac{1}{\sqrt{2\pi\sigma^{-2}}}\int_{\mathbb{R}} \exp\Big(- \lambda^2\sigma^2\frac{t+1}{2}\Big)\mathrm{d}\lambda \\ &= \exp\left(\frac{S_t^2}{2\sigma^2(t+1)}\right)\frac{\sqrt{2\pi\sigma^{-2}/(t+1)}}{\sqrt{2\pi\sigma^{-2}}}\,. \end{aligned}$$

Thus, we deduce that

$$|S_t| = \sigma\sqrt{2(t+1)\log\left(\sqrt{t+1}M_t\right)}\,.$$

We conclude by applying a simple Markov's inequality:

$$\mathbb{P}\left(\tau|\mu_\tau - \mu| \geq \sigma\sqrt{2(\tau+1)\log\left(\sqrt{\tau+1}/\delta\right)}\right) = \mathbb{P}(M_\tau \geq 1/\delta) \leq \mathbb{E}[M_\tau]\delta\,.$$

□

A.2. Time-Uniform Laplace Concentration for Bernoulli Distributions

We now want to make use of the special structure of Bernoulli variables to derive refined time-uniform concentration inequalities. Let us first recall that if $(X_i)_{i\leq n}$ are i.i.d. according to a Bernoulli distribution $\mathcal{B}(p)$ with parameter $p \in [0,1]$, then it holds by the Chernoff-method that for all $\varepsilon \geq 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - p) \geq \varepsilon\right) \leq \exp\Big(- n\mathtt{kl}(p+\varepsilon, p)\Big)\,,$$

where $\mathtt{kl}(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ denotes the Bernoulli Kullback-Leibler divergence. The reverse map of the Cramér transform $\varepsilon \mapsto \mathtt{kl}(p+\varepsilon, p)$ is unfortunately not explicit, and one may consider Taylor's approximation of it to derive approximate but explicit high-probability confidence bounds. More precisely, the following has been shown (see (Kearns & Saul, 1998; Weissman et al., 2003; Berend & Kontorovich, 2013; Raginsky & Sason, 2013)):

Lemma 6 (Sub-Gaussianity of Bernoulli random variables) *For all $p \in [0,1]$, the left and right tails of the Bernoulli distribution are controlled in the following way*

$$\forall \lambda \in \mathbb{R}, \qquad \log \mathbb{E}_{X\sim\mathcal{B}(p)}\big[\exp(\lambda(X-p))\big] \leq \frac{\lambda^2}{2}g(p)\,,$$

where $g(p) = \frac{1/2-p}{\log(1/p-1)}$. The control of the right-tail can be further refined for $p \in [\frac{1}{2}, 1]$ as follows:

$$\forall \lambda \in \mathbb{R}^+, \qquad \log \mathbb{E}_{X\sim\mathcal{B}(p)}\big[\exp(\lambda(X-p))\big] \leq \frac{\lambda^2}{2}p(1-p)\,.$$

We note that the left and right tails are not controlled in a symmetric way. This yields, introducing the function $\underline{g}(p) = \begin{cases} g(p) & \text{if } p < 1/2 \\ p(1-p) & \text{otherwise} \end{cases}$, the following asymmetrical confidence set

Corollary 1 (Time-uniform Bernoulli concentration) *Let* $(X_i)_{i\le n} \overset{i.i.d.}{\sim} \mathcal{B}(p)$. *Then, for all* $\delta \in (0,1)$,

$$\mathbb{P}\left(\forall n \in \mathbb{N},\ -\sqrt{g(p)}\beta_n(\delta) \le \frac{1}{n}\sum_{i=1}^{n} X_i - p \le \sqrt{\underline{g}(p)}\beta_n(\delta)\right) \ge 1-2\delta\,,$$

where $\beta_n(\delta) := \sqrt{\frac{2}{n}\left(1+\frac{1}{n}\right)\log(\sqrt{n+1}/\delta)}$.

Proof of Corollary 1:

Let us introduce the following quantities

$$\begin{aligned}
\forall \lambda \in \mathbb{R}^+, \quad M_t^\lambda &= \exp\left(\sum_{s=1}^{t}\left(\lambda(X_s-p) - \frac{\lambda^2 \underline{g}(p)}{2}\right)\right),\\
\forall \lambda \in \mathbb{R}, \quad M_t'^\lambda &= \exp\left(\sum_{s=1}^{t}\left(\lambda(X_s-p) - \frac{\lambda^2 g(p)}{2}\right)\right).
\end{aligned}$$

Note that M_t^λ is a non-negative super-martingale for all $\lambda \in \mathbb{R}^+$, and $M_t'^\lambda$ is a non-negative super-martingale for all $\lambda \in \mathbb{R}$. Furthermore, $\mathbb{E}[M_t^\lambda] \le 1$ and $\mathbb{E}[M'^\lambda_t] \le 1$.

Let Λ be a random variable with density

$$f_p(\lambda) = \begin{cases} \frac{\exp(-\lambda^2 \underline{g}(p)/2)}{\int_{\mathbb{R}^+}\exp(-z^2\underline{g}(p)/2)\mathrm{d}z} = \sqrt{\frac{2\underline{g}(p)}{\pi}}\exp(-\lambda^2\underline{g}(p)/2) & \text{if } \lambda \in \mathbb{R}^+,\\ 0 & \text{else.}\end{cases}$$

Let $M_t = \mathbb{E}[M_t^\Lambda | \mathcal{F}_t]$ and note that

$$\begin{aligned}
M_t &= \sqrt{\frac{2\underline{g}(p)}{\pi}}\int_{\mathbb{R}^+}\exp\left(\lambda S_t - \frac{\lambda^2\underline{g}(p)t}{2} - \frac{\lambda^2\underline{g}(p)}{2}\right)\mathrm{d}\lambda\\
&= \sqrt{\frac{2\underline{g}(p)}{\pi}}\int_{\mathbb{R}^+}\exp\Big(-\Big[\lambda\sqrt{\frac{\underline{g}(p)(t+1)}{2}} - \frac{S_t}{\sqrt{2\underline{g}(p)(t+1)}}\Big]^2 + \frac{S_t^2}{2\underline{g}(p)(t+1)}\Big)\mathrm{d}\lambda\\
&= \exp\left(\frac{S_t^2}{2\underline{g}(p)(t+1)}\right)\sqrt{\frac{2\underline{g}(p)}{\pi}}\int_{\mathbb{R}^+}\exp\Big(-\Big(\lambda - \frac{S_t}{\underline{g}(p)(t+1)}\Big)^2\underline{g}(p)\frac{t+1}{2}\Big)\mathrm{d}\lambda\\
&= \exp\left(\frac{S_t^2}{2\underline{g}(p)(t+1)}\right)\sqrt{\frac{2\underline{g}(p)}{\pi}}\int_{c_t}\exp\Big(-\lambda^2\underline{g}(p)\frac{t+1}{2}\Big)\mathrm{d}\lambda \qquad \text{where } c_t = -\frac{S_t}{\underline{g}(p)(t+1)}\\
&\ge \exp\left(\frac{S_t^2}{2\underline{g}(p)(t+1)}\right)\sqrt{\frac{2\underline{g}(p)}{\pi}}\sqrt{\frac{\pi}{2(t+1)\underline{g}(p)}} \qquad \text{if } S_t \ge 0\\
&= \exp\left(\frac{S_t^2}{2\underline{g}(p)(t+1)}\right)\frac{1}{\sqrt{t+1}}\,.
\end{aligned}$$

Note also that M_t is still a non-negative super-martingale satisfying $\mathbb{E}[M_t] \le 1$ for all t. Likewise, considering Λ' to be a random variable with density

$$f_p'(\lambda) = \begin{cases} \frac{\exp(-\lambda^2 g(p)/2)}{\int_{\mathbb{R}^-}\exp(-z^2 g(p)/2)\mathrm{d}z} = \sqrt{\frac{2g(p)}{\pi}}\exp(-\lambda^2 g(p)/2) & \text{if } \lambda \in \mathbb{R}^-,\\ 0 & \text{else.}\end{cases}$$

Introducing $M'_t = \mathbb{E}[{M'}_t^{\Lambda'}|\mathcal{F}_t]$, it comes

$$M'_t \quad \geq \quad \exp\left(\frac{S_t^2}{2g(p)(t+1)}\right)\frac{1}{\sqrt{t+1}} \quad \text{if } S_t \leq 0.$$

M'_t is a non-negative super-martingale satisfying $\mathbb{E}[M_t] \leq 1$ for all t. Thus, we deduce that

$$\frac{|S_t|}{t} \leq \begin{cases} \sqrt{2\underline{g}(p)\frac{(1+1/t)}{t}\log\left(M_t\sqrt{1+t}\right)} & \text{if } S_t \geq 0 \\ \sqrt{2g(p)\frac{(1+1/t)}{t}\log\left(M'_t\sqrt{1+t}\right)} & \text{if } S_t \leq 0\,, \end{cases}$$

which implies

$$-\sqrt{2g(p)\frac{(1+1/t)}{t}\log\left(M'_t\sqrt{1+t}\right)} \leq \frac{S_t}{t} \leq \sqrt{2\underline{g}(p)\frac{(1+1/t)}{t}\log\left(M_t\sqrt{1+t}\right)}\,.$$

Combining the previous steps, we thus obtain for each $\delta \in (0,1)$,

$$\begin{aligned} \mathbb{P}\bigg(\exists t, \frac{S_t}{t} \geq \sqrt{2\underline{g}(p)\frac{(1+1/t)}{t}\log\left(\sqrt{1+t}/\delta\right)} \text{ or } \frac{S_t}{t} \leq -\sqrt{2g(p)\frac{(1+1/t)}{t}\log\left(\sqrt{1+t}/\delta\right)}\bigg) \\ \leq \quad \mathbb{P}\Big(\exists t, M_t \geq 1/\delta \text{ or } M'_t \geq 1/\delta\Big) \\ \leq \quad \mathbb{P}\big(\exists t, M_t \geq 1/\delta\big) + \mathbb{P}\big(\exists t, M'_t \geq 1/\delta\big) \\ \leq \quad \delta(\mathbb{E}[\max_t M_t] + \mathbb{E}[\max_t M'_t]) \\ \leq \quad 2\delta\,. \end{aligned}$$

The last inequality holds by an application of Doob's property for non-negative super-martingales, and using that $\mathbb{E}[M_1] = \mathbb{E}[M'_1] = 1$. □

A.3. Comparison of Time-Uniform Concentration Bounds

In this section, we give additional details about the concentration inequalities used to derive the confidence bounds in UCRL3. We first present the following result from (Maillard, 2019), which makes use of a generic peeling approach:

Lemma 7 ((Maillard, 2019, Lemma 2.4)) *Let $Z = (Z_t)_{t\in\mathbb{N}}$ be a sequence of random variables generated by a predictable process, and $\mathcal{F} = (\mathcal{F}_t)_t$ be its natural filtration. Let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be a convex upper-envelope of the cumulant generating function of the conditional distributions with $\varphi(0) = 0$, and let $\varphi_\star$ denote its Legendre-Fenchel transform, that is:*

$$\begin{aligned} \forall \lambda \in \mathcal{D}, \forall t, \qquad & \log \mathbb{E}\left[\exp\left(\lambda Z_t\right)|\mathcal{F}_{t-1}\right] \leq \varphi(\lambda)\,, \\ \forall x \in \mathbb{R}, \qquad & \varphi_\star(x) = \sup_{\lambda\in\mathbb{R}}(\lambda x - \varphi(\lambda))\,, \end{aligned}$$

where $\mathcal{D} = \{\lambda \in \mathbb{R} : \forall t, \log \mathbb{E}\left[\exp(\lambda Z_t)|\mathcal{F}_{t-1}\right] \leq \varphi(\lambda) < \infty\}$. Assume that $\mathcal{D}$ contains an open neighborhood of 0. Let $\varphi^{-1}_{\star,+} : \mathbb{R} \to \mathbb{R}_+$ (resp. $\varphi^{-1}_{\star,-}$) be its reverse map on $\mathbb{R}_+$ (resp. $\mathbb{R}_-$), that is

$$\varphi^{-1}_{\star,-}(z) := \sup\{x \leq 0 : \varphi_\star(x) > z\} \quad \textit{and} \quad \varphi^{-1}_{\star,+}(z) := \inf\{x \geq 0 : \varphi_\star(x) > z\}\,.$$

Let N_n be a stopping time that is $\mathcal{F}$-measurable and almost surely bounded by n. Then, for all $\eta \in (1,n]$ and $\delta \in (0,1)$,

$$\mathbb{P}\left[\frac{1}{N_n}\sum_{t=1}^{N_n} Z_t \geq \varphi^{-1}_{\star,+}\left(\frac{\eta}{N_n}\log\left(\left\lceil\frac{\log(n)}{\log(\eta)}\right\rceil\frac{1}{\delta}\right)\right)\right] \leq \delta\,,$$

$$\mathbb{P}\left[\frac{1}{N_n}\sum_{t=1}^{N_n} Z_t \leq \varphi^{-1}_{\star,-}\left(\frac{\eta}{N_n}\log\left(\left\lceil\frac{\log(n)}{\log(\eta)}\right\rceil\frac{1}{\delta}\right)\right)\right] \leq \delta\,.$$

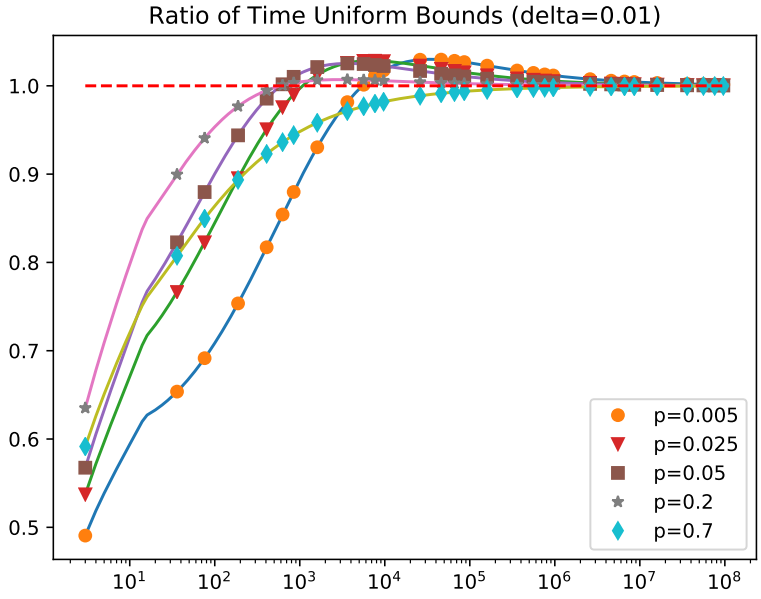

Figure 10. Plot of $n \mapsto r(p, n, \delta)$ for several values of p, with $\delta = 0.01$. We plot the horizontal line $r(p, n, \delta) = 1$ for reference: Above this line, the second Bernstein bound is less tight than the first obe, whereas below this line, the second Bernstein bound is sharper.

Moreover, if N is a possibly unbounded stopping time that is $\mathcal{F}$-measurable, then for all $\eta > 1$ and $\delta \in (0, 1)$,

$$\mathbb{P}\left[\frac{1}{N}\sum_{t=1}^{N} Z_t \geq \varphi_{\star,+}^{-1}\left(\frac{\eta}{N}\log\left[\frac{\log(N)\log(\eta N)}{\delta \log^2(\eta)}\right]\right)\right] \leq \delta\,,$$
$$\mathbb{P}\left[\frac{1}{N}\sum_{t=1}^{N} Z_t \leq \varphi_{\star,-}^{-1}\left(\frac{\eta}{N}\log\left[\frac{\log(N)\log(\eta N)}{\delta \log^2(\eta)}\right]\right)\right] \leq \delta\,.$$

In order to derive the confidence intervals for individual elements $p(s'|s,a), (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ of transition function, we directly apply the above lemma to sub-Gamma random variables. Let us first recall that sub-Gamma random variables satisfy $\varphi(\lambda) \leq \frac{\lambda^2 v}{2(1-b\lambda)}$, for all $\lambda \in (0, 1/b)$; see, e.g., (Boucheron et al., 2013, Chapter 2.4). Therefore,

$$\varphi_{\star,+}^{-1}(z) = \sqrt{2vz} + bz \quad \text{and} \quad \varphi_{\star,-}^{-1}(z) = -\sqrt{2vz} - bz\,.$$

We finally note that for a Bernoulli distributed random variable with parameter q, we have $v = q(1-q)$ and $b = 1$.

Dann et al. (2017) introduce an alternative time-uniform Bernstein bound. In order to compare the methods, we introduce the following two functions

$$C^{\text{Bernstein-D}}(p, n, \delta) = p + \sqrt{\frac{2p}{n}\ell\ell_n(\delta)} + \frac{\ell\ell_n(\delta)}{n} \tag{4}$$
$$\text{where} \quad \ell\ell_n(\delta) = 2\log\log(\max(e, n)) + \log(3/\delta)$$
$$C^{\text{Bernstein-M}}(p, n, \delta) = p + \sqrt{\frac{2p(1-p)}{n}\ell_n(\delta)} + \frac{\ell_n(\delta)}{3n} \tag{5}$$
$$\text{where} \quad \ell_n(\delta) = \eta\log\left(\frac{\log(n)\log(\eta n)}{\log^2(\eta)\delta}\right) \text{ with } \eta = 1.12\,.$$

Figure 10 plots the ratio $r(p, n, \delta) = C^{\text{Bernstein-M}}(p, n, \delta)/C^{\text{Bernstein-D}}(p, n, \delta)$ as a function of n for different values of p and

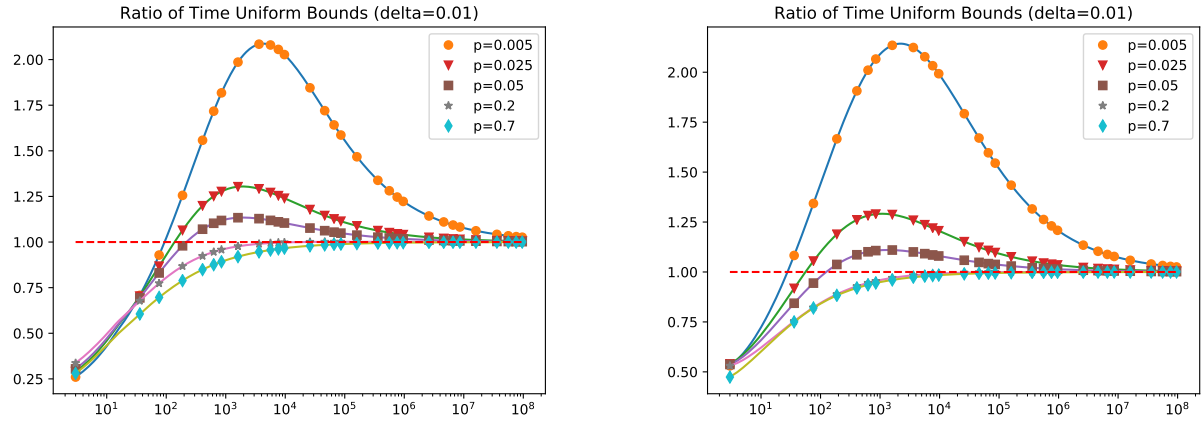

Figure 11. Plot of $n \mapsto r(p, n, \delta)$ for several values of p, with $\delta = 0.01$. We plot the horizontal line $r(p, n, \delta) = 1$ for reference: Above this line, the Gaussian-Laplace bound is looser than the Bernstein bound, while below this line, the Gaussian-Laplace bound is sharper. Left: Using $C^{\text{Bernstein-D}}$ (the first Bernstein bound). Right: Using $C^{\text{Bernstein-M}}$ (the second Bernstein bound).

for the fixed value of $\delta = 0.01$. This shows the clear advantage of using the considered technique over that of (Dann et al., 2017).

In order to better understand the benefit of using a sub-Gaussian tail control for Bernoulli, we further introduce the following function

$$C^{\text{ex-Gaussian-Laplace}}(p, n, \delta) \quad = \quad p + \sqrt{\frac{2\underline{g}(p)(1 + \frac{1}{n}) \log(2\sqrt{n+1}/\delta)}{n}}\,, \tag{6}$$

and plot in Figure 11 the ratio $r(p, n, \delta) = C^{\text{ex-Gaussian-Laplace}}(p, n, \delta)/C^{\text{e-Bernstein-peeling}}(p, n, \delta)$ as a function of n for different values of p and for the fixed value of $\delta = 0.01$. It shows that up to 10^2 samples (for one state-action pair), (6) is sharper than (4) for $p > 0.005$. Hence, this justifies using (6) in practice.

B. Extended Value Iteration

Proof of Lemma 2:

By the discussion in Section 3.2 prior to Algorithm 3, we have that

$$\begin{aligned}
\mathbb{E}_{S\sim p}[f(S)] &= \sum_{s'\in\widetilde{\mathcal{S}}} f(s')p(s') + \sum_{s'\notin\widetilde{\mathcal{S}}} f(s')p(s') \\
&\leq \overline{f}(\widetilde{\mathcal{S}}) + \sum_{s'\notin\widetilde{\mathcal{S}}} f(s')p(s') \\
&\leq \overline{f}(\widetilde{\mathcal{S}}) + \min\left(\kappa, \overline{f}(\widetilde{\mathcal{S}}), \overline{f}(\mathcal{S}\setminus\widetilde{\mathcal{S}})\right)
\end{aligned}$$

where the first inequality holds with high probability by Remark 1, and the second one is guaranteed by the stopping rule of `NOSS` (Algorithm 2). Indeed, `NOSS` by construction builds a minimal set $\widetilde{\mathcal{S}}$ containing the empirical support $\widehat{\mathcal{S}}_n$ (plus eventually one point), and satisfies the condition $\overline{f}(\mathcal{S}\setminus\widetilde{\mathcal{S}}) < \min\left(\kappa, \overline{f}(\widetilde{\mathcal{S}})\right)$ required to exit the loop. □

Proof of Lemma 3:

Let us denote by $\star$ an optimal policy. Let $\boldsymbol{g}_\star : \mathcal{S} \to \mathbb{R}$ denote the constant function equal to $g_\star$, and $\boldsymbol{\kappa}_t$ the constant function equal to κ_t. Using vector notations, we have on the one hand

$$\begin{aligned}
\boldsymbol{g}_\star &= \overline{P}_\star[\mu_\star + P_\star u_n^+ - u_n^+] \\
&\le \overline{P}_\star[\mu_\star^+ + P_{\star,n}^+ u_n^+ + \boldsymbol{\kappa}_t - u_n^+] \text{ w.p. } 1-\delta \\
&\le \overline{P}_\star[\mu^+_{\pi^+_{n+1}} + P^+_{\pi^+_{n+1},n} u_n^+ - u_n^+] + \overline{P}_\star \boldsymbol{\kappa}_t \text{ by optimality of } \pi^+_{n+1} \\
&= \overline{P}_\star[u_{n+1}^+ - u_n^+] + \overline{P}_\star \boldsymbol{\kappa}_t \,.
\end{aligned}$$

On the other hand, for the MDP computed by `EVI-NOSS`, it holds

$$\mathbf{g}^+_{n+1} = \overline{P}^+_{n+1}[\mu^+_{\pi^+_{n+1}} + P^+_{n+1} u_n^+ - u_n^+] = \overline{P}^+_{n+1}[u^+_{n+1} - u_n^+]$$

Hence, combining these two results, we obtain that with probability higher than $1-\delta$,

$$\begin{aligned}
\mathbf{g}_\star - \mathbf{g}^+_{n+1} &\le \overline{P}_\star[u^+_{n+1} - u_n^+] - \overline{P}^+_{n+1}[u^+_{n+1} - u_n^+] + \overline{P}_\star \kappa_t \\
&\le \mathbb{S}(u^+_{n+1} - u_n^+) + \|\overline{P}_\star\|_1 \|\kappa_t\|_\infty \\
&\le \varepsilon + \kappa_t \,.
\end{aligned}$$

□

C. Regret Analysis of `UCRL3`: Proof of Lemma 4 and Theorem 1

In this section, we prove Lemma 4 and Theorem 1.

Proof of Lemma 4:

Recall the definition of the Gini index for pair (s,a): $G_{s,a} := \sum_{x\in\mathcal{S}} p(x|s,a)(1-p(x|s,a))$. Applying Cauchy-Schwarz gives

$$L_{s,a} = \Big(\sum_{x\in\mathcal{K}_{s,a}} \sqrt{p(x)(1-p(x))} \Big)^2 \le K_{s,a} \sum_{x\in\mathcal{K}_{s,a}} p(x)(1-p(x)) = K_{s,a} G_{s,a} \,.$$

Furthermore, in view of the concavity of $z \mapsto \sum_{x\in\mathcal{S}} z(x)(1-z(x))$, the maximal value of $G_{s,a}$ is achieved when $p(x|s,a) = \frac{1}{K_{s,a}}$ for $x \in \mathcal{K}_{s,a}$. Hence, $G_{s,a} \le 1 - 1/K_{s,a}$. Therefore, $L_{s,a} \le K_{s,a} G_{s,a} \le K_{s,a} - 1$. □

We next prove Theorem 1. Our proof follows similar lines as in the proof of (Jaksch et al., 2010, Theorem 2). We start with the following time-uniform concentration inequality to control a bounded martingale difference sequence, which follows from Lemma 5:

Corollary 2 (Time-uniform Azuma-Hoeffding) *Let $(X_t)_{t\ge1}$ be a martingale difference sequence such that for all t, $X_t \in [a,b]$ almost surely for some $a,b \in \mathbb{R}$. Then, for all $\delta \in (0,1)$, it holds*

$$\mathbb{P}\left(\exists T \in \mathbb{N} : \sum_{t=1}^T X_t \ge (b-a)\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} \right) \le \delta \,.$$

Proof of Theorem 1:

Let $\delta \in (0,1)$. To simplify notations, we define the short-hand $J_k := J_{t_k}$ for various random variables that are fixed within a given episode k and omit their dependence on δ (for example $\mathcal{M}_k := \mathcal{M}_{t_k,\delta}$). We let $m(T)$ denote the number of episodes initiated by the algorithm up to time T. By applying Corollary 2, we deduce that

$$\mathfrak{R}(T) = \sum_{t=1}^{T} g^\star - \sum_{t=1}^{T} r_t \le \sum_{s,a} N_{m(T)}(s,a)(g^\star - \mu(s,a)) + \sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\,,$$

with probability at least $1-\delta$. We have

$$\begin{aligned}\sum_{s,a} N_{m(T)}(s,a)(g^\star - \mu(s,a)) &= \sum_{k=1}^{m(T)} \sum_{s,a} \sum_{t=t_k}^{t_{k+1}-1} \mathbb{I}\{s_t = s, a_t = a\}\big(g^\star - \mu(s,a)\big) \\ &= \sum_{k=1}^{m(T)} \sum_{s,a} \nu_k(s,a)\big(g^\star - \mu(s,a)\big)\,.\end{aligned}$$

Introducing $\Delta_k := \sum_{s,a} \nu_k(s,a)\big(g^\star - \mu(s,a)\big)$ for $1 \le k \le m(T)$, we get

$$\mathfrak{R}(T) \le \sum_{k=1}^{m(T)} \Delta_k + \sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\,,$$

with probability at least $1-\delta$. A given episode k is called *good* if $M \in \mathcal{M}_k$ (that is, the set of plausible MDPs contains the true model), and *bad* otherwise.

Control of the regret due to bad episodes. By Lemma 1, the set $\mathcal{M}_k$ contains the true MDP with probability higher than $1-\delta$ uniformly for all T, and for all episodes $k = 1, \dots, m(T)$. As a consequence, with probability at least $1-\delta$, $\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \notin \mathcal{M}_k\} = 0$.

Control of the regret due to good episodes. To upper bound regret in good episodes, we closely follow (Jaksch et al., 2010) and decompose the regret to control the transition and reward functions. Consider a good episode k (hence, $M \in \mathcal{M}_k$). By choosing π_k^+ and $\widetilde{M}_k$, using Lemma 3, we get that

$$g_k := g_{\pi_k^+}^{\widetilde{M}_k} \ge g^\star - \frac{1}{\sqrt{t_k}} - \overline{\kappa}_k\,,$$

with probability greater than $1-\delta$, where $\overline{\kappa}_k = \frac{\gamma \mathbb{S}(u_k) K}{\max_{s,a} N_k(s,a)^{2/3}}$. Hence, with probability greater than $1-\delta$,

$$\Delta_k \le \sum_{s,a} \nu_k(s,a)\big(g_k - \mu(s,a)\big) + \sum_{s,a} \nu_k(s,a)\Big(\frac{1}{\sqrt{t_k}} + \overline{\kappa}_k\Big)\,. \tag{7}$$

Using the same argument as in the proof of (Jaksch et al., 2010, Theorem 2), the value function $u_k^{(i)}$ computed by `EVI-NOSS` at iteration i satisfies: $\max_s u_k^{(i)}(s) - \min_s u_k^{(i)}(s) \le D$. The convergence criterion of `EVI-NOSS` implies

$$|u_k^{(i+1)}(s) - u_k^{(i)}(s) - g_k| \le \frac{1}{\sqrt{t_k}}\,, \qquad \forall s \in \mathcal{S}\,. \tag{8}$$

Using the Bellman operator on the optimistic MDP, we have:

$$u_k^{(i+1)}(s) = \widetilde{\mu}_k(s, \pi_k^+(s)) + \sum_{s'} \widetilde{p}_k(s'|s, \pi_k^+(s)) u_k^{(i)}(s')\,.$$

Substituting this into (8) gives

$$\left|\Big(g_k - \widetilde{\mu}_k(s,\pi_k^+(s))\Big) - \Big(\sum_{s'} \widetilde{p}_k(s'|s,\pi_k^+(s))u_k^{(i)}(s') - u_k^{(i)}(s)\Big)\right| \le \frac{1}{\sqrt{t_k}}\,, \qquad \forall s \in \mathcal{S}\,.$$

Defining $\mathbf{g}_k = g_k\mathbf{1}$, $\widetilde{\boldsymbol{\mu}}_k := \big(\widetilde{\mu}_k(s,\pi_k^+(s))\big)_s$, $\widetilde{\mathbf{P}}_k := \big(\widetilde{p}_k\big(s'|s,\pi_k^+(s)\big)\big)_{s,s'}$ and $\nu_k := \big(\nu_k\big(s,\pi_k^+(s)\big)\big)_s$, we can rewrite the above inequality as:

$$\left|\mathbf{g}_k - \widetilde{\boldsymbol{\mu}}_k - (\widetilde{\mathbf{P}}_k - \mathbf{I})u_k^{(i)}\right| \le \frac{1}{\sqrt{t_k}}\mathbf{1}\,.$$

Combining this with (7) yields

$$\begin{aligned}
\Delta_k &\le \sum_{s,a}\nu_k(s,a)\big(g_k - \mu(s,a)\big) + \sum_{s,a}\nu_k(s,a)\Big(\frac{1}{\sqrt{t_k}} + \overline{\kappa}_k\Big)\\
&= \sum_{s,a}\nu_k(s,a)\big(g_k - \widetilde{\mu}_k(s,a)\big) + \sum_{s,a}\nu_k(s,a)\big(\widetilde{\mu}_k(s,a) - \mu(s,a)\big) + \sum_{s,a}\nu_k(s,a)\Big(\frac{1}{\sqrt{t_k}} + \overline{\kappa}_k\Big)\\
&\le \nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})u_k^{(i)} + \sum_{s,a}\nu_k(s,a)\big(\widetilde{\mu}_k(s,a) - \mu(s,a)\big) + \sum_{s,a}\nu_k(s,a)\Big(\frac{2}{\sqrt{t_k}} + \overline{\kappa}_k\Big)\,.
\end{aligned}$$

Similarly to (Jaksch et al., 2010), we define $w_k(s) := u_k^{(i)}(s) - \frac{1}{2}(\min_s u_k^{(i)}(s) + \max_s u_k^{(i)}(s))$ for all $s \in \mathcal{S}$. Then, in view of the fact that $\widetilde{\mathbf{P}}_k$ is row-stochastic, we obtain

$$\Delta_k \le \nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k + \sum_{s,a}\nu_k(s,a)\big(\widetilde{\mu}_k(s,a) - \mu(s,a)\big) + \sum_{s,a}\nu_k(s,a)\Big(\frac{2}{\sqrt{t_k}} + \overline{\kappa}_k\Big)\,,$$

with probability at least $1-\delta$. The second term in the right-hand side can be upper bounded as follows: $M \in \mathcal{M}_k$ implies

$$\begin{aligned}
\widetilde{\mu}_k(s,a) - \mu(s,a) &\le 2b^r_{t,\delta/(3SA(1+S))}(s,a)\\
&\le \beta_{N_k(s,a)}\big(\tfrac{\delta}{3SA(1+S)}\big)\\
&= \sqrt{\frac{2}{N_k(s,a)}\Big(1+\frac{1}{N_k(s,a)}\Big)\log\Big(3SA(S+1)\sqrt{N_k(s,a)+1}/\delta\Big)}\\
&\le \sqrt{\frac{4}{N_k(s,a)}\log\big(6S^2A\sqrt{T+1}/\delta\big)}\,,
\end{aligned}$$

where we have used $1 \le N_k(s,a) \le T$ and $S \ge 2$ in the last inequality. Furthermore, using $t_k \ge \max_{s,a} N_k(s,a)$ and $\mathbb{S}(u_k) \le D$ yields

$$\sum_{s,a}\nu_k(s,a)\Big(\frac{2}{\sqrt{t_k}} + \overline{\kappa}_k\Big) \le 2\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{N_k(s,a)}} + \gamma DK\sum_{s,a}\frac{\nu_k(s,a)}{N_k(s,a)^{2/3}}\,.$$

Putting together, we obtain

$$\Delta_k \le \nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k + \Big(\sqrt{4\log\big(6S^2A\sqrt{T+1}/\delta\big)} + 2\Big)\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{N_k(s,a)}} + \gamma DK\sum_{s,a}\frac{\nu_k(s,a)}{N_k(s,a)^{2/3}}\,, \tag{9}$$

with probability at least $1-\delta$. In what follows, we derive an upper bound on $\nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k$. Similarly to (Jaksch et al., 2010), we consider the following decomposition:

$$\nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k = \underbrace{\nu_k(\widetilde{\mathbf{P}}_k - \mathbf{P}_k)w_k}_{L_1(k)} + \underbrace{\nu_k(\mathbf{P}_k - \mathbf{I})w_k}_{L_2(k)}\,.$$

The following lemmas provide upper bounds on $L_1(k)$ and $L_2(k)$:

Lemma 8 *Consider a good episode k. Then,*

$$L_1(k) \le \sqrt{2\ell_T\big(\tfrac{\delta}{6S^2A}\big)} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{N_k(s,a)}} D_s \sqrt{L_{s,a}} + 4DS\ell_T\big(\tfrac{\delta}{6S^2A}\big) \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)} \,.$$

Lemma 9 *For all T, it holds with probability at least $1-\delta$,*

$$\sum_{k=1}^{m(T)} L_2(k) \mathbb{I}\{M \in \mathcal{M}_k\} \le D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)} + DSA \log_2(\tfrac{8T}{SA}) \,.$$

Applying Lemmas 8 and 9, and summing over all good episodes, we obtain the following bound that holds with probability higher than $1-2\delta$, uniformly over all $T \in \mathbb{N}$:

$$\begin{aligned}
\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \in \mathcal{M}_k\} &\le \sum_{k=1}^{m(T)} L_1(k) + \sum_{k=1}^{m(T)} L_2(k) \\
&\quad + \Big(\sqrt{4\log\big(6S^2A\sqrt{T+1}/\delta\big)} + 2\Big) \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{N_k(s,a)}} + \gamma DK \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^{2/3}} \\
&\le \sqrt{2\ell_T\big(\tfrac{\delta}{6S^2A}\big)} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{N_k(s,a)}} D_s \sqrt{L_{s,a}} + 4DS\ell_T\big(\tfrac{\delta}{6S^2A}\big) \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)} \\
&\quad + \Big(\sqrt{4\log\big(6S^2A\sqrt{T+1}/\delta\big)} + 2\Big) \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{N_k(s,a)}} \\
&\quad + D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)} + DSA\log_2(\tfrac{8T}{SA}) + \gamma DK \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^{2/3}} \,. \qquad (10)
\end{aligned}$$

To simplify the above bound, we provide the following lemma:

Lemma 10 *We have:*

$$(i) \qquad \sum_{s,a} \sum_{k=1}^{m(T)} \frac{\nu_k(s,a)}{\sqrt{N_k(s,a)}} \le \big(\sqrt{2}+1\big)\sqrt{SAT} \,.$$

$$(ii) \qquad \sum_{s,a} \sum_{k=1}^{m(T)} \frac{\nu_k(s,a)}{\sqrt{N_k(s,a)}} D_s \sqrt{L_{s,a}} \le \big(\sqrt{2}+1\big)\sqrt{\sum_{s,a} D_s^2 L_{s,a} T} \,.$$

$$(iii) \qquad \sum_{s,a} \sum_{k=1}^{m(T)} \frac{\nu_k(s,a)}{N_k(s,a)} \le 2SA\log\big(\tfrac{T}{SA}\big) + SA \,.$$

$$(iv) \qquad \sum_{s,a} \sum_{k=1}^{m(T)} \frac{\nu_k(s,a)}{N_k(s,a)^{2/3}} \le 6(SA)^{2/3}T^{1/3} + 2SA \,.$$

Putting everything together, it holds that with probability at least $1-4\delta$,

$$\begin{aligned}
\mathfrak{R}(T) &\le \big(\sqrt{2}+1\big)\Big(\sqrt{4\log\big(6S^2A\sqrt{T+1}/\delta\big)} + 2\Big)\sqrt{SAT} + \big(D\sqrt{2} + \sqrt{\tfrac{1}{2}}\big)\sqrt{(T+1)\log(\sqrt{T+1}/\delta)} \\
&\quad + \sqrt{2\ell_T\big(\tfrac{\delta}{6S^2A}\big)}\big(\sqrt{2}+1\big)\sqrt{T\sum_{s,a} D_s^2 L_{s,a}} \\
&\quad + 4DS\ell_T\big(\tfrac{\delta}{6S^2A}\big)\Big(2SA\log\big(\tfrac{T}{SA}\big) + SA\Big) \\
&\quad + 60DKS^{2/3}A^{2/3}T^{1/3} + DSA\log_2(\tfrac{8T}{SA}) + 20DKSA \,.
\end{aligned}$$

Noting that for $S, A \geq 2$, it is easy to verify that for $T \geq 3$, $\ell_T\left(\frac{\delta}{6S^2A}\right) \leq 2\log\left(6S^2A\sqrt{T+1}/\delta\right)$. Hence, after simplification we obtain that for all $T \geq 3$, with probability at least $1-4\delta$,

$$\mathfrak{R}(T) \leq \left(5\sqrt{\sum_{s,a} D_s^2 L_{s,a}} + 10\sqrt{SA} + 2D\right)\sqrt{T\log\left(\tfrac{6S^2A\sqrt{T+1}}{\delta}\right)} + 60DKS^{2/3}A^{2/3}T^{1/3} + \mathcal{O}\Big(DS^2A\log^2\left(\tfrac{T}{\delta}\right)\Big)\,.$$

Finally we remark that

$$5\sqrt{\sum_{s,a} D_s^2 L_{s,a}} + 10\sqrt{SA} \leq 20\sqrt{SA + \sum_{s,a} D_s^2 L_{s,a}} \leq 20\sqrt{2\sum_{s,a}\left(D_s^2 L_{s,a} \vee 1\right)}\,,$$

so that $\mathfrak{R}(T) = \mathcal{O}\Big(\Big[\sqrt{\sum_{s,a}\left(D_s^2 L_{s,a} \vee 1\right)} + D\Big]\sqrt{T\log(\sqrt{T}/\delta)}\Big)$. □

C.1. Proof of Technical Lemmas

Proof of Lemma 8:

To derive an upper bound on $L_1(k)$, first notice that

$$\begin{aligned} L_1(k) &= \sum_{s,x} \nu_k(s,\pi_k^+(s))\Big(\widetilde{p}_k(x|s,\pi_k^+(s)) - p(x|s,\pi_k^+(s))\Big)w_k(x) \\ &\leq \sum_{s,a}\nu_k(s,a)\sum_x \Big(\widetilde{p}_k(x|s,a) - p(x|s,a)\Big)w_k(x)\,. \end{aligned}$$

Fix s and a, and introduce short-hands $N_k := N_k(s,a)$, $\widetilde{p}_k := \widetilde{p}_k(\cdot|s,a)$, $\widehat{p}_k := \widehat{p}_k(\cdot|s,a)$, and $p := p(\cdot|s,a)$. We have

$$\begin{aligned} \sum_x \Big(\widetilde{p}_k(x|s,a) - p_k(x|s,a)\Big)w_k(x) &= \sum_x (\widetilde{p}_k(x) - p(x))w_k(x) \\ &\leq \underbrace{\sum_x |\widehat{p}_k(x) - p(x)||w_k(x)|}_{F_1} + \underbrace{\sum_x |\widetilde{p}_k(x) - \widehat{p}_k(x)||w_k(x)|}_{F_2}\,. \end{aligned}$$

To upper bound F_1, we first show that $\max_{x\in\mathrm{supp}(\widetilde{p}_k(\cdot|s,a))}|w_k(x)| \leq \frac{D_s}{2}$. To show this, we note that similarly to (Jaksch et al., 2010), we can combine all MDPs in $\mathcal{M}_k$ to form a single MDP $\widetilde{\mathcal{M}}_k$ with continuous action space $\mathcal{A}'$. In this extended MDP, in each state $s \in \mathcal{S}$, and for each $a \in \mathcal{A}$, there is an action in $\mathcal{A}'$ with mean $\widetilde{\mu}(s,a)$ and transition probability $\widetilde{p}(\cdot|s,a)$ satisfying (1). Similarly to (Jaksch et al., 2010), we note that $u_k^{(i)}(s)$ amounts to the total expected i-step reward of an optimal non-stationary i-step policy starting in state s on the MDP $\widetilde{\mathcal{M}}_k$ with the extended action set. The local diameter of state s of this extended MDP is at most D_s, since by assumption k is a good episode and hence $\mathcal{M}_k$ contains the true MDP M, and therefore, the actions of the true MDP are contained in the continuous action set of $\widetilde{\mathcal{M}}_k$. Now, if there were states $s_1, s_2 \in \cup_a \mathrm{supp}(\widetilde{p}_k(\cdot|s,a))$ with $u_k^{(i)}(s_1) - u_k^{(i)}(s_2) > D_s$, then an improved value for $u_k^{(i)}(s_1)$ could be achieved by the following non-stationary policy: First follow a policy that moves from s_1 to s_2 most quickly, which takes at most D_s steps on average. Then follow the optimal i-step policy for s_2. We thus have $u_k^{(i)}(s_1) \geq u_k^{(i)}(s_2) - D_s$, since at most D_s of the i rewards of the policy for s_2 are missed. This is a contradiction, and so the claim follows.

To upper bound F_1, noting that k is a good episode yields:

$$\begin{aligned}
F_1 &\leq \sqrt{\frac{2\ell_{N_k}}{N_k}} \sum_x \sqrt{p(x)(1-p(x))}|w_k(x)| + \frac{S\ell_{N_k}}{3N_k}\|w_k\|_\infty \\
&\leq \max_{x\in\mathcal{K}_{s,a}} |w_k(x)| \sqrt{\frac{2\ell_{N_k}}{N_k}} \sum_x \sqrt{p(x)(1-p(x))} + \frac{DS\ell_{N_k}}{6N_k} \\
&\leq D_s \sqrt{\frac{\ell_{N_k}}{2N_k}} \sum_x \sqrt{p(x)(1-p(x))} + \frac{DS\ell_{N_k}}{6N_k} \\
&= D_s \sqrt{\frac{\ell_{N_k}}{2N_k} L_{s,a}} + \frac{DS\ell_{N_k}}{6N_k}\,,
\end{aligned}$$

where we have used that $\|w_k\|_\infty \leq \frac{D}{2}$ and $\max_{x\in\mathcal{K}_{s,a}} |w_k(x)| \leq \frac{D_s}{2}$.

To upper bound F_2, we will need the following lemma:

Lemma 11 *Consider x and y satisfying $|x-y| \leq \sqrt{2y(1-y)\zeta} + \zeta/3$. Then,*

$$\sqrt{y(1-y)} \leq \sqrt{x(1-x)} + 2.4\sqrt{\zeta}\,.$$

Applying Lemma 11 twice and using the relation $\max_{x\in\mathrm{supp}(\widetilde{p}_k(\cdot|s,a))} |w_k(x)| \leq \frac{D_s}{2}$ yield:

$$\begin{aligned}
F_2 &\leq \sqrt{\frac{2\ell_{N_k}}{N_k}} \sum_x \sqrt{\widetilde{p}_k(x)(1-\widetilde{p}_k(x))}|w_k(x)| + \frac{DS\ell_{N_k}}{6N_k} \\
&\leq D_s \sqrt{\frac{\ell_{N_k}}{2N_k}} \sum_x \sqrt{\widetilde{p}_k(x)(1-\widetilde{p}_k(x))} + \frac{DS\ell_{N_k}}{6N_k} \\
&\leq D_s \sqrt{\frac{\ell_{N_k}}{2N_k}} \sum_x \sqrt{\widehat{p}_k(x)(1-\widehat{p}_k(x))} + 2.4\sqrt{2}\frac{DS\ell_{N_k}}{N_k} + \frac{DS\ell_{N_k}}{6N_k} \\
&\leq D_s \sqrt{\frac{\ell_{N_k}}{2N_k}} \sum_x \sqrt{p(x)(1-p(x))} + \frac{3.6DS\ell_{N_k}}{N_k}\,.
\end{aligned}$$

Combininig the bounds on F_1 and F_2, and noting that

$$\ell_{N_k(s,a)}\big(\tfrac{\delta}{3(1+S)SA}\big) \leq \ell_{N_k(s,a)}\big(\tfrac{\delta}{6S^2A}\big) \leq \ell_T\big(\tfrac{\delta}{6S^2A}\big)$$

complete the proof. □

Proof of Lemma 11:

By Taylor's expansion, we have

$$\begin{aligned}
y(1-y) &= x(1-x) + (1-2x)(y-x) - (y-x)^2 \\
&= x(1-x) + (1-x-y)(y-x) \\
&\leq x(1-x) + |1-x-y|\left(\sqrt{2y(1-y)\zeta} + \tfrac{1}{3}\zeta\right) \\
&\leq x(1-x) + \sqrt{2y(1-y)\zeta} + \tfrac{1}{3}\zeta\,.
\end{aligned}$$

Using the fact that $a \le b\sqrt{a} + c$ implies $a \le b^2 + b\sqrt{c} + c$ for nonnegative numbers a, b, and c, we get

$$\begin{aligned} y(1-y) &\le x(1-x) + \tfrac{1}{3}\zeta + \sqrt{2\zeta\left(x(1-x) + \tfrac{1}{3}\zeta\right)} + 2\zeta \\ &\le x(1-x) + \sqrt{2\zeta x(1-x)} + 3.15\zeta \\ &= \left(\sqrt{x(1-x)} + \sqrt{\tfrac{1}{2}\zeta}\right)^2 + 2.65\zeta\,, \end{aligned} \tag{11}$$

where we have used $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ valid for all $a, b \ge 0$. Taking square-root from both sides and using the latter inequality give the desired result:

$$\sqrt{y(1-y)} \le \sqrt{x(1-x)} + \sqrt{\tfrac{1}{2}\zeta} + \sqrt{2.65\zeta} \le \sqrt{x(1-x)} + 2.4\sqrt{\zeta}\,.$$

□

Proof of Lemma 9:

Similarly to the proof of (Jaksch et al., 2010, Theorem 2), we define the sequence $(X_t)_{t\ge1}$ with $X_t := (p(\cdot|s_t, a_t) - \mathbf{e}_{s_{t+1}})w_{k(t)}\mathbb{I}\{M \in \mathcal{M}_{k(t)}\}$, for all t, where $k(t)$ denotes the episode containing time step t. For any k with $M \in \mathcal{M}_k$, we have that:

$$\begin{aligned} L_2(k) = \nu_k(\mathbf{P}_k - \mathbf{I})w_k &= \sum_{t=t_k}^{t_{k+1}-1} (p(\cdot|s_t, a_t) - \mathbf{e}_{s_t})w_k \\ &= \sum_{t=t_k}^{t_{k+1}-1} \Big(p(\cdot|s_t, a_t) - \mathbf{e}_{s_{t+1}} + \mathbf{e}_{s_{t+1}} - \mathbf{e}_{s_t}\Big)w_k = \sum_{t=t_k}^{t_{k+1}-1} X_t + w_k(s_{t+1}) - w_k(s_t) \le \sum_{t=t_k}^{t_{k+1}-1} X_t + D\,, \end{aligned}$$

so that $\sum_{k=1}^{m(T)} L_2(k) \le \sum_{t=1}^{T} X_t + m(T)D$. Using $\|w_k\|_\infty = \frac{D}{2}$ and applying the Hölder inequality give

$$|X_t| \le \|p(\cdot|s_t, a_t) - \mathbf{e}_{s_{t+1}}\|_1 \frac{D}{2} \le \Big(\|p(\cdot|s_t, a_t)\|_1 + \|\mathbf{e}_{s_{t+1}}\|_1\Big)\frac{D}{2} = D\,.$$

So, X_t is bounded by D, and also $\mathbb{E}[X_t|s_1, a_1, \ldots, s_t, a_t] = 0$, so that $(X_t)_{t\ge1}$ is martingale difference sequence. Therefore, by Corollary 2, we get:

$$\mathbb{P}\left(\exists T : \sum_{t=1}^{T} X_t \ge D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)}\right) \le \delta\,,$$

thus concluding the proof. □

C.2. Proof of Supporting Lemmas

Proof of Lemma 10:

Inequalities (i)-(iii) easily follow from Lemma 12, which is stated at the end of this proof, and using Jensen's inequality. Next we prove the inequality (iv).

Given $t \geq 1$, let $k(t)$ denote the episode containing time step t. Following similar steps as in the proof of (Ouyang et al., 2017, Lemma 5), we have

$$\begin{aligned}
\sum_{s,a}\sum_{k=1}^{m(T)} \frac{\nu_k(s,a)}{N_k(s,a)^{2/3}} &= \sum_{s,a}\sum_{t=1}^{T} \frac{\mathbb{I}\{(s_t,a_t)=(s,a)\}}{N_{k(t)}(s,a)^{2/3}} \\
&\leq 2\sum_{s,a}\sum_{t=1}^{T} \frac{\mathbb{I}\{(s_t,a_t)=(s,a)\}}{N_t(s,a)^{2/3}} \\
&= 2\sum_{s,a}\left(\mathbb{I}\{N_{m(T)}(s,a)\geq 1\} + \sum_{j=1}^{N_{m(T)}(s,a)} j^{-2/3}\right) \\
&\leq 2SA + 6\sum_{s,a} N_{m(T)}(s,a)^{1/3} \\
&\leq 2SA + 6SA\left(\sum_{s,a}\frac{N_{m(T)}(s,a)}{SA}\right)^{1/3} \\
&= 2SA + 6S^{2/3}A^{2/3}T^{1/3}\,,
\end{aligned}$$

where we have used that for any $L \geq 1$, $\sum_{j=1}^{L} j^{-2/3} \leq 1 + \int_1^L z^{-2/3}\mathrm{d}z \leq 3L^{1/3}$, and where the last step follows from Jensen's inequality. □

Lemma 12 ((Jaksch et al., 2010, Lemma 19),(Talebi & Maillard, 2018, Lemma 24)) *For any sequence of numbers $z_1, z_2, \ldots, z_n$ with $0 \leq z_k \leq Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$, it holds*

$$(i) \qquad \sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \leq \left(\sqrt{2}+1\right)\sqrt{Z_n}\,.$$

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

D. Further Details for Experiments

Tie-breaking rule to compute optimistic policies. All the considered algorithms (UCRL2, KL-UCRL, UCRL2B, UCRL3) resort to a form of EVI internal procedure, that computes at each iteration n a policy π_n^+ maximizing the current optimistic value u_n^+ (see Algorithm 1). In practice, several policies may satisfy this, hence a tie-breaking rule is required. For fairness, we used the same tie-breaking rule for all algorithms. It consists, for a state s, to break ties by defining the policy to choose an action uniformly randomly amongst $\mathrm{Argmin}_{a\in\mathcal{A}} N_k(s,a)$. Such breaking rules aim to stabilize the algorithm.

Atypical sequences. The concentration inequalities we have employed for UCRL3 are mostly tight. Unfortunately, concentration inequalities are also known to be loose in the specific case of atypical sequences of observations. Namely, the specific situation when $n = N_t(s,a) > 1$ and all observed samples from (s,a) equal s_0, corresponds to observing a sequence of n ones from a Bernoulli distribution with parameter $\theta = p(s_0|s,a)$. Note that for n i.i.d. observations, this event should be of probability θ^n. In such a situation where $\widehat{p}_t(s_0|s,a) = 1$, all concentration inequalities yield conservative lower bounds on $p(s_0|s,a)$. We replace these lower bounds with $(1/2)^n$ for this very specific situation.

Extended Value Iterations with lazy support updates The EVI-NOSS procedure proceeds in steps, first computing an optimistic support, then updating u and π using the Bellman optimal operator at every single step. In order to reduce computation, we use a lazy implementation that keeps updating u and π at each step but updates the support only once every L-steps. This also tends to reduce the number of steps before convergence in practice. In our experiments, we chose $L = 5$.

Code release The full code implementation is made publicly available as an article companion following this link: `https://gitlab.inria.fr/omaillar/average-reward-reinforcement-learning`. It is coded in Python 3, and is designed to be compatible with OpenAI gym discrete environments.

E. Numerical Experiments with PSRL

In this section, we provide further numerical comparison with a version of the PSRL algorithm (Osband et al., 2013) for average-reward RL. PSRL is a popular algorithm originally designed and analysed for episodic RL, and to the best of our knowledge, its (frequentist) regret guarantees in the context of average-reward regret minimization are still unclear. In this section we will show that PSRL can be a competitive strategy in several environments but also, unfortunately, completely fails in some others. We believe might provide pointers to the lack of theoretical guarantees for this strategy and suggests further modifications could help obtain the best of both worlds (UCRL3 and PSRL).

We study a variant of PSRL, which maintains for each state-action pair (s, a) a Dirichlet distribution to model the transition distribution $p(\cdot|s, a)$, and a Beta distribution to model the reward distribution $\nu(s, a)$. The Beta distribution is classically used as a prior to model Bernoulli distributions. Here, we only know that the rewards are supported on $[0, 1]$, but we can use the popular *Bernoullization* trick. That is, using n rewards sampled from $\nu(s, a)$, we use a Beta distribution $\mathrm{Beta}(S + \alpha, F + \alpha)$, where S stands for the pseudo-success-counts equal to the sum of the n rewards, and F denotes the pseudo-failure-counts equal to $n - S$. The Dirichlet distribution is initialized with uniform weights equal to α, and we use $\alpha = 1$ for both Dirichlet and Beta initial parameters.

We report in the next figures the results of PSRL against UCRL3 and some algorithms that enjoy controlled (frequentist) regret guarantees – In the figures, we referred to this variant as PSRL-AvR.

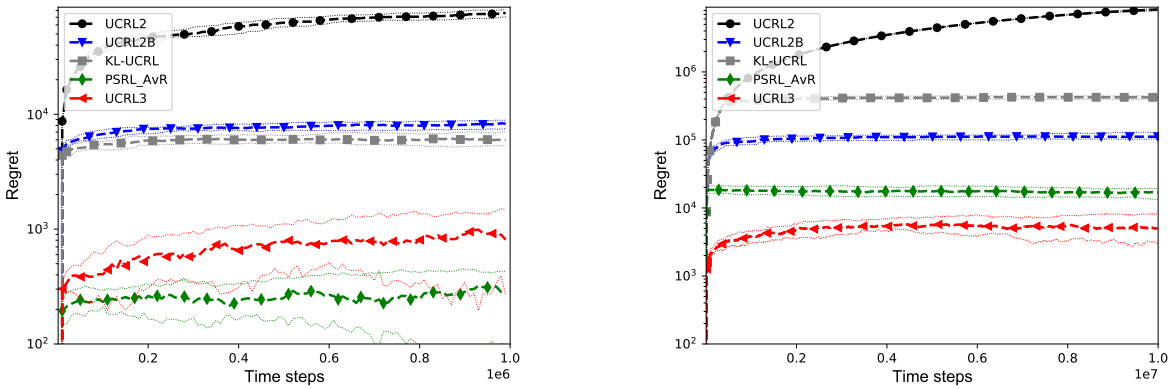

Figure 12. Regret for the 6-state (left) and 25-state (right) *RiverSwim* environments

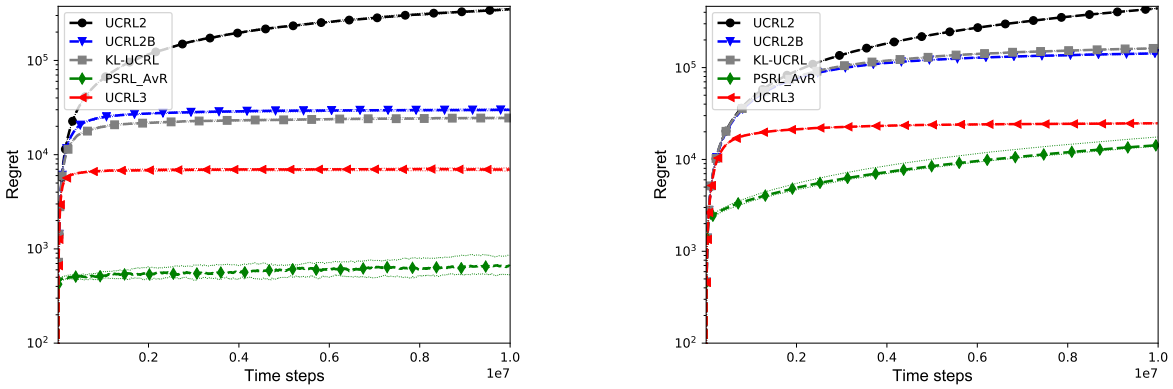

Figure 13. Regret for the 2-room (left) and 4-room (right) goal-state environments

The performance of PSRL is good in the 6-state RiverSwim environment, but degrades with a larger number of states when compared to UCRL3. The performance of PSRL in 2-room and 4-room MDPs are striking. These environments are goal-state (a.k.a. goal-oriented) MDPs with very sparse rewards. We conjecture PSRL favors such environment. In a Garnet MDP, we observe that PSRL is not necessarily competitive. Figure 14, left, shows the results in a 100-state random MDP

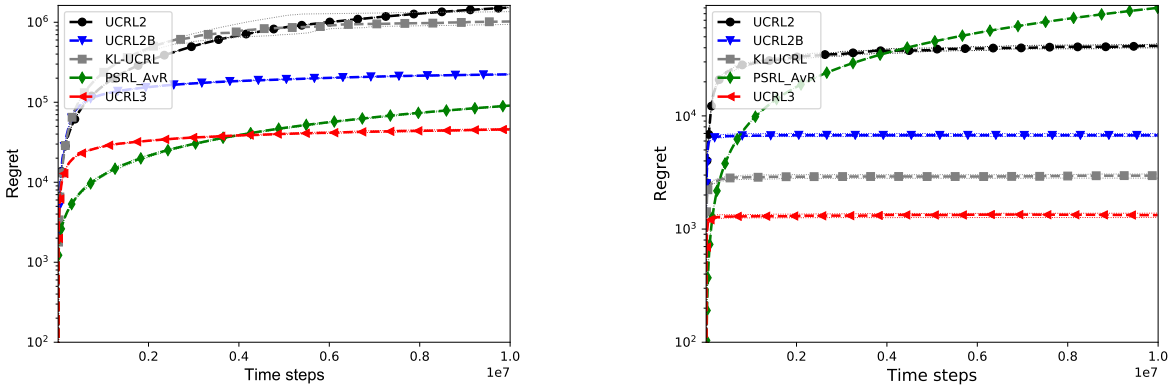

Figure 14. Regret in a 100-state randomly generated MDP with sparse rewards (left), and a 10-state randomly generated MDP with rich rewards (right)

with relatively sparse rewards. PSRL has a competitve initial phase, but is later outperformed by UCRL3, which suggests it is unable to find an optimal policy when T is not very large. Figure 14, right, shows the result of an experiment in a small 10-state Garnet MDP, but where most rewards are constrained to be far from 0. Here, we observe that PSRL achieves a very poor performance in such a case. The MDP is depicted in Figure 15 for completeness.

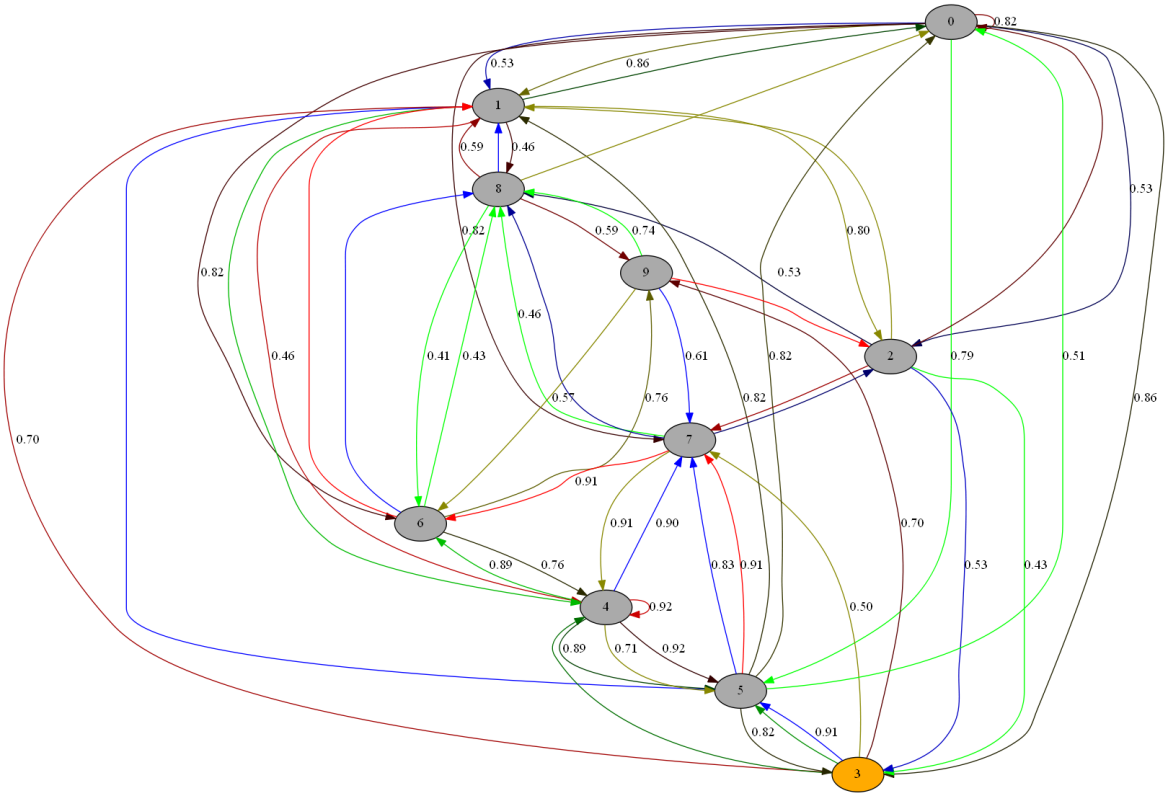

Figure 15. A randomly-generated reward-rich MDP with 10 states: One color per action, shaded according to the corresponding probability mass, labels indicate mean reward, and the current state is highlighted in orange.