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, Vienna, Austria, 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 $\mathfrak{R}(\mathbb{A}, T) := Tg^\star - \sum_{t=1}^{T} r_t(s_t, a_t)$, 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 reports 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)	$\mathcal{O}\Big(D\sqrt{S\sum_{s,a}(\mathbb{V}_{s,a}\vee 1)T\log(\log(T)/\delta)}\Big)$	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) derives 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) introduces 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) presents 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) presents 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) presents 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. In (Dann et al., 2017), the authors used 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 $\widetilde{\mathcal{S}} = \widehat{\mathcal{S}} \cup \operatorname{argmax}_{s\in\mathcal{S}} f(s)$,
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 $t_0 = 0$, $t = 1$, $k = 1$, and observe the initial state s_1
for episodes $k = 1, 2, \ldots$ **do**
  Set $t_k = t$
  Set $N_{t_k}(s,a) = N_{t_{k-1}}(s,a) + v_k(s,a)$ 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)
  Compute (see Algorithm 3)

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

  while $v_k(s_t, \pi^+_{t_k}(s_t)) < N_{t_k}(s_t, \pi^+_{t_k}(s_t))$ **do**
    Play action $a_t = \pi^+_{t_k}(s_t)$, and observe the next state s_{t+1} and reward $r_t(s_t,a_t)$
    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):*

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

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

$$\mathfrak{R}(\textit{UCRL3}, T) \le c\sqrt{T \log\left(\tfrac{6S^2A\sqrt{T+1}}{\delta}\right)} + 30DKS^{2/3}A^{2/3}T^{1/3} + \mathcal{O}\Big(DS^2A\log^2\left(\tfrac{T}{\delta}\right)\Big),$$

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

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

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 $K_{s,a}$.

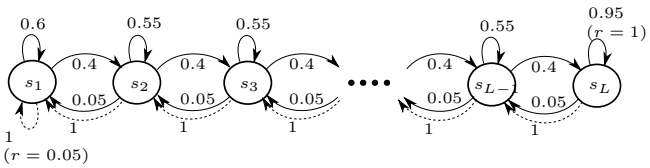

Figure 2. The L-state *RiverSwim* MDP

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 F, 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 E 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.

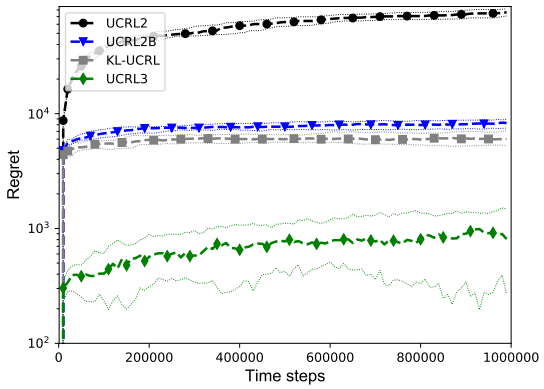

Figure 3. Regret for the 6-state *RiverSwim* environment

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.

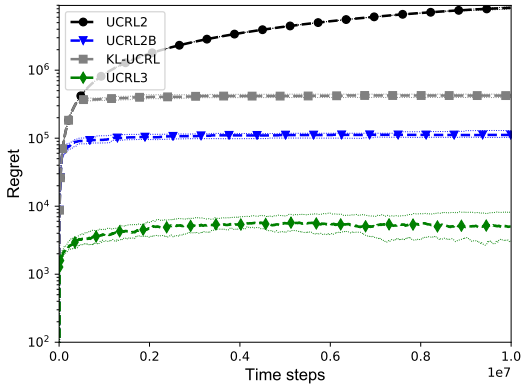

Figure 4. Regret for the 25-state *RiverSwim* environment

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*.*

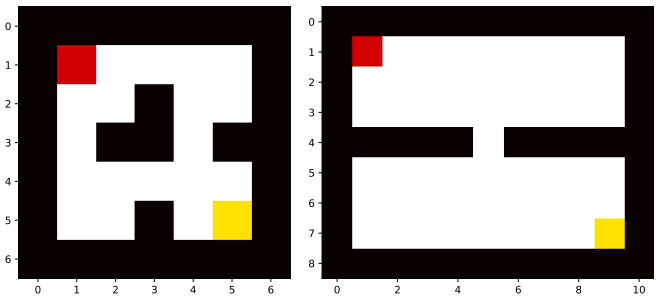

Figure 5. A 4-room (left) and a 2-room (right) grid-world environment, with 20 and 55 states: the starting state is shown in red, and the rewarding state is shown in yellow. From the yellow state, all actions bring the learner to the red state. Other transitions are noisy as in a *frozen-lake* environment.

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.

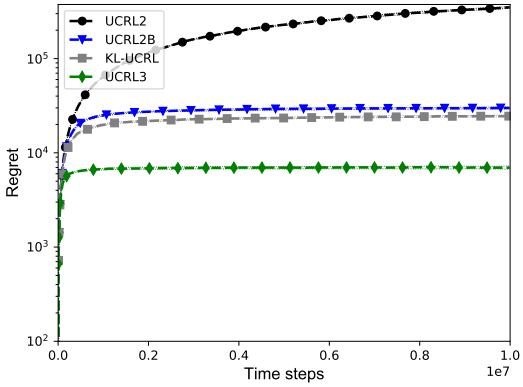

Figure 6. Regret for the 4-room environment

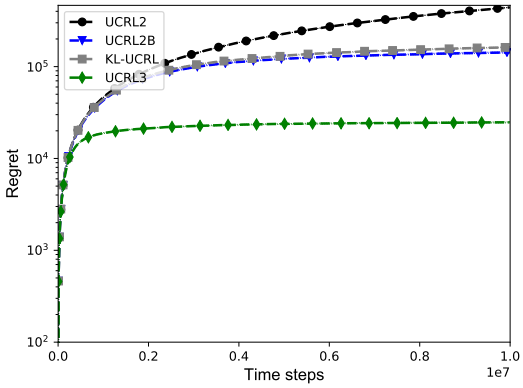

Figure 7. Regret for the 2-room environment

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

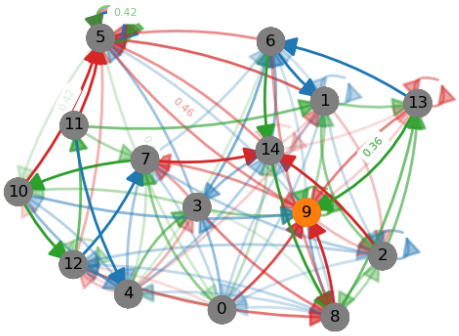

Figure 8. A randomly-generated MDP with 15 states: One color per action, shaded according to the corresponding probability mass, labels indicate mean reward, and the current state is highlighted in orange.

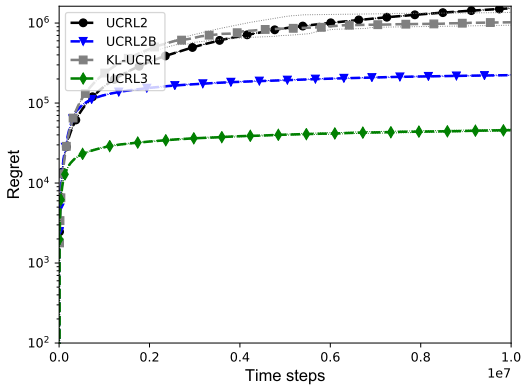

Figure 9. Regret in one 100-state randomly generated MDP

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). This work was done when 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.