Tightening Exploration in Upper Confidence Reinforcement Learning

Hippolyte Bourel\textsuperscript{1} Odalric-Ambrym Maillard\textsuperscript{1} Mohammad Sadegh Talebi\textsuperscript{2}

Abstract

The upper confidence reinforcement learning (UCRL) strategy 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 strategy and its variants have remained until now mostly theoretical as numerical experiments on simple environments exhibit long burn-in phases before the learning takes place. In pursuit of practical efficiency, we present UCRL\textsuperscript{3}, following the lines of UCRL\textsuperscript{2}, 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. Further, to tighten exploration, it uses an adaptive computation of the support of each transition distributions, which in turn enables us to revisit the extended value iteration procedure to optimize over distributions with reduced support by disregarding low probability transitions, while still ensuring near-optimism. We demonstrate, through numerical experiments on standard environments, that reducing exploration this way yields a substantial numerical improvement compared to UCRL\textsuperscript{2} and its variants. On the theoretical side, these key modifications enable us to derive a regret bound for UCRL\textsuperscript{3} improving on UCRL\textsuperscript{2}, that for the first time makes appear notions of local diameter and effective support, thanks to variance-aware concentration bounds.

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 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \nu)$ be an undiscounted MDP, where $\mathcal{S}$ denotes the discrete state-space with cardinality $|\mathcal{S}|$, and $\mathcal{A}$ denotes the discrete action-space with cardinality $|\mathcal{A}|$. $p$ 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 $K_{s,a}^t$ the set of successor states of the state-action pair $(s,a)$, that is $K_{s,a} := \{x \in \mathcal{S} : p(x|s,a) > 0\}$, and further define $K_{s,a} := |K_{s,a}|$. Finally, $\nu$ is a reward distribution function on $[0,1]$ with mean function denoted by $\mu$. 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 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 reward function $\nu$ 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 $\nu$, to that gathered by the learner. Following (Jaksch et al., 2010), we define the regret of a learning algorithm $\hat{\pi}$ after $T$ steps as $\mathcal{R}(\hat{\pi}, T) := Tg^* - \sum_{t=1}^T r_t(s_t,a_t)$, where $g^*$ denotes the average-reward (or gain) attained by an optimal policy. Alternatively, the objective of the learner is to minimize the regret, which calls for balancing between 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, 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 long-term reward. A popular algorithm 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 $\tilde{O}(DS\sqrt{AT})$ with high probability, in any communicating MDP with $S$ states, $A$ actions, and diameter $D^A$. (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, 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; 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. Although these algorithms enjoy delicate and strong theoretical regret guarantees, their numerical assessment 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 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 has been reported to significantly improve on UCRL2. 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) theoretical guarantees. Though our result does not imply that optimistic strategies are the best, it shows that they can be much stronger competitors than the 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 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 an novel procedure, which we call NOSS, which 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 consider a near-optimistic, as opposed to fully optimistic, for planning. Our adaptive procedure enables us to control this additional error. Further, this enables us to make appear in the regret analysis a notion of local diameter as well as a local effective support (Section 3.3), which in turn reduces the regret bounds. We define the local diameter below.

**Definition 1 (Local Diameter of State $s$)** Consider state $s \in S$. For $s_1, s_2 \in \bigcup_{a \in A} K_{s,a}$ with $s_1 \neq s_2$, let $T^\pi(s_1, s_2)$ denote the number of steps it takes to get from $s_1$ to $s_2$ starting from $s_1$ and following policy $\pi$. Then, the local diameter of MDP $M$ for $s$, denoted by $D^\pi_{local}(M)$, is defined as

$$D^\pi_{local} := \max_{s_1, s_2 \in \bigcup_{a \in A} K_{s,a}} \min_{\pi} \min_{s} \mathbb{E}[T^\pi(s_1, s_2)].$$

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 this adaptive support estimation procedure 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.** 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 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 a wider class of MDPs are
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, \( \nu_{s,a} \) denotes the variance of optimal bias function of the true MDP, when the state is distributed according to \( p(\cdot|s,a) \). For UCRL3, \( L_{s,a} := (\sum_{x \in S} p(x|s,a)(1 - p(x|s,a)))^2 \) denotes the local effective support of \( p(\cdot|s,a) \).

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). 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 \( O(\sqrt{DS\max_{s,a}\nu_{s,a}T}) \) for the class of communicating MDPs. A more refined regret bound for KL-UCRL for ergodic MDPs is presented in (Talebi & Maillard, 2018). (Bartlett & Tewari, 2009) presents REGAL and report a \( O(D'S\sqrt{AT}) \) 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 bias function of the true MDP. (Fruit et al., 2018b) presents SCAL, which similarly to REGAL works for 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 scaling as \( O(D\sqrt{\sum_{s,a}\nu_{s,a}T}) \). Recently, (Zhang & Ji, 2019) presents EBF achieving a regret of \( O(\sqrt{HSA}) \) assuming that the learner knows an upper bound \( H \) on the span of the optimal bias function of the true MDP.\(^\dagger\) 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 from 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, reported in (Agrawal & Jia, 2017b), scales as \( O(\sqrt{DS\max_{s,a}\nu_{s,a}T}\log(T)) \) 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 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 undiscounted RL considered here (in particular regarding closing the gap between lower and upper bounds).

In Figure 1, we report the regret upper bounds of algorithms for regret minimization in the average-reward setting. We do not report in this table, REGAL and EBF, as no efficient implementation is currently known. Let us remark that the presented regret bound for UCRL3 does not contradict the worst-case lower bound \( \Omega(\sqrt{DSAT}) \) of (Jaksch et al., 2010). Indeed, for the *hard-to-learn* MDP used in the proof of 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 \( y \). \( \Delta_S \) represents the set of all probability distributions defined on \( S \). For a distribution \( p \in \Delta_S \) and a vector-function \( f = (f(s))_{s \in S} \), we denote by \( Pf \) its application on \( f \), defined by \( Pf = \mathbb{E}_{X \sim p}[f(X)] \). We introduce \( \Delta_S^{S \times A} := \{ q : q(\cdot|s,a) \in \Delta_S, (s,a) \in S \times A \} \), and for \( p \in \Delta_S^{S \times A} \), 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 \( S(f) = \max_s f(s) - \min_s f(s) \).

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 \( t \). Let us define \( \hat{\mu}_t(s,a) \) 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 \( \hat{\nu}_t(\cdot|s,a) \) as the empirical distribution built using \( N_t(s,a) \) i.i.d. observations from \( p(\cdot|s,a) \). We further define \( N_t(s,a)^+ := \max\{N_t(s,a), 1\} \).

### 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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Regret bound</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCRL2 (Jaksch et al., 2010)</td>
<td>( \mathcal{O}(DS\sqrt{AT}\log(T)) )</td>
<td>valid for fixed ( T ) provided as input.</td>
</tr>
<tr>
<td>KL-UCRL (Filippi et al., 2010)</td>
<td>( \mathcal{O}(DS\sqrt{AT}\log(\log(T))\nu_{s,a}) )</td>
<td>restricted to ergodic MDPs.</td>
</tr>
<tr>
<td>KL-UCRL (Talebi &amp; Maillard, 2018)</td>
<td>( \mathcal{O}(DS\sqrt{\sum_{s,a}\nu_{s,a}}\nu_{s,a}T\log(T)) )</td>
<td>without knowledge of the span.</td>
</tr>
<tr>
<td>SCAL+ (QIAN et al., 2019)</td>
<td>( \mathcal{O}(DS\sqrt{\sum_{s,a}\nu_{s,a}T\log(T)}\nu_{s,a}) )</td>
<td>note the extra ( \sqrt{\log(T)} ) term.</td>
</tr>
<tr>
<td>UCRL2B (Fruit et al., 2019)</td>
<td>( \mathcal{O}(\sqrt{DS\max_{s,a}\nu_{s,a}T\log(T)}\nu_{s,a}) )</td>
<td></td>
</tr>
<tr>
<td>UCRL3 (This Paper)</td>
<td>( \mathcal{O}(\sqrt{DS\max_{s,a}\nu_{s,a}T\log(T)}\nu_{s,a}) )</td>
<td></td>
</tr>
</tbody>
</table>

---

\( \dagger \) We remark that the universal constants of the leading term here are fairly large.
the following two sets: For each \((s, a) \in S \times A\),

\[
\begin{align*}
&\text{UCRL}^2 \left( s, a \right) = \\
&\left\{ \mu' \in [0, 1] : \left| \hat{\mu}_t(s, a) - \mu' \right| \leq \sqrt{\frac{3.5 \log \left( \frac{2S\hat{A}}{\delta} \right)}{N_t(s, a)}} \right\},
\end{align*}
\]

\[
\begin{align*}
&\text{UCRL}^2 \left( s', a \right) = \\
&\left\{ p' \in \Delta S : \| \hat{p}_t(s, a) - p' \|_1 \leq \sqrt{\frac{14S \log \left( \frac{2A}{\delta} \right)}{N_t(s, a)}} \right\}.
\end{align*}
\]

At a high level, \textbf{UCRL2} maintains the set of MDPs \( M_{t, \delta} = \{ \tilde{M} = (S, \mathcal{A}, \tilde{p}, \tilde{v}) \} \), where for each \((s, a) \in S \times \mathcal{A}\), \( \tilde{\mu}(s, a) \in \text{UCRL}^2(s, a) \) (with \( \tilde{\mu} \) denoting the mean of \( \tilde{v} \)) and \( \tilde{p}(s, a) \in \text{UCRL}^2(s, a) \). It then implements the optimistic principle by trying to compute \( \pi^+_{t+k} = \arg \max_{\pi, S \rightarrow A} \max \{ \tilde{g}^M_{\pi} : M \in M_{t+k} \} \), where \( \tilde{g}^M_{\pi} \) is the average-gain for policy \( \pi \) in MDP \( M \). This is carried out approximately by \textbf{EVI} that builds a near-optimal policy \( \pi^+_{t+k} \) and an MDP \( \tilde{M}_t \) such that \( \tilde{g}^M_{\pi^+_{t+k}} \geq \max_{\pi, M \in M_{t+k}} \tilde{g}^M_{\pi} - \frac{1}{\sqrt{t}} \).

Finally, \textbf{UCRL2} does not recompute \( \pi^+_{t+k} \) at each time step. Instead, it proceeds in internal episodes, indexed by \( k \in \mathbb{N} \), where a near-optimalistic policy \( \pi^+_{t+k} \) is computed only at the starting time of each episode. Letting \( t_k \) denote the starting time of episode \( k \), the algorithm computes \( \pi^+_{t+k} := \pi^+_{t_k} \) 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 \( k > 1 \),

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

where \( v_{t_1} : \mathcal{A} \rightarrow \mathbb{R} \) denotes the number of observations of \((s, a) \in \mathcal{A}\) between time \( t_1 \) and \( t_2 \), and where we recall that for \( x \in \mathbb{R}, x^+ := \max\{x, 1\} \). The \textbf{EVI} algorithm writes as presented in Algorithm 1.

**Algorithm 1 Extended Value Iteration (EVI)**

\[
\begin{align*}
\text{Input: } & e_t \\
& u_0 \equiv 0, u_{-1} \equiv -\infty, n = 0 \\
& \text{while } E(u_n - u_{n-1}) > e_t \text{ do} \\
& \quad \text{Compute } \left\{ \mu^+ : s, a \rightarrow \max_{\mu' : \mu' \in \text{UCRL}^2(s, a)} \left\{ p^+_n : s, a \rightarrow \arg \max \left\{ P^a_{n+1}(s, s') : a \in A \right\} \right\} \right\} \\
& \quad \text{Update } \left\{ \pi^+_{n+1}(s) = \max_\pi \left( \mu^+(s, a) + P^a_{n+1}(s, s') : a \in A \right) \right\} \\
& \quad n = n + 1 \\
& \text{end while}
\end{align*}
\]

3. The \textbf{UCRL3} Algorithm

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

(i) While being a theoretically appealing strategy, \textbf{UCRL2} suffers from conservative confidence intervals, yielding bad empirical performances. Indeed the random stopping times \( N_t(s, a) \) are handled using simple union bounds, causing large confidence bounds. The first modification we introduce has thus the same design as \textbf{UCRL2}, but it replaces these confidence bounds with tighter time-uniform concentration inequalities. Further, unlike \textbf{UCRL2}, it 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 \textbf{SCAL} or \textbf{UCRL2B}. Indeed one drawback of \( L_1 \) confidence bounds is that they require an upper bound on the size of the support of the distribution. Without further knowledge, only \( S \) can be provided. In \textbf{UCRL2}, this causes the factor \( S \) to appear inside the square-root, due to a union bound over \( 2^S \) terms. Deriving \( L_1 \) confidence bounds adaptive to the support size seems challenging. In stark contrast, entry-wise confidence bounds can be used without knowing the support: when \( p(s'|s, a) \) has a support much smaller than \( S \), this may lead to a substantial improvement. Hence, the modified strategy \textbf{UCRL3} relies on time-uniform Bernoulli concentration bounds (presented in Section 3.1 below).

(ii) In order to further tighten exploration, the second idea is to revisit \textbf{EVI} to compute a refined optimistic policy at each round. Indeed, the optimization procedure used in \textbf{EVI} considers all plausible transition probabilities without support restriction, causing unwanted exploration. We introduce a novel value iteration procedure, which we call \textbf{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 now introduce the following high probability confidence sets for the mean rewards: For each \((s, a) \in S \times A\),

\[
c_{t, b_0}(s, a) = \left\{ \mu' \in [0, 1] : \left| \hat{\mu}_t(s, a) - \mu' \right| \leq b_t r_{t, b_0/(S \mathcal{A})}(s, a) \right\},
\]

where we introduced the notation

\[
b_t r_{t, b_0/(S \mathcal{A})}(s, a) := \max \left\{ \frac{1}{2} \beta N_t(s, a) \left( \frac{\delta_1}{S^A} \right), \sqrt{2\tilde{\sigma}^2_t(s, a) \frac{\ell_{N_t(s, a)}(\delta_1)}{N_t(s, a)} + \frac{7\ell_{N_t(s, a)}(\delta_1)}{3N_t(s, a)}} \right\},
\]

with \( \tilde{\sigma}^2_t(s, a) \) denoting the empirical variance of the reward function of \((s, a) \) built using the observations gathered up to time \( t \), and where \( \ell_n(\delta) = \eta \log \left( \frac{\log(n) \log(m)}{\log^2(\eta) \delta} \right) \) with \( \eta = 1.12. \)

Any \( \eta > 1 \) is valid, and \( \eta = 1.12 \) yields a small bound.
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 
\[
P\left( \exists t \in \mathbb{N}, (s, a) \in \mathcal{S} \times \mathcal{A}, \, \mu(s, a) \notin c_{t, \delta_0}(s, a) \right) \leq 3\delta_0.
\]

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 \((s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}\),
\[
C_{t, \delta_0}(s, a, s') = \left\{ q \in [0, 1]: \right\}
\]
\[
|\tilde{p}(s'|s, a) - q| \leq \sqrt{\frac{2q(1-q)}{N_t(s, a)} c_{N_t(s, a)} \left( \frac{\delta_0}{|S|} \right) + \frac{c_{N_t(s, a)} \left( \frac{\delta_0}{|S|} \right)}{3N_t(s, a)}},
\]
and 
\[
-\sqrt{g(q)} \leq \tilde{p}(s'|s, a) - q \leq \sqrt{g(q)}
\]
where 
\[
g(p) = \begin{cases} g(p) & \text{if } p < 0.5, \\ p(1-p) & \text{else,} \end{cases}
\]

The first inequality comes from Bernstein concentration inequalities, 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 (a.k.a. the Laplace method) 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, 2019; Raginsky & Sason, 2013). We discuss this in great detail in Appendix A.2.

UCRL3 finally considers the set of plausible MDPs \( M_{t, \delta} = \{ \tilde{M} = (\mathcal{S}, \mathcal{A}, \tilde{p}, \tilde{\nu}) \} \), where for each \((s, a) \in \mathcal{S} \times \mathcal{A}\),
\[
\tilde{\mu}(s, a) \in c_{t, \delta_0}(s, a),
\]
\[
\tilde{p}(\cdot|s, a) \in C_{t, \delta_0}(s, a) = \left\{ p' \in \Delta_S: \forall s', p'(s') \in C_{t, \delta_0}(s, a, s') \right\}.
\]

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

**Lemma 1 (Time-uniform confidence bounds)** For any MDP with rewards bounded in \([0, 1]\), mean function \(\mu\) and transition function \(p\), for all \( \delta \in (0, 1) \), it holds
\[
P\left( \exists t \in \mathbb{N}, (s, a) \in \mathcal{S} \times \mathcal{A}, \, \mu(s, a) \notin c_{t, \delta_0}(s, a) \text{ or } p(\cdot|s, a) \notin C_{t, \delta_0}(s, a) \right) \leq \delta.
\]

### 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 \(q \in C_{t, \delta}(s, a)\)). 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 \(S\), 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 \( \tilde{S} \subset S \) 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 \((s, a)\):
\[
\tilde{f}_{s,a}(\tilde{S}) = \max_{\tilde{p} \in \tilde{X}} \sum_{s' \in \tilde{S}} f(s') \tilde{p}(s'), \quad \text{where (2)}
\]
\[
\tilde{X} = \left\{ \tilde{p}: \forall s' \in \tilde{S}, \tilde{p}(s') \in C_{t, \delta}(s, a, s') \text{ and } \sum_{s' \in \tilde{S}} \tilde{p}(s') \leq 1 \right\}.
\]

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

The original EVI (Algorithm 1) computes \( \tilde{f}_{s,a}(S) \) for the function \( f = u_i \) at each iteration \( i \). When \( p = p(\cdot|s, a) \) has a sparse support included in \( \tilde{S} \), \( C_{t, \delta}(s, a, s') \) often does not reduce to \( \{0\} \) for \( s' \notin \tilde{S} \), while one may prefer to force a solution with a sparse support. A naïve way to proceed is to define \( \tilde{S} \) as the empirical support (i.e., the support of \( \tilde{p}_t(\cdot|s, a) \)). Doing so, one however solves a different optimization problem than the one using the full set \( S \), which means we may lose the optimistic property (i.e., \( \tilde{f}_{s,a}(\tilde{S}) \geq \mathbb{E}_{X \sim p(\cdot|s, a)} [f(X)] \) may not hold) and get an uncontrolled error. Indeed, the following decomposition
\[
\mathbb{E}_{X \sim p}[f(X)] = \sum_{s' \in \tilde{S}} f(s') p(s') + \sum_{s' \notin \tilde{S}} f(s') p(s') ,
\]
shows that computing an optimistic value restricted on \( \tilde{S} \) 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 \( \tilde{S} \) contains the support of \( \tilde{p}_n \), thanks to Bernstein’s confidence bounds, it is easy to see\(^8\) that the first term in the above decomposition contains terms scaling as \( \tilde{O}(n^{-1/2}) \), while the error term contains only terms scaling as \( \tilde{O}(n^{-1}) \). On the other hand, the error term sums \(|\tilde{S}\setminus\tilde{S}|\) many elements, which can be large in case \(\rho\) is sparse, and thus may even exceed \(\bar{T}_{s,a}(\tilde{S})\) 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 \(f = u_t\) (i.e., the value function at iterate \(i\)), the support \(\tilde{S}\) of the empirical distribution \(\tilde{p}_t(\cdot|s, a)\), high-probability confidence sets \(C: \{C_{i,s,a}(s, a, s', s' \in \tilde{S}^t)\}\), and a parameter \(\kappa \in (0, 1)\). It then adaptively augments \(\tilde{S}\) in order to find a set \(\tilde{S}\), whose corresponding value function \(\bar{T}_{t,s,a}(\tilde{S})\) is near-optimistic, as formalized in the following lemma:

**Algorithm 2 NOSS\((f, \tilde{S}, C, \kappa)\)**

Let \(\tilde{S} = \tilde{S} \cup \arg\max_{s \notin \tilde{S}} f(s)\).

Define \(\tilde{T}\) using \(f\) and confidence sets \(C\) (see (2)).

while \(\tilde{T}(\tilde{S}\setminus \tilde{S}) \geq \min(\kappa, \tilde{T}(\tilde{S}))\) do

Let \(\tilde{s} \in \arg\max_{s \notin \tilde{S}} f(s)\)

\(\tilde{S} = \tilde{S} \cup \{\tilde{s}\}\)

end while

return \(\tilde{S}\)

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

\[\bar{T}_{t,s,a}(\tilde{S}) \geq \mathbb{E}_{X \sim p(\cdot|s, a)}[f(X)] - \min\{\kappa, \tilde{T}_{s,a}(\tilde{S}), \bar{T}_{s,a}(\tilde{S}\setminus \tilde{S})\} \]

in other words, the value function \(\bar{T}_{s,a}(\tilde{S})\) 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 in EVI

\[p^+_n : s, a \mapsto \arg\max_{p' \in C_{i,s,a}(s, a)} p'_n u_n, p' \in C_{i,s,a}(s, a)\]

with NOSS: For a state-action pair \((s, a)\) and at each iterate \(n\) of EVI, UCRL3 applies NOSS (Algorithm 2) to the function \(u_n - \min_s u_n(s)\) (that is, the relative optimistic value function), and empirical distribution \(\tilde{p}_t(\cdot|s, a)\). 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 \(\kappa\) to

\[\kappa = \kappa_{t,n}(s, a) = \frac{\gamma S(u_n)\sup(\tilde{p}_t(\cdot|s, a))}{\max_s N_t(s, a)^{2/3}}, \quad \text{where} \quad \gamma = 10.\]

The scaling with the size of the support and span of the considered function is intuitive. The reason to further normalize by \(\max_s N_t(s, a)^{2/3}\) is to deal with the case when \(N_t(s, a)\) is small: First, in the case of Bernstein’s bounds, and since \(\tilde{S}\) contains at least the empirical support, \(\min \{\tilde{T}_{s,a}(\tilde{S}), \bar{T}_{s,a}(\tilde{S}\setminus \tilde{S})\}\) should essentially scale as \(\tilde{O}(N_t(s, a)^{-1})\). Hence for pairs such that \(N_t(s, a)\) is large, \(\kappa\) 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 \(\max_s N_t(s, a)^{2/3}\) 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)** The EVI-NOSS algorithm satisfies that, using the stopping criterion \(\tilde{S}(u_{n+1} - u_n) \leq \varepsilon\), the average-reward gain \(g_{n+1}\) corresponding to the policy \(\pi_{n+1}\) and the MDP \(\tilde{M} = (S, A, \mu_{n+1}^{s,a}, P_{n+1}^{s,a})\) computed at the last iteration \(n + 1\), is near-optimistic, in the sense that with probability higher than \(1 - \delta\), uniformly over all \(t\), \(g_{n+1} \geq g^* - \varepsilon - \tilde{\pi}\), where \(\tilde{\pi} = \tilde{\pi}_{t,n} = \frac{\gamma S(u_n)\sup(\tilde{p}_t(\cdot|s, a))}{\max_s N_t(s, a)^{2/3}}\).

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:

\[L_{s,a} := \left(\sum_{x \in S} \sqrt{p(x|s, a)(1 - p(x|s, a))}\right)^2,\]

\[P_{n+1}^{s,a} := \left(\sum_{x \in S} \sqrt{p(x|s, a)(1 - p(x|s, a))}\right)^2,\]

\[\tilde{\pi}_{t,n} = \frac{\gamma S(u_n)\sup(\tilde{p}_t(\cdot|s, a))}{\max_s N_t(s, a)^{2/3}}\]

\[\tilde{\pi} = \tilde{\pi}_{t,n} = \frac{\gamma S(u_n)\sup(\tilde{p}_t(\cdot|s, a))}{\max_s N_t(s, a)^{2/3}}\]

\[\tilde{\pi} = \tilde{\pi}_{t,n} = \frac{\gamma S(u_n)\sup(\tilde{p}_t(\cdot|s, a))}{\max_s N_t(s, a)^{2/3}}\]

\[\tilde{\pi} = \tilde{\pi}_{t,n} = \frac{\gamma S(u_n)\sup(\tilde{p}_t(\cdot|s, a))}{\max_s N_t(s, a)^{2/3}}\]
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 \(v_0 = \delta/(3 + 3S)\). Set \(t_0 = 0\), \(t = 1\), \(k = 1\), and observe the initial state \(s_1\).
for episodes \(k \geq 1\) do
Set \(N_k = t_k\)
Set \(N_k(s, a) = N_k(s, a) + v_k(s, a)\) for all \((s, a)\)
Compute empirical estimates \(\hat{b}_k(s, a)\) and \(\hat{p}_k(s, a)\) for all \((s, a)\)
Compute (see Algorithm 3)
\[
\pi_k^+ = \text{EVI-NOSS} \left( \hat{b}_k, \pi_k, \delta_0, C_{t_k}, \delta_0, \max N_k(s, a), \frac{1}{\sqrt{t}} \right)
\]
while \(v_k(s_1, \pi_k^+(s_1)) < N_k(s_1, \pi_k^+(s_1))^+\) do
Play action \(a_t = \pi_k^+(s_t)\), and observe the next state \(s_{t+1}\)
and reward \(r_t(s, a_t)\).
Set \(\pi_k(s_t, a_t) = v_k(s_t, a_t) + 1\)
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 \((s, a)\).\(^8\) 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 \((s, a)\):
\[
L_{s,a} \leq K_{s,a} G_{s,a} \leq K_{s,a} - 1 \leq S - 1.
\]

Theorem 1 (Regret of UCRL3) With probability higher than \(1 - 4\delta\), uniformly over all \(T \geq 3\),
\[
\mathcal{R}(\text{UCRL3}, T) \leq c \sqrt{T \log \left( \frac{6S^2 A T}{\delta} \right)} + 10 D S^2 A \sqrt{T^3} + 2D. \text{Therefore, the regret of UCRL3 asymptotically grows as}
\[
\mathcal{O}(\sqrt{\sum_{s,a} (D^2 L_{s,a} \lor 1) + D} \sqrt{T \log(\sqrt{T}/\delta)}).
\]

we now discuss the regret bound of UCRL3 with respect to that of UCRL2B. As shown in Table 1, the latter algorithm attains a regret bound of \(\mathcal{O}(\sqrt{T \log(\sqrt{T}/\delta)}\). The two regret bounds are not directly comparable: The regret bound of UCRL2B depends on \(\sqrt{T}\) 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 \(K_{s,a}\).

\(^8\) We recall that for a pair \((s, a)\), we define \(K_{s,a} := \text{supp}(p(.|s, a))\), and denote its cardinality by \(K_{s,a}\).

4. Numerical Experiments

In this section we provide illustrative numerical experiments that show the benefit of UCRL3 over UCRL2 and some of its popular alternatives. We conduct numerical experiments to examine the performance of the proposed variants of UCRL2, and compare it to that of state-of-the-art algorithms such as UCRL2, KL-UCRL, and UCRL2B. For all algorithms, we set \(\delta = 0.05\) and use the same tie-breaking rule (see Appendix E).

In the first set of experiments, we consider a \(S\)-state RiverSwim environment (corresponding to the MDP shown in Figure 2). To better understand Theorem 1 for such environments, 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 \(\hat{\mathcal{R}}\), we introduce \(\mathcal{R}(\hat{\mathcal{R}})\) to denote the regret bound normalized by \(\sqrt{T \log(\sqrt{T}/\delta)}\) ignoring universal constants. For instance, \(\mathcal{R}(\text{UCRL2}) = D \sqrt{SA}\).\(^9\) In Table 2, we compare \(\hat{\mathcal{R}}\) for various algorithms, for \(S\)-state RiverSwim for several values of \(S\). Note that \(\mathcal{R}(\text{UCRL2B})\) grows with \(T\) unlike \(\mathcal{R}\) for UCRL2, SCAL\(^*\), and UCRL3. Note that even

\(^9\) 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.
Table 2. Comparison of the quantity $\overline{R}$ of various algorithms for $S$-state RiverSwim: $\overline{R}(\text{UCRL}2) = D \sum_{s,a} K_{s,a}$, $\overline{R}(\text{SCAL}^+) = D \sum_{s,a} K_{s,a} \log(T)$ for $T = 100$, and $\overline{R}(\text{UCRL}3) = \sqrt{\sum_{s,a} (D^2 L_{s,a})^+} + D$.

as shown in Figure 5, yielding MDPs with $A = 4$ actions and $S = 20$ states (respectively $S = 55$) when taking off walls. In such grid-worlds, the learner starts in the upper-left corner. A reward of 1 is placed in the lower-right corner, when this rewarding state is reached, the learner is sent back to the initial state. The learner can perform 4 actions: Going up, left, down, or right, there are probabilities of 0.1 to stay in the same state, and 0.1 to go in each of the two perpendicular directions (or stay if this leads to a wall), giving a 0.7 probability to go in the chosen direction.

In Figure 3, we plot the regret under UCRL2, KL-UCRL, UCRL2B, and UCRL3 examined in a RiverSwim environment with $S = 6$ states. The curves show the results averaged over 50 independent runs along with the first and 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 larger 25-state RiverSwim environment; On this environment, UCRL2 only starts learning after close to $10^7$ time steps.

We further provide results in larger MDPs. We consider two frozen lake environments of respective sizes $7 \times 7$ and $9 \times 11$ as shown in Figure 5, yielding MDPs with $A = 4$ actions and $S = 20$ states (respectively $S = 55$) when taking off walls. In such grid-worlds, the learner starts in the upper-left corner. A reward of 1 is placed in the lower-right corner, when this rewarding state is reached, the learner is sent back to the initial state. The learner can perform 4 actions: Going up, left, down, or right, there are probabilities of 0.1 to stay in the same state, and 0.1 to go in each of the two perpendicular directions (or stay if this leads to a wall), giving a 0.7 probability to go in the chosen direction.

**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 $S^2 A$ ($S A (S - 1)$ for the transition function, and $S A$ 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.

Figures 6 (respectively 7) show the regret performance of UCRL2, KL-UCRL, UCRL2B, and UCRL3 on these 2-room (respectively 4-room) grid-worlds MDPs. The full code and implementation details are made available to the community. 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 $3 \times 10^3$. UCRL3 still outperforms other state-of-the-art strategies 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 Environment Test-bench) (Bhatnagar et al., 2009), in which we can specify the number of states, actions, average size of the support of transition distributions, sparsity of the reward function, as well as the minimal non-zero probability mass and minimal non-zero mean-reward.
Comparing \textit{UCRL3} against \textit{UCRL2B} in experiments reveal the gain achieved is not only due to Bernstein’s confidence intervals. Let us recall that on top of using Bernstein’s confidence intervals, \textit{UCRL3} also uses a refinement using sub-Gaussianity of Bernoulli distributions as well as the \textit{EVI-NOSS} instead of \textit{EVI} for planning. Experimental results verify that both tight confidence sets (see also Figure 11 in the appendix) and \textit{EVI-NOSS} play an essential role in achieving small empirical regret.

\section{5. Conclusion}

We studied reinforcement learning in finite Markov decision processes (MDPs) under the average-reward criterion, and introduced \textit{UCRL3}, a refined variant of \textit{UCRL2} (Jaksch et al., 2010), that efficiently balances exploration and exploitation in communicating MDPs. The design of \textit{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 and high-probability regret bound for \textit{UCRL3} scaling as \(\tilde{O}(\sqrt{(D + \sqrt{\sum_{s,a}(D^2 s L_{s,a} + 1)} \sqrt{T})})\), 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 \textit{UCRL2}. Through numerical experiments we showed that \textit{UCRL3} significantly outperforms existing variants of \textit{UCRL2} in standard environments.

\section{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 Educa-
tion and Research, Inria, and the French Agence Nationale de la Recherche (ANR), under grant ANR-16-CE40-0002 (project BADASS). This work was done when M. S. Talebi was a postdoctoral researcher in Inria Lille – Nord Europe.

References


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 \(\sigma\)-sub-Gaussian noise if

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

where \(\mathcal{F}_{t-1}\) denotes the \(\sigma\)-algebra generated by \(Y_1, \ldots, Y_{t-1}\).

**Lemma 5 (Uniform confidence intervals)** Let \(Y_1, \ldots, Y_t\) be a sequence of i.i.d. real-valued random variables with mean \(\mu\), such that \(Y_t - \mu\) is \(\sigma\)-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

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

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

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

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

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

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

\[
E[M_t^\lambda] = E[M_1^\lambda] + \mathbb{E} \left[ \sum_{t=1}^{\infty} M_{t+1}^\lambda - M_t^\lambda \right] = 1 + \sum_{t=1}^{\infty} \mathbb{E}[(M_{t+1}^\lambda - M_t^\lambda) \{\tau > t\}]
\]

\[
= 1 + \sum_{t=1}^{\infty} \mathbb{E}[(E[M_{t+1}^\lambda | \mathcal{F}_t] - M_t^\lambda) \{\tau > t\}]
\]

\[
\leq 1.
\]

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 \(M_t = E[M_t^\lambda | \mathcal{F}_\infty]\). Note that the standard deviation of \(\Lambda\) is \(\sigma^{-1}\) due to the fact we consider \(\sigma\)-sub-Gaussian random variables. We immediately get \(E[M_t] = E[E[M_t^\lambda | \Lambda]] \leq 1\). For convenience, let \(S_t = t(\mu_t - \mu)\). By construction
of $M_t$, we have

\[ M_t = \frac{1}{\sqrt{2\pi \sigma^2}} \int_{\mathbb{R}} \exp \left( \lambda \frac{S_t}{\sigma} - \frac{\lambda^2 \sigma^2}{2} \right) d\lambda \]

\[ = \frac{1}{\sqrt{2\pi \sigma^2}} \int_{\mathbb{R}} \exp \left( - \left[ \frac{t+1}{2} - \frac{S_t}{\sigma \sqrt{2(t+1)}} \right]^2 + \frac{S_t^2}{2 \sigma^2(t+1)} \right) d\lambda \]

\[ = \exp \left( \frac{S_t^2}{2 \sigma^2(t+1)} \right) \frac{1}{\sqrt{2\pi \sigma^2}} \int_{\mathbb{R}} \exp \left( - \frac{\lambda^2 \sigma^2}{2} \right) d\lambda \]

\[ = \exp \left( \frac{S_t^2}{2 \sigma^2(t+1)} \right) \frac{\sqrt{2\pi \sigma^2}}{\sqrt{2\pi \sigma^2}} . \]

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:

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

\[ \square \]

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 $\varepsilon \geq 0$,

\[ \mathbb{P} \left( \frac{1}{n} \sum_{i=1}^{n} (X_i - p) \geq \varepsilon \right) \leq \exp \left( - n k_1(p, \varepsilon, p) \right) , \]

where $k_1(p, q) = p \log(p/q) + (1-p) \log((1-p)/(1-q))$ denotes the Bernoulli Kullback-Leibler divergence. The reverse map of the Cramèr transform $\varepsilon \mapsto k_1(p + \varepsilon, p)$ 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

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

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

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

We note that the left and right tails are not controlled in a symmetric way. This yields, introducing the function $\tilde{g}(p) = \begin{cases} g(p) & \text{if } p < 1/2 \\ p(1-p) & \text{otherwise} \end{cases}$, the following asymmetrical confidence set
Corollary 1 (Time-uniform Bernoulli concentration) Let \((X_i)_{i \leq n} \overset{i.i.d.}{\sim} \mathcal{B}(p)\). Then, for all \(\delta \in (0, 1)\),
\[
\mathbb{P} \left( \forall n \in \mathbb{N}, \ -\sqrt{g(p)} \beta_n(\delta) \leq \frac{1}{n} \sum_{i=1}^{n} X_i - p \leq \sqrt{g(p)} \beta_n(\delta) \right) \geq 1 - 2\delta,
\]
where \(\beta_n(\delta) := \sqrt{\frac{2}{n}} (1 + \frac{1}{n}) \log(\sqrt{n + 1} / \delta)\).

Proof of Corollary 1:

Let us introduce the following quantities
\[
\forall \lambda \in \mathbb{R}^+, \quad M_t^\lambda = \exp \left( \sum_{s=1}^{t} \left( \lambda (X_s - p) - \frac{\lambda^2 g(p)}{2} \right) \right),
\]
\[
\forall \lambda \in \mathbb{R}, \quad M_t'^\lambda = \exp \left( \sum_{s=1}^{t} \left( \lambda (X_s - p) - \frac{\lambda^2 g(p)}{2} \right) \right).
\]

Note that \(M_t^\lambda\) is a non-negative super-martingale for all \(\lambda \in \mathbb{R}^+\), and \(M_t'^\lambda\) is a non-negative super-martingale for all \(\lambda \in \mathbb{R}\). Further, \(\mathbb{E}[M_t^\lambda] \leq 1\) and \(\mathbb{E}[M_t'^\lambda] \leq 1\).

Let \(\lambda\) be a random variable with density
\[
f_p(\lambda) = \begin{cases} \frac{\exp(-\lambda^2 g(p)/2)}{\int_{\mathbb{R}^+} \exp(-z^2 g(p)/2) dz} = \sqrt{\frac{2g(p)}{\pi}} \exp(-\lambda^2 g(p)/2) & \text{if } \lambda \in \mathbb{R}^+, \\ 0 & \text{else.} \end{cases}
\]

Let \(M_t = \mathbb{E}[M_t^\lambda | \mathcal{F}_t]\) and note that
\[
M_t &= \sqrt{\frac{2g(p)}{\pi}} \int_{\mathbb{R}^+} \exp \left( \lambda S_t - \frac{\lambda^2 g(p) t}{2} - \frac{\lambda^2 g(p)}{2} \right) d\lambda \\
&= \sqrt{\frac{2g(p)}{\pi}} \int_{\mathbb{R}^+} \exp \left( - \left[ \lambda \sqrt{\frac{g(p)(t+1)}{2}} - \frac{S_t}{\sqrt{2g(p)(t+1)}} \right]^2 + \frac{S_t^2}{2g(p)(t+1)} \right) d\lambda \\
&= \exp \left( \frac{S_t^2}{2g(p)(t+1)} \right) \sqrt{\frac{2g(p)}{\pi}} \int_{\mathbb{R}^+} \exp \left( - \left( \lambda - \frac{S_t}{g(p)(t+1)} \right)^2 g(p) \frac{t+1}{2} \right) d\lambda \\
&\geq \exp \left( \frac{S_t^2}{2g(p)(t+1)} \right) \sqrt{\frac{2g(p)}{\pi}} \sqrt{\frac{\pi}{2(t+1)g(p)}} \quad \text{if } S_t \geq 0 \\
&= \exp \left( \frac{S_t^2}{2g(p)(t+1)} \right) \frac{1}{\sqrt{t+1}}.
\]

Note also that \(M_t\) is still a non-negative super-martingale satisfying \(\mathbb{E}[M_t] \leq 1\) for all \(t\). Likewise, considering \(\Lambda'\) to be a random variable with density
\[
f_{\Lambda'}(\lambda) = \begin{cases} \frac{\exp(-\lambda^2 g(p)/2)}{\int_{\mathbb{R}^+} \exp(-z^2 g(p)/2) dz} = \sqrt{\frac{2g(p)}{\pi}} \exp(-\lambda^2 g(p)/2) & \text{if } \lambda \in \mathbb{R}^-, \\ 0 & \text{else.} \end{cases}
\]
Introducing $M'_t = \mathbb{E}[M_t^{\Lambda_t} \mid \mathcal{F}_t]$, it comes

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

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

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

which implies

$$-\sqrt{2g(p)\frac{(1+1/t)}{t} \log (M_t' \sqrt{1+t})} \leq \frac{S_t}{t} \leq \sqrt{2g(p)\frac{(1+1/t)}{t} \log (M_t \sqrt{1+t})}.$$

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

$$\mathbb{P}\left( \exists t, \frac{S_t}{t} \geq \sqrt{2g(p)\frac{(1+1/t)}{t} \log (\sqrt{1+t}/\delta)} \text{ or } \frac{S_t}{t} \leq -\sqrt{2g(p)\frac{(1+1/t)}{t} \log (\sqrt{1+t}/\delta)} \right)$$

$$\leq \mathbb{P}\left( \exists t, M_t \geq 1/\delta \text{ or } M'_t \geq 1/\delta \right)$$

$$\leq \mathbb{P}(\exists t, M_t \geq 1/\delta) + \mathbb{P}(\exists t, M'_t \geq 1/\delta)$$

$$\leq \delta(\mathbb{E}[\max_t M_t] + \mathbb{E}[\max_t M'_t])$$

$$\leq 2\delta.$$

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

---

**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 $\mathcal{F} = (\mathcal{F}_t)$ 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 $\varphi_*$ denote its Legendre-Fenchel transform, that is:

$$\forall \lambda \in \mathcal{D}, \forall t, \quad \log \mathbb{E} \left[ \exp (\lambda Z_t) \mid \mathcal{F}_{t-1} \right] \leq \varphi(\lambda),$$

$$\forall x \in \mathbb{R}, \quad \varphi_*(x) = \sup_{\lambda \in \mathbb{R}} (\lambda x - \varphi(\lambda)),$$

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

$$\varphi_*^{-1}(z) := \sup\{ x \leq 0 : \varphi_*(x) > z \} \quad \text{and} \quad \varphi_*^{-1}(z) := \inf\{ x \geq 0 : \varphi_*(x) > z \}.$$

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 $\delta \in (0,1)$,

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

$$\mathbb{P}\left[ \frac{1}{N_n} \sum_{t=1}^{N_n} Z_t \leq \varphi_*^{-1}(\frac{\eta}{N_n} \log \left( \left\lceil \log(n) \log(\eta) \right\rceil \frac{1}{\delta} \right)) \right] \leq \delta.$$
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 one, whereas below this line, the second Bernstein bound is sharper.

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

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

$$
P \left[ \frac{1}{N} \sum_{t=1}^{N} Z_t \leq \frac{1}{N} \log \left( \frac{\log(N) \log(\eta N)}{\delta \log^2(\eta)} \right) \right] \leq \delta.$$

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

$$\varphi_{s,+}^{-1}(z) = \sqrt{2vz} + bz$$

and

$$\varphi_{s,-}^{-1}(z) = -\sqrt{2vz} - bz.$$

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

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

$$C^{ Bernstein-D}(p,n,\delta) = p + \sqrt{\frac{2p}{n} \ell_n(\delta) + \frac{\ell_n(\delta)}{n}}$$

where

$$\ell_n(\delta) = 2 \log \log(\max(e,n)) + \log(3/\delta)$$

(4)

and

$$C^{ Bernstein-M}(p,n,\delta) = p + \sqrt{\frac{2p(1-p)}{n} \ell_n(\delta) + \frac{\ell_n(\delta)}{3n}}$$

where

$$\ell_n(\delta) = \eta \log \left( \frac{\log(n) \log(\eta n)}{\log^2(\eta)} \right)$$

with $\eta = 1.12$.

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

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

$$C_{\text{ex-Gaussian-Laplace}}(p, n, \delta) = p + \sqrt{\frac{2g(p)(1 + \frac{1}{n}) \log(2\sqrt{n} + 1/\delta)}{n}},$$

and plot in Figure 11 the ratio $r(p, n, \delta) = C_{\text{ex-Gaussian-Laplace}}(p, n, \delta)/C_{\text{peeling}}(p, n, \delta)$ 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

$$\mathbb{E}_{S \sim p}[f(S)] = \sum_{s' \in \tilde{S}} f(s')p(s') + \sum_{s' \notin \tilde{S}} f(s')p(s')$$

$$\leq \overline{f}(\tilde{S}) + \sum_{s' \notin \tilde{S}} f(s')p(s')$$

$$\leq \overline{f}(\tilde{S}) + \min (\kappa, \overline{f}(\tilde{S}), \overline{f}(S \setminus \tilde{S}))$$

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 $\tilde{S}$ containing the empirical support $\hat{S}_n$ (plus eventually one point), and satisfies the condition $\overline{f}(S \setminus \tilde{S}) < \min (\kappa, \overline{f}(\tilde{S}))$ required to exit the loop. \qed
We first provide the following time-uniform concentration inequality to control a bounded martingale difference sequence, which follows from Lemma 5:

In this section, we prove Theorem 1. Our proof follows similar lines as in the proof of (Jaksch et al., 2010, Theorem 2).

**Corollary 2 (Time-uniform Azuma-Hoeffding)** Let \( \kappa_t \) the constant function equal to \( g_* \), and \( \kappa_t \) the constant function equal to \( \kappa_t \). Using vector notations, we have on the one hand

\[
\mathbf{g}_t = \mathcal{P}_t [\mu_* + P_* u_t^+ - u_t^+] \\
\leq \mathcal{P}_t [\mu_*^+ + P_{t, n} u_n^+ - u_n^+] \text{ w.p. } 1 - \delta \\
\leq \mathcal{P}_t [\mu_{n+1}^+ + P_{n+1} u_n^+ - u_n^+] + \mathcal{P}_t \kappa_t \text{ by optimality of } \pi_{n+1}^+ \\
= \mathcal{P}_t [u_{n+1}^+ - u_n^+] + \mathcal{P}_t \kappa_t .
\]

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

\[
g_{n+1}^+ = \mathcal{P}_{n+1} [\mu_{n+1}^+ + P_{n+1} u_n^+ - u_n^+] = \mathcal{P}_{n+1} [u_{n+1}^+ - u_n^+]
\]

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

\[
g_* - g_{n+1}^+ \leq \mathcal{P}_t [u_{n+1}^+ - u_n^+] - \mathcal{P}_{n+1} [u_{n+1}^+ - u_n^+] + \mathcal{P}_t \kappa_t \leq \mathcal{S}(u_{n+1}^+ - u_n^+) + \|\mathcal{P}_t\|_\infty \|\kappa_t\|_\infty \leq \varepsilon + \kappa_t .
\]

**Proof of Theorem 1:**

Let us denote by \( \kappa \) an optimal policy. Let \( \mathbf{g}_* : \mathcal{S} \rightarrow \mathbb{R} \) denote the constant function equal to \( g_* \), and \( \kappa_t \) the constant function equal to \( \kappa_t \). Using vector notations, we have on the one hand

\[
g_* = \mathcal{P}_t [\mu_* + P_* u_t^+ - u_t^+] \\
\leq \mathcal{P}_t [\mu_*^+ + P_{t, n} u_n^+ - u_n^+] \text{ w.p. } 1 - \delta \\
\leq \mathcal{P}_t [\mu_{n+1}^+ + P_{n+1} u_n^+ - u_n^+] + \mathcal{P}_t \kappa_t \text{ by optimality of } \pi_{n+1}^+ \\
= \mathcal{P}_t [u_{n+1}^+ - u_n^+] + \mathcal{P}_t \kappa_t .
\]

**C. Regret Analysis of UCRL3: Proof of Theorem 1**

In this section, we prove Theorem 1. Our proof follows similar lines as in the proof of (Jaksch et al., 2010, Theorem 2).

We first provide 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 \( (X_t) \) 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

\[
P\left( \exists T \in \mathbb{N} : \sum_{t=1}^{T} X_t \geq (b - a) \sqrt{\frac{1}{2} (T + 1) \log(\sqrt{T + 1} / \delta)} \right) \leq \delta .
\]

**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 \( \delta \) (for example \( \mathcal{M}_k := \mathcal{M}_{t_k, t_k} \)). Denote by \( m(T) \) the number of episodes initiated by the algorithm up to time \( T \). By applying Corollary 2, we deduce that

\[
\mathcal{R}(T) = \sum_{t=1}^{T} g_* - \sum_{t=1}^{T} r_t (s_t, a_t) \leq \sum_{s,a} N_{m(T)} (s, a) (g_* - \mu(s, a)) + \sqrt{\frac{1}{2} (T + 1) \log(\sqrt{T + 1} / \delta)} ,
\]

with probability at least \( 1 - \delta \). We have

\[
\sum_{s,a} N_{m(T)} (s, a) (g_* - \mu(s, a)) = \sum_{k=1}^{m(T)} \sum_{s,a} \sum_{t=t_k}^{t_{k+1}} I(s_t = s, a_t = a) (g_* - \mu(s, a)) = \sum_{k=1}^{m(T)} \sum_{s,a} \nu_k (s, a) (g_* - \mu(s, a)) .
\]
Introducing $\Delta_k := \sum_{s,a} \nu_k(s,a)(g_* - \mu(s,a))$ for $1 \leq k \leq m(T)$, we get

$$\mathcal{R}(T) \leq \sum_{k=1}^{m(T)} \Delta_k + \sqrt{\frac{2}{2}(T+1) \log \left(\frac{\sqrt{T} + 1}{\delta}\right)},$$

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, for all $T$, and for all episodes $k = 1, \ldots, m(T)$, the set $\mathcal{M}_k$ contains the true MDP with probability higher than $1 - \delta$. As a consequence, with probability at least $1 - \delta$, $\sum_{k=1}^{m(T)} \Delta_k \{M \notin \mathcal{M}_k\} = 0$.

**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 $\pi_k^+$ and $\mathcal{M}_k$, using Lemma 3, we get that

$$g_k := g_{\pi_k^+} \geq g_* - \frac{1}{\sqrt{T_k}} - \pi_k^-,$$

where $\pi_k^- = \frac{\nu_k}{\max_{s,a} \nu_k(s,a)}$. Hence, with probability greater than $1 - \delta$,

$$\Delta_k \leq \sum_{s,a} \nu_k(s,a)(g_k - \mu(s,a)) + \sum_{s,a} \nu_k(s,a)\left(\frac{1}{\sqrt{T_k}} + \pi_k^+\right). \quad (7)$$

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: $\max_s u_k^{(i)}(s) - \min_s u_k^{(i)}(s) \leq D$. The convergence criterion of EVI-NOSS implies

$$|u_k^{(i+1)}(s) - u_k^{(i)}(s) - g_k| \leq \frac{1}{\sqrt{T_k}}, \quad \forall s \in \mathcal{S}. \quad (8)$$

Using the Bellman operator on the optimistic MDP:

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

Substituting this into (8) gives

$$\left| \left( g_k - \tilde{\mu}_k(s,a) \right) - \left( \sum_{s'} \tilde{p}_k(s'|s,a)u_k^{(i)}(s') - u_k^{(i)}(s) \right) \right| \leq \frac{1}{\sqrt{T_k}}, \quad \forall s \in \mathcal{S}.$$

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

$$\left| g_k - \tilde{\mu}_k - (\tilde{p}_k - \mathbf{1})u_k^{(i)} \right| \leq \frac{1}{\sqrt{T_k}} \mathbf{1}.$$

Combining this with (7) yields

$$\Delta_k \leq \sum_{s,a} \nu_k(s,a)(g_k - \mu(s,a)) + \sum_{s,a} \nu_k(s,a)\left(\frac{2}{\sqrt{T_k}} + \pi_k^+\right)$$

$$= \sum_{s,a} \nu_k(s,a)(g_k - \tilde{\mu}_k(s,a)) + \sum_{s,a} \nu_k(s,a)(\tilde{\mu}_k(s,a) - \mu(s,a)) + \sum_{s,a} \nu_k(s,a)\left(\frac{2}{\sqrt{T_k}} + \pi_k^+\right)$$

$$\leq \nu_k(\tilde{p}_k - \mathbf{1})u_k^{(i)} + \sum_{s,a} \nu_k(s,a)(\tilde{\mu}_k(s,a) - \mu(s,a)) + \sum_{s,a} \nu_k(s,a)\left(\frac{2}{\sqrt{T_k}} + \pi_k^+\right).$$
Similarly to (Jaksch et al., 2010), we define \( w_k(s) := u_k^{(i)}(s) - \frac{1}{T} \left( \min_s u_k^{(i)}(s) + \max_s u_k^{(i)}(s) \right) \) for all \( s \in \mathcal{S} \). Then, in view of the fact that \( \mathbf{P}_k \) is row-stochastic, we obtain

\[
\Delta_k \leq \nu_k(\mathbf{P}_k - \mathbf{I}) w_k + \sum_{s,a} \nu_k(s,a) (\tilde{\mu}_k(s,a) - \mu(s,a)) + \sum_{s,a} \nu_k(s,a) \left( \frac{2}{\sqrt{k}} + \tau_k \right). \tag{9}
\]

The second term in the right-hand side can be upper bounded as follows: \( M \in \mathcal{M}_k \) implies

\[
\tilde{\mu}_k(s,a) - \mu(s,a) \leq \beta N_k(s,a) (3SA(1+S)) (s,a) \\
\leq \gamma \log \left( \frac{6SA(S+1)N_k(s,a)}{\delta} \right) \\
\leq \gamma \log \left( \frac{6SA \sqrt{T + 1}}{\delta} \right),
\]

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

\[
\sum_{s,a} \nu_k(s,a) \left( \frac{1}{\sqrt{t_k}} + \tau_k \right) \leq 2 \sum_{s,a} \nu_k(s,a) + \gamma D K \sum_{s,a} \nu_k(s,a) N_k(s,a)^{2/3}.
\]

Putting together, we obtain

\[
\Delta_k \leq \nu_k(\mathbf{P}_k - \mathbf{I}) w_k + \left( \sqrt{4 \log \left( \frac{6SA \sqrt{T + 1}}{\delta} \right)} + 2 \right) \sum_{s,a} \nu_k(s,a) N_k(s,a)^{2/3} + \gamma D K \sum_{s,a} \nu_k(s,a) N_k(s,a)^{2/3}. \tag{10}
\]

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

\[
\nu_k(\mathbf{P}_k - \mathbf{I}) w_k = \nu_k(\mathbf{P}_k - \mathbf{P}_k) w_k \tag{L_1(k)} + \nu_k(\mathbf{P}_k - \mathbf{I}) w_k \tag{L_2(k)}.
\]

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

**Lemma 8** Consider a good episode \( k \). Then,

\[
L_1(k) \leq \sqrt{2 \log \left( \frac{\delta}{6SA} \right)} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{N_k(s,a)}} D^2_k \left( 2 \log \left( \frac{\delta}{6SA} \right) \right) \sum_{s,a} \nu_k(s,a) N_k(s,a)^{2/3}.
\]

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

\[
\sum_{k=1}^{m(T)} L_2(k) I\{ M \in \mathcal{M}_k \} \leq D \sqrt{2(T + 1) \log(\sqrt{T + 1} / \delta)} + DS A \log_2 \left( \frac{2T}{\delta} \right).
\]

Applying Lemmas 8 and 9, and summing over all good episodes, we obtain the following bound that holds with probability
To simplify the above bound, we provide the following lemma:

**Lemma 10** We have:

\[
\begin{align*}
&\sum_{k=1}^{m(T)} \Delta_k I\{M \in \mathcal{M}_k\} \leq \sum_{k=1}^{m(T)} L_1(k) + \sum_{k=1}^{m(T)} L_2(k) \\
&+ \left(\sqrt{4 \log (6S^2A \sqrt{T+1/\delta}) + 2}\right) \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^+} + \gamma DK \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^{2/3}} \\
&\leq \sqrt{2\ell_T(\frac{\delta}{6S^2A})} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^+} D_s \sqrt{L_{s,a}} + 4DS\ell_T(\frac{\delta}{6S^2A}) \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^+} \\
&+ \left(\sqrt{4 \log (6S^2A \sqrt{T+1/\delta}) + 2}\right) \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^+} \\
&+ D\sqrt{2(T+1) \log(\sqrt{T+1/\delta}) + DSA \log_2(\frac{8T}{\delta\gamma}) + \gamma DK \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^{2/3}}}
\end{align*}
\]  

(11)

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

**Lemma 10** We have:

\[
\begin{align*}
&\sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^+} \leq (\sqrt{2} + 1) \sqrt{SAT}. \\
&(ii) \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^+} D_s \sqrt{L_{s,a}} \leq (\sqrt{2} + 1) \sum_{s,a} D_k^2 L_{s,a} T. \\
&(iii) \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^+} \leq 2SA \log_2(\frac{T}{SA}) + SA. \\
&(iv) \sum_{k=1}^{m(T)} \sum_{s,a} \frac{\nu_k(s,a)}{N_k(s,a)^+} \leq 3(SA)^{2/3} T^{1/3} + 2SA.
\end{align*}
\]

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

\[
\mathcal{R}(T) \leq \left(\sqrt{4 \log (6S^2A \sqrt{T+1/\delta}) + 2}\right) \sqrt{SAT} + (D\sqrt{2} + \sqrt{\frac{1}{2}}) \sqrt{(T+1) \log(\sqrt{T+1/\delta})} \\
+ \sqrt{2\ell_T(\frac{\delta}{6S^2A})} (\sqrt{2} + 1) \sum_{s,a} D_k^2 L_{s,a} T \sqrt{T} + 4D\ell_T(\frac{\delta}{6S^2A}) \left(2SA \log_2(\frac{T}{SA}) + SA\right) \\
+ 10DSA^{2/3} T^{1/3} + 20D S A.
\]

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

\[
\mathcal{R}(T) \leq \left(5 \sqrt{\sum_{s,a} D_k^2 L_{s,a} + 4\sqrt{SA} + 2D}\right) \sqrt{T \log(\frac{6S^2A \sqrt{T+1}}{\delta})} + 10DSA^{2/3} T^{1/3} + O\left(DS^2 A \log^2(\frac{T}{\delta})\right).
\]

Finally we remark that

\[
5 \sqrt{\sum_{s,a} D_k^2 L_{s,a} + 4\sqrt{SA}} \leq 10 \sqrt{SA} + \sum_{s,a} D_k^2 L_{s,a} \leq 10 \sqrt{2 \sum_{s,a} (D_k^2 L_{s,a} + 1)},
\]
so that $\mathcal{R}(T) = \mathcal{O}\left(\left[\sum_{s,a} (D^2 L_{s,a} + 1) + D\right] \sqrt{T \log(\sqrt{T}/\delta)}\right)$.

C.1. Proof of Technical Lemmas

Proof of Lemma 8:

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

$$L_1(k) = \sum_{s,x} \nu_k(s, \pi^+_k(s)) \left(\tilde{p}_k(x|s, \pi^+_k(s)) - p(x|s, \pi^+_k(s))\right) w_k(x) $$

$$\leq \sum_{s,a} \nu_k(s, a) \sum_x (\tilde{p}_k(x|s, a) - p(x|s, a)) w_k(x).$$

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

$$\sum_x (\tilde{p}_k(x|s, a) - p(x|s, a)) w_k(x) = \sum_x (\tilde{p}_k(x) - p(x)) w_k(x)$$

$$\leq \sum_x |\tilde{p}_k(x) - p(x)||w_k(x)| + \sum_x |\tilde{p}_k(x) - \tilde{p}_k(x)||w_k(x)|.$$

To upper bound $F_1$, we first show that $\max_{x \in \text{supp}(\tilde{p}_k(\cdot|s, a))} |w_k(x)| \leq \frac{D_k}{2}$. 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 $\mathcal{M}_k$ with continuous action space $\mathcal{A}'$. In this extended MDP, in each state $s \in S$, and for each $a \in A$, there is an action in $\mathcal{A}'$ with mean $\tilde{\mu}(s, a)$ and transition $\tilde{\pi}(\cdot|s, a)$ satisfying (1). Similarly to (Jaksch et al., 2010), we note that $u^{(i)}_k(s)$ amounts to the total expected $i$-step reward of an optimal non-stationary $i$-step policy starting in state $s$ on the MDP $\mathcal{M}_k$ with 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 $\mathcal{M}$, and therefore, the actions of the true MDP are contained in the continuous action set of the extended MDP $\mathcal{M}_k$. Now, if there were states $s_1, s_2 \in \bigcup_a \text{supp}(\tilde{p}_k(\cdot|s, a))$ with $u^{(i)}_k(s_1) - u^{(i)}_k(s_2) > D_s$, then an improved value for $u^{(i)}_k(s_1)$ could be achieved by the following non-stationary policy: First follow a policy which 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 $u^{(i)}_k(s_1) \geq u^{(i)}_k(s_2) - D_s$, 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:

$$F_1 \leq \sqrt{\frac{2\ell N_k}{N_k}} \sum_x \sqrt{p(x)(1 - p(x))} ||w_k(x)|| + \frac{S\ell N_k}{3N_k} ||w_k||_\infty$$

$$\leq \max_{x \in \mathcal{K}_{s,a}} |w_k(x)| \sqrt{\frac{2\ell N_k}{N_k}} \sum_x \sqrt{p(x)(1 - p(x))} + \frac{DS\ell N_k}{6N_k}$$

$$\leq D_s \sqrt{\frac{\ell N_k}{2N_k} \sum_x p(x)(1 - p(x))} + \frac{DS\ell N_k}{6N_k}$$

$$= D_s \sqrt{\frac{\ell N_k}{2N_k} L_{s,a}} + \frac{DS\ell N_k}{6N_k},$$

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

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 $\max_{x \in \text{supp}(\tilde{p}_k(\cdot|s,a))} |w_k(x)| \leq \frac{D_s}{2}$ yield:
\[
F_2 \leq \sqrt{\frac{2\ell_{N_k}}{N_k}} \sum_x \sqrt{\tilde{p}_k(x)(1-\tilde{p}_k(x))|w_k(x)|} + \frac{DS\ell_{N_k}}{6N_k}
\leq D_s\sqrt{\frac{\ell_{N_k}}{2N_k}} \sum_x \sqrt{\tilde{p}_k(x)(1-\tilde{p}_k(x))} + \frac{DS\ell_{N_k}}{6N_k}
\leq D_s\sqrt{\frac{\ell_{N_k}}{2N_k}} \sum_x \sqrt{p(x)(1-p(x))} + 2.4\sqrt{2} \frac{DS\ell_{N_k}}{N_k} + \frac{DS\ell_{N_k}}{6N_k}
\leq D_s\sqrt{\frac{\ell_{N_k}}{2N_k}} \sum_x \sqrt{p(x)(1-p(x))} + 3.6\frac{DS\ell_{N_k}}{N_k}.
\]

Combining the bounds on $F_1$ and $F_2$, and noting that
\[
\ell_{N_k(s,a)} \left( \frac{\delta}{3(1+S)SA} \right) \leq \ell_{N_k(s,a)} \left( \frac{\delta}{6S^2A} \right) \leq \ell_T \left( \frac{\delta}{6S^2A} \right)
\]
complete the proof. □

Proof of Lemma 11:

By Taylor’s expansion, we have
\[
y(1-y) = x(1-x) + (1-2x)(y-x) - (y-x)^2
= x(1-x) + (1-x-y)(y-x)
\leq x(1-x) + |1-x-y| \left( \sqrt{2y(1-y)}\zeta + \frac{1}{3}\zeta \right)
\leq x(1-x) + \sqrt{2y(1-y)}\zeta + \frac{1}{3}\zeta.
\]

Using the fact that $a \leq b\sqrt{a} + c$ implies $a \leq b^2 + b\sqrt{c} + c$ for nonnegative numbers $a$, $b$, and $c$, we get
\[
y(1-y) \leq x(1-x) + \frac{1}{3}\zeta + \sqrt{2\zeta \left( x(1-x) + \frac{1}{3}\zeta \right)} + 2\zeta
\leq x(1-x) + \sqrt{2\zeta x(1-x) + 3.15\zeta}
= \left( \sqrt{x(1-x)} + \sqrt{\frac{1}{2}\zeta} \right)^2 + 2.65\zeta,
\]
where we have used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ valid for all $a, b \geq 0$. Taking square-root from both sides and using the latter inequality give the desired result:
\[
\sqrt{y(1-y)} \leq \sqrt{x(1-x)} + \sqrt{\frac{1}{2}\zeta + \sqrt{2.65\zeta}} \leq \sqrt{x(1-x)} + 2.4\sqrt{\zeta}.
\]
□
Proof of Lemma 9:

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

\[
L_2(k) = \nu_k (P_k - I) w_k = \sum_{t=t_k}^{t_k+1} (p(\cdot|s_t, a_t) - e_{s_t})w_k
\]

\[
= \sum_{t=t_k}^{t_k+1} \left(p(\cdot|s_t, a_t) - e_{s_t+1} + e_{s_t+1} - e_{s_t}\right)w_k
\]

\[
= \sum_{t=t_k}^{t_k+1} X_t + w_k(s_{t+1}) - w_k(s_t) \leq \sum_{t=t_k}^{t_k+1} X_t + D,
\]

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

\[
|X_t| \leq \|p(\cdot|s_t, a_t) - e_{s_t+1}\|_1 \frac{D}{2} \leq \left(\|p(\cdot|s_t, a_t)\|_1 + \|e_{s_t+1}\|_1\right)\frac{D}{2} = D.
\]

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 \((X_t)_t\) is martingale difference sequence. Therefore, by Corollary 2, we get:

\[
P\left(\exists T: \sum_{t=1}^{T} X_t \geq D\sqrt{2(T + 1) \log(\sqrt{T + 1}/\delta)}\right) \leq \delta.
\]

Hence, we deduce that with probability at least \(1 - \delta\), the result holds. \(\Box\)

C.2. Proof of Supporting Lemmas

Proof of Lemma 10:

Inequalities (i)-(iii) easily follow by applying Lemma 12, which is stated at the end of this proof, and using Jensen’s inequality. Next we prove the inequality (iv).
Following similar steps as in the proof of (Ouyang et al., 2017, Lemma 5), we have

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

where we have used that for any \(L \geq 1\), \(\sum_{j=1}^{L} j^{-2/3} \leq 1 + \int_{1}^{L} z^{-2/3} \, dz \leq \frac{3}{2} L^{1/3}\), and where the last step follows from Jensen’s inequality. \(\square\)

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

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

**D. Other Technical Lemmas**

**Proof of Lemma 4:**

The notion of local effective support could be related to Gini index of the transition distribution of \((s, a)\) as follows. Given pair \((s, a)\), let us introduce the Gini index \(G_{s,a} := \sum_{x \in \mathcal{S}} p(x|s, a)(1 - p(x|s, a))\). Given \(\varepsilon > 0\), we introduce \(K_{s,a}^{\varepsilon} := \{x \in \mathcal{S} : p(x|s, a)(1 - p(x|s, a)) \geq \varepsilon\}\), of cardinality \(K_{s,a}^{\varepsilon}\). We have that \(L_{s,a} \leq \min_{\varepsilon > 0} \left( \sqrt{K_{s,a}^{\varepsilon} G_{s,a} + K_{s,a} \varepsilon} \right)^2\).

Indeed, applying Cauchy-Schwarz gives

\[
L_{s,a} = \sum_{x \in K_{s,a}^{\varepsilon}} \sqrt{p(x)(1 - p(x))} + \sum_{x \notin K_{s,a}^{\varepsilon}} \sqrt{p(x)(1 - p(x))} \\
\leq \sqrt{K_{s,a}^{\varepsilon} \sum_{x \in K_{s,a}^{\varepsilon}} p(x)(1 - p(x)) + (K_{s,a} - K_{s,a}^{\varepsilon}) \varepsilon} \\
\leq \sqrt{K_{s,a}^{\varepsilon} G_{s,a} + K_{s,a} \varepsilon},
\]
where we have used the Cauchy-Schwarz inequality.

Note that the inequality above implies \( L_{s,a} \leq K_{s,a} G_{s,a} \). Furthermore, in view of the concavity of \( z \mapsto \sum_{x \in S} z(x)(1 - z(x)) \), the maximal value of \( G_{s,a} \) is achieved when \( p(x|s,a) = \frac{1}{K_{s,a}} \) for \( x \in K_{s,a} \). Hence, \( G_{s,a} \leq 1 - \frac{1}{K_{s,a}} \). Therefore, \( L_{s,a} \leq K_{s,a} G_{s,a} \leq K_{s,a} - 1 \).

\[\square\]

E. 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 \( \pi_n^+ \) 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 \( \text{Argmin}_{a \in A} N_k(s,a) \). 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 \( \theta^n \). In such a situation where \( \hat{p}_t(s_0|s,a) = 1 \), 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 \( \pi \) using the Bellman optimal operator at every single step. In order to reduce computation, we use a lazy implementation that keeps updating \( u \) and \( \pi \) 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 python3 and is designed to be compatible with openAI gym discrete environments.