Proceedings of Machine Learning Research vol ???:1–80, 2025 38th Annual Conference on Learning Theory

Logarithmic regret of exploration in average reward Markov decision processes

Victor Boone VICTOR.BOONE@INRIA.FR
Bruno Gaujal BRUNO.GAUJAL@INRIA.FR
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

Editors: Nika Haghtalab and Ankur Moitra

Abstract

In average reward Markov decision processes, state-of-the-art algorithms for regret minimization follow a well-established framework: They are model-based, optimistic and episodic. First, they maintain a confidence region from which optimistic policies are computed using a well-known subroutine called Extended Value Iteration (`EVI`). Second, these policies are used over time windows called episodes, each ended by the Doubling Trick (`DT`) rule or a variant thereof. In this work, without modifying `EVI`, we show that there is a significant advantage in replacing (`DT`) by another simple rule, that we call the Vanishing Multiplicative (`VM`) rule. When managing episodes with (`VM`), the algorithm's regret is, both in theory and in practice, as good if not better than with (`DT`), while the one-shot behavior is greatly improved. More specifically, the management of bad episodes (when sub-optimal policies are being used) is much better under (`VM`) than (`DT`) by making the regret of exploration logarithmic rather than linear. These results are made possible by a new in-depth understanding of the contrasting behaviors of confidence regions during good and bad episodes.

Keywords: Markov decision processes, average reward, regret minimization, optimism

1. Introduction

Regret minimization in average reward Markov decision processes is a classical problem with a rich literature and landscape of methods. Regarding theoretical guarantees (especially in the minimax setting), the most successful line of algorithms adapts the famous `UCB` algorithm of Auer (2002) to Markov decision processes. This includes Auer and Ortner (2006); Tewari and Bartlett (2007); Auer et al. (2009); Bartlett and Tewari (2009); Filippi et al. (2010); Fruit et al. (2018); Tossou et al. (2019); Fruit et al. (2020); Bourel et al. (2020); Zhang and Ji (2019); Boone and Zhang (2024) in particular, that are the focus of this work. All these algorithms are episodic and follow the *optimism-in-the-face-of-uncertainty* principle: During learning, they maintain a confidence region of plausible environments from which decisions are taken. Specifically, they deploy policies achieving the highest average gain among all MDPs in the confidence region. This policy is used for a whole time interval called an episode, and is only updated when deemed necessary.

This paper is not about improving the regret guarantees of these algorithms. Instead, we are interested in improving their long term behavior over a single run. In particular, we argue that state-of-the-art algorithms renew their policy too lazily, leading to long sequences of sub-optimal play, even when the learning process is well advanced. This phenomenon appears strikingly during experiments: When running the classical `KLUCRL` of Filippi et al. (2010), the algorithm

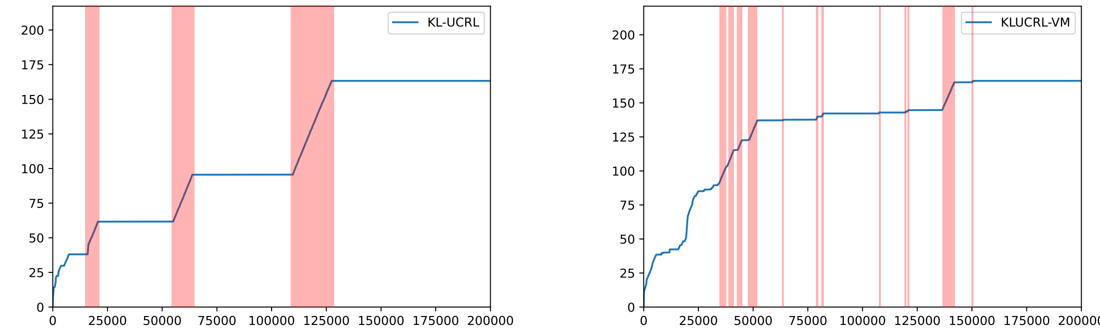


Figure 1: The left plot displays the regret of `KLUCRL` Filippi et al. (2010) over a single run with highlighted periods of sub-optimal play that are increasing in duration. In comparison, the right plot displays the regret of our proposed algorithm, where periods of sub-optimal play are much shorter resulting in a smoother regret curve.

displays periods of sub-optimal play that last for increasingly long durations, even after the initial burn-in phase is ended, see Figure 1. Such episodes of sub-optimal play are generally inevitable and correspond to the explorative part of the learning task; The learner has to make sure that seemingly bad actions are bad indeed. The issue rather lies in the fact that the current design of all these algorithms makes such episodes increase exponentially in size. This phenomenon was recently pointed out by Boone and Gaujal (2023) and measured by a new performance metric called the **regret of exploration** (see Definition 3). The authors further suggest a way to obtain regret of exploration guarantees by refining the management of episodes. However, their solution is computationally heavy and is only shown to work in the very restricted setting of Markov decision processes with deterministic transition kernels.

Contribution In this paper, we go beyond Boone and Gaujal (2023) and provide a solution with better guarantees, both theoretically and experimentally. We introduce a new simpler rule to end episodes, and show that the performance under the new episode rule guarantees logarithmic regret of exploration for two classes of MDPs: ergodic, and communicating MDPs with prior information on the support of the transition kernel. Our analysis is generic and focuses on when and how the confidence region used by an optimistic algorithm is well-behaved so that episodes of sub-optimal play are short and isolated. We further show that the regret guarantees remain mostly intact, both in the model independent (minimax) and model dependent settings.

2. Preliminaries

General notations Given a finite set $\mathcal{X}$, we denote $\mathcal{P}(\mathcal{X})$ the set of probability measures over $\mathcal{X}$. For $q \in \mathcal{P}(\mathcal{X})$ and $f : \mathcal{X} \to \mathbf{R}$ a measurable map, we write $qf := \int f(x)dq(x)$ the average of f against q. The Kullback-Leibler divergence from distribution q' to q' is denoted $\mathrm{KL}(q||q')$ and we further write $\mathrm{kl}(p, p') := \mathrm{KL}(\mathrm{Ber}(p)||\mathrm{Ber}(p')) = p\log(\frac{p}{p'}) + (1-p)\log(\frac{1-p}{1-p'})$ the divergence from Bernoulli distribution of parameters p' to p. Given a finite set $\mathcal{S}$, we denote

$e = (1, \ldots, 1) \in \mathbf{R}^{\mathcal{S}}$ the constant unitary vector and $(e_s)_{s \in \mathcal{S}}$ the canonical basis of $\mathbf{R}^{\mathcal{S}}$. The **span semi-norm** of a vector $u \in \mathbf{R}^{\mathcal{S}}$ is $\mathrm{sp}(u) := \max(u) - \min(u)$.

2.1. Markov decision processes in average reward

This work uses standard notations for Markov decision processes in average reward in the style of (Puterman, 1994, §8-9). A **Markov decision process** (or **model**) consists in a tuple $M \equiv (\mathcal{S}, \mathcal{A}, p, r)$ made of a state space $\mathcal{S}$ and an action space $\mathcal{A} \equiv \bigcup_{s \in \mathcal{S}} \mathcal{A}(s)$ together forming a state-action pair space $\mathcal{Z} := \bigcup_{s \in \mathcal{S}} \{s\} \times \mathcal{A}(s)$, a transition kernel $p : \mathcal{Z} \to \mathcal{P}(\mathcal{S})$ and reward distributions $r : \mathcal{Z} \to \mathcal{P}(\mathbf{R})$. The reward-kernel pair is $q := (r, p) : \mathcal{Z} \to \mathcal{P}(\mathbf{R}) \times \mathcal{P}(\mathcal{S})$.

Assumption 1 *The pair space $\mathcal{Z}$ is known and finite and rewards are Bernoulli.*

Since rewards have Bernoulli distributions (Assumption 1), we will use a harmless abuse of notations and write $r(z) \in [0, 1]$ for the mean reward function at $z \in \mathcal{Z}$.

2.1.1. INTERACTING WITH A MARKOV DECISION PROCESS USING POLICIES

The set of stationary deterministic policies is $\Pi \equiv \mathcal{S} \to \mathcal{A}$. We denote S_t, A_t, R_t the random state, action and reward observed at time t, and $Z_t := (S_t, A_t)$ is the associated pair. By construction, $S_{t+1} \sim p(Z_t)$ and $R_t \sim \mathrm{Ber}(r(Z_t))$. The (observed) history of play is $O_t := (S_1, A_1, R_1, \ldots, S_t)$ and $\mathcal{O}$ is the space of all possible histories. Fixing the environment M, the policy $\pi \in \Pi$ and the initial state $s \in \mathcal{S}$ properly defines a probability space on the set of histories, or, said more loosely, determines the distribution of $(S_t, A_t, R_t)_{t \geq 1}$. We write $\mathbf{E}_s^{M,\pi}[-]$ and $\mathbf{P}_s^{M,\pi}(-)$ the associated expectation and probability operators.

The **visit count** of a pair $z \in \mathcal{Z}$ is written $N_T(z) := \sum_{t=1}^{T-1} \mathbf{1}(Z_t = z)$.

The **gain** and **bias** functions of a policy $\pi \in \Pi$ from the initial state $s \in \mathcal{S}$ are denoted $g^\pi(s; M)$ and $h^\pi(s; M)$ and given by the formulas

$$
\begin{aligned}
g^\pi(s; M) &:= \lim_{T \to \infty} \mathbf{E}_s^{M,\pi}\left[\frac{1}{T}\sum_{t=1}^{T} R_t\right], \\
h^\pi(s; M) &:= \underset{T \to \infty}{\text{Cesaro-lim}}\, \mathbf{E}_s^{M,\pi}\left[\sum_{t=1}^{T} (R_t - g^\pi(S_t; M))\right].
\end{aligned}
$$

The optimal gain and bias functions are respectively $g^*(M)$ and $h^*(M)$ and the set of gain optimal policies, i.e., policies $\pi \in \Pi$ such that $g^\pi(M) = g^*(M)$, are denoted $\Pi^*(M)$. In particular, $g^*(M) := \max_{\pi \in \Pi} g^\pi(M)$ and $h^*(M) := \max_{\pi \in \Pi^*(M)} h^\pi(M)$. We define the **Bellman gaps** $\Delta^*(-; M) : \mathcal{Z} \to \mathbf{R}$ as the gaps in Bellman's optimality equations:

$$
\Delta^*(s, a; M) := g^*(s; M) + h^*(s; M) - r(s, a) - p(s, a) h^*(M). \tag{1}
$$

A pair $z \in \mathcal{Z}$ is said **weakly-optimal**, written $z \in \mathcal{Z}^*(M)$, if it has Bellman gap $\Delta^*(z; M) = 0$; and **sub-optimal** otherwise, written $z \in \mathcal{Z}^-(M)$. A pair $z \in \mathcal{Z}$ is said **optimal**, written $z \in \mathcal{Z}^{**}(M)$, if $z \in \mathcal{Z}^*(M)$ and it is visited infinitely often (almost surely) under some gain optimal policy. Note that $\mathcal{Z}^{**}(M) \subseteq \mathcal{Z}^*(M)$ by definition.

2.1.2. COMMUNICATING MARKOV DECISION PROCESSES

All throughout the paper, the models that we consider are always **communicating** (Assumption 2).

Assumption 2 *In this work, all Markov decision processes are communicating, i.e., that every state is reachable from any other under the right policy, meaning that the diameter is finite:*

$$D(M) := \max_{s \neq s'} \min_{\pi \in \Pi} \mathbf{E}_s^{M,\pi}[\inf\{t \geq 1 : S_t = s'\}] < \infty. \tag{2}$$

Assumption 2 is pretty common nowadays. It is the core assumption made in the seminal paper of Auer et al. (2009) and most subsequent works; It is much more general than the ergodic assumption of Agrawal (1990); Burnetas and Katehakis (1997); Pesquerel and Maillard (2022); It is not completely general either. This assumption is absolutely necessary for the well-behavior of the `EVI`-subroutine of Auer et al. (2009) (discussed downstream), which is the common foundation of the algorithms of interest in this paper.

Under Assumption 2, the optimal gain is a constant vector with $g^*(s; M) \in \mathbf{R}e$ so that we write $g^*(M) \in \mathbf{R}$ in place of $g^*(s; M)$; And Bellman gaps are non-negative, i.e., $\Delta^*(z; M) \geq 0$ for all $z \in \mathcal{Z}$. In the sequel, the dependency in M is dropped when unambiguous.

2.2. Reinforcement learning and regret minimization

A **learning algorithm** is formally a measurable map $\Lambda : \mathcal{O} \to \mathcal{P}(\mathcal{A})$, mapping histories of observations to probabilistic choices of actions. Similarly to policies, fixing the environment M, a learner Λ and the initial state $s \in \mathcal{S}$ properly defines the distribution of $(S_t, A_t, R_t)_{t \geq 1}$ and we write $\mathbf{E}_s^{M,\Lambda}[-]$ and $\mathbf{P}_s^{M,\Lambda}(-)$ the associated expectation and probability operators. The objective of the learner is to maximize $R_1 + \ldots + R_T$, and their ability to do so is measured by the **regret**, that compares the amount of reward that a gain optimal policy $\pi^* \in \Pi^*(M)$ (dependent on M) and the learner are able to collect within the same time budget. Following standard MDP theory, $|\mathbf{E}_s^{M,\pi^*}[R_1 + \ldots + R_T] - Tg^*| \leq \mathrm{sp}(h^*)$ so that in this setting, the regret is usually defined as $Tg^* - \sum_{t=1}^{T} R_t$, see Auer et al. (2009). In this work, we consider a pseudo-regret instead (Definition 1), to remove random noise over which the learner as no control. The expected regret defined below is equal to the classical one, up to an inconsequential additive constant, with $|\mathrm{Reg}(T; M, \Lambda, s) - \mathbf{E}_s^{M,\Lambda}[Tg^* - \sum_{t=1}^{T} R_t]| \leq \mathrm{sp}(h^*)$.

Definition 1 *The* ***pseudo-regret*** *of an algorithm Λ over M is the random variable given by:*

$$\Delta(1, T) \equiv \Delta(T) := \sum_{t=1}^{T} \Delta^*(Z_t; M) \tag{3}$$

and the ***expected regret*** *is* $\mathrm{Reg}(T; M, \Lambda, s) := \mathbf{E}_s^{M,\Lambda}[\Delta(T)]$.

The lower the regret, the better the learner performs. A learner Λ is said **no-regret** relatively to a set of models $\mathcal{M}^0$, or Hannan consistent Hannan (1957), if for all communicating $M \in \mathcal{M}^0$ and regardless of the initial state $s \in \mathcal{S}$, $\mathrm{Reg}(T; M, \Lambda, s) = \mathrm{o}(T)$. $\mathcal{M}^0$ will be called the **ambient set** and is a form of prior information. We further assume that $\mathcal{M}^0$ is in product form.

Assumption 3 *The ambient set $\mathcal{M}^0$ is in* ***product form****, i.e., it is of the form $\mathcal{M}^0 \equiv \prod_{z \in \mathcal{Z}} (\mathcal{R}_z^0 \times \mathcal{P}_z^0)$ where $\mathcal{R}_z^0 \subseteq [0, 1]$ and $\mathcal{P}_z^0 \subseteq \mathcal{P}(\mathcal{S})$.*

2.3. Optimistic model-based and `EVI`-based algorithms

There is a large literature on algorithms with regret guarantees. In this paper, we focus on **optimistic model-based algorithms**, which is a line of algorithms adapted from the well-known `UCB` Auer (2002). They follow the *optimism-in-the-face-of-uncertainty* (OFU) principle: when unsure about the value of an action or a policy, estimate that value as the highest that is statistically plausible. Ever since `UCRL` Auer and Ortner (2006), the main incarnation of this principle is the following. Over time, maintain a confidence region $\mathcal{M}(t)$ that contains M with high probability and work in an episodic fashion. An **episode** is a time segment $\{t_k, \dots, t_{k+1} - 1\}$ during which the algorithm plays a fixed policy $\pi_{t_k} \in \Pi$. This policy is computed as the policy achieving the highest gain on the best plausible model at time t_k. More formally, we define the **optimistic gain** of π in $\mathcal{M}(t)$ from $s \in \mathcal{S}$ as

$$g^\pi(s; \mathcal{M}(t)) := \sup_{M' \in \mathcal{M}(t)} g^\pi(s; M'). \tag{4}$$

An **optimistic policy** π is such that $g^\pi(\mathcal{M}(t)) := \sup_{\pi \in \Pi} g^\pi(\mathcal{M}(t))$. Perhaps surprisingly, optimistic policies are easy to compute from $\mathcal{M}(t)$ via a process called Extended Value Iteration (`EVI`), see Auer et al. (2009). Algorithm 1 provides the general architecture of these algorithms.

Algorithm 1 `EVI`-based algorithms.

for $t = 1, 2, \dots$ **do**
  if the current policy π_{t_k} is obsolete **then**
    Increase k, set $t_k \leftarrow t$ and compute $\pi_{t_k} \leftarrow \texttt{EVI}(\mathcal{M}(t_k))$;
  end if
  Set $\pi_t \leftarrow \pi_{t_k}$ and play $A_t := \pi_t(S_t)$;
end for

Selected focus: `KLUCRL` The scheme of `EVI`-based algorithms (Algorithm 1) can be improved along two axis. The first is the choice of confidence region. According to Sanov's theorem, the tightest way to construct $\mathcal{M}(t)$ so that it contains M with high probability is to rely on KL divergences (see Section A.1), leading to a region $\mathcal{M}(t) \equiv \prod_{z \in \mathcal{Z}} \mathcal{R}_z(t) \times \mathcal{P}_z(t)$ with:

$$\begin{aligned} \mathcal{R}_z(t) &:= \{\tilde{r}_z \in [0,1] : N_z(t) \mathrm{KL}(\hat{r}_z(t) || \tilde{r}_z) \le \log(2et)\} \cap \mathcal{R}^0_z \\ \mathcal{P}_z(t) &:= \{\tilde{p}_z \in \mathcal{P}(\mathcal{S}) : N_z(t) \mathrm{KL}(\hat{p}_z(t) || \tilde{p}_z) \le |\mathcal{S}| \log(2et)\} \cap \mathcal{P}^0_z \end{aligned} \tag{5}$$

where $\hat{r}_z(t)$ and $\hat{p}_z(t)$ are the empirically observed reward and kernels after t learning steps. The expression of (5) is tuned so that $\mathbf{P}(\exists t \ge T : M \notin \mathcal{M}(t)) \le 2|\mathcal{Z}|T^{-1}$, see Section B.2. Our work could be adapted to other types of confidence region, e.g., ℓ_1 or ℓ_2 norms, or Bernstein's style inequalities but the above will be the selected focus for its superiority over the others, both theoretically and empirically. Note that in (5), $\mathcal{M}(t)$ is constrained to the ambiant set of MDPs $\mathcal{M}^0$. This is a form of prior information.

The second potential improvement axis is the way to determine whether the current policy is obsolete or not. Most of the literature relies on the **doubling trick** (`DT`) or variants thereof, that essentially wait for a pair to increase its visit count multiplicatively—by 2 for the doubling trick.

$$N_t(S_t, \pi_{t_k}(S_t)) \ge \max\{2N_{t_k}(S_t, \pi_{t_k}(S_t)), 1\}. \tag{DT}$$

Choosing $\mathcal{M}(t)$ as in (5) and managing episodes with (`DT`) leads to our variant of the algorithm `KLUCRL` of Filippi et al. (2010) that can take the prior information $\mathcal{M}^0$ into account. From now on and to streamline the discussion, `KLUCRL` is the main focus.

3. The regret of exploration of episodic algorithms

In this work, we move beyond regret minimization, by investigating additional regret guarantees localized in time. To that end, the regret notations of Definition 1 are overloaded as such: we denote $\Delta(\tau, \tau') \equiv \Delta(\tau, \tau'; M) := \sum_{t=\tau}^{\tau'} \Delta^*(Z_t)$ the pseudo-regret endured from τ to τ' where $\tau \leq \tau'$ are two stopping times of the stochastic process. We further write $\mathrm{Reg}(\tau, \tau'; M) = \mathbf{E}_s^{M,\Lambda}[\Delta(\tau, \tau'; M)]$ the associated expected regret.

3.1. The definition of the regret of exploration beyond ergodic Markov decision processes

We start by generalizing the definition of the regret of exploration of Boone and Gaujal (2023) beyond ergodic models. To measure the instantaneous performance of an algorithm that has exploration phases, one may be tempted to monitor the regret starting at times when the algorithm drops an optimal policy for a sub-optimal one, i.e., at times

$$\{t_k : \pi_{t_k - 1} \in \Pi^*(M) \text{ and } \pi_{t_k} \notin \Pi^*(M)\}. \tag{6}$$

Equation (6) captures the idea. However, beyond ergodic environments, (6) is not precise enough and is ill-behaved in general. The main reason is that deployed policies can be partially optimal and multi-chain. For instance, (6) fails to capture time-instants where $\pi_{t_k - 1}$ is gain optimal from the current state but not globally, while such times should be considered as exploration times too. Indeed, the behavior of the algorithm does not depend on the actions chosen by the policy from states that can never be reached. To account for such cases, the final definition of **exploration times** (Definition 2) is slightly more complex.

Definition 2 (Exploration) *An episode k is an **exploration episode** and t_k is an **exploration time** if the two following conditions are satisfied:* (1) $g^*(M) = g^{\pi_{t_k - 1}}(S_{t_k}; M)$*; and* (2) *we have* $\mathbf{P}_s^{\pi_{t_k}}(\exists t \geq 1 : \Delta^*(Z_t; M) \neq 0) > 0$*. The set of exploration episodes is denoted* $\mathcal{K}^-$.

Written differently, t_k is an exploration time if the learning agent drops a policy that is gain optimal from the current state for a policy that may use a sub-optimal pair if iterated over and over from the current state. When the underlying model is ergodic, the exploration times given by Definition 2 are equivalent to those defined using (6) and by Boone and Gaujal (2023).

We enumerate $\mathcal{K}^-$ as $(t_{k(i)})$ where $k(i)$ denotes the i-th exploration episode and $t_{k(i)}$ is the associated i-th initial exploration time. Formally, $t_{k(1)} := \inf \mathcal{K}^-$ and $t_{k(i+1)} := \inf\{t_k > t_{k(i)} : k \in \mathcal{K}^-\}$. The **regret of exploration** is defined as the worst expected regret at exploration times asymptotically.

Definition 3 (Regret of exploration) *Let $(t_{k(i)})$ be the enumeration of exploration times. The **regret of exploration** is given by:*

$$\mathrm{RegExp}(T) \equiv \mathrm{RegExp}(T; M) := \limsup_{i \to \infty} \mathrm{Reg}(t_{k(i)}, t_{k(i)} + T; M). \tag{7}$$

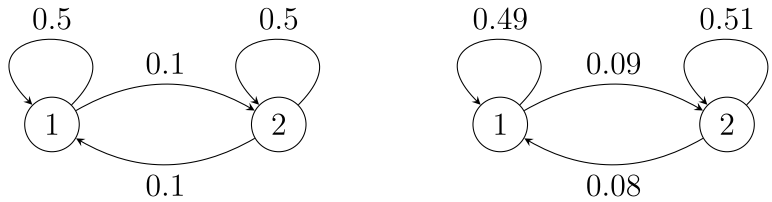


Figure 2: Examples of non-degeneracy (Definition 4). A degenerate Markov decision process (to the left) and a non-degenerate Markov decision process (to the right). Both models have deterministic transitions represented with arrows. Labels are reward means.

3.2. Explorative Markov decision processes

The regret of exploration is only worth studying if there are infinitely many exploration times $t_{k(i)}$. This is not always the case. In fact, there exist Markov decision processes for which infinite exploration is somehow unnecessary, making them conceptually easier to learn than bandits. In Appendix G, we provide a complete characterization of the set of Markov decision processes for which the number of exploration episodes is infinite and where the regret of exploration is well-defined: For such models, there exist **consistent** (see Salomon et al. (2013)) learners Λ achieving $\mathrm{Reg}(T; M, \Lambda) = \mathrm{o}(\log(T))$. This result is surprisingly difficult to establish and is peripheral to our work, hence completely deferred to Appendix G. Let us insist on the fact that given a class $\mathcal{M}$ of environments, the regret of exploration may be ill-defined for large sub-spaces of $\mathcal{M}$, even for reasonable learning algorithms. This motivates the following definitions.

Definition 4 (Non-degeneracy) *A model $M \equiv (\mathcal{Z}, r, p)$ is said* ***non-degenerate*** *if* (1) *it has a unique Bellman optimal policy, i.e., there is a unique $\pi \in \Pi$ that satisfies the first two orders optimality equations:*

$$\forall s \in \mathcal{S}, \quad g^\pi(s) = \max_{a \in \mathcal{A}(s)} \{p(s,a)g^\pi\}$$
$$\forall s \in \mathcal{S}, \quad g^\pi(s) + h^\pi(s) = \max_{a \in \mathcal{A}(s)} \{r(s,a) + p(s,a)h^\pi\}$$

and (2) *that unique Bellman optimal policy is unichain.*

In Figure 2, we provide an example of degenerate and non-degenerate Markov decision processes. On the left model, there are multiple Bellman optimal policies, as one can achieve optimal gain by either looping on the left or the right loop. As a matter of fact, every policy excepted $1 \leftrightarrow 2$ is Bellman optimal. On the right model, we add a small noise to the reward function. That noise breaks ties and the right loop becomes better than the other, so that the Bellman optimal policy is indeed unique ($1 \to 2 \to 2$) and unichain.

The main observation from Figure 2 can be generalized: All "noisy" versions of the non-degenerate model (to the left) are non-degenerate. This means that "almost-all" communicating Markov decision processes are non-degenerate and makes the non-degeneracy assumption mild. More details are found in Appendix F.

From now on, we will focus on non-degenerate Markov decision processes.

Definition 5 (Explorative models) *Given a space of Markov decision processes $\mathcal{M}$, its **explorative sub-space** $\mathcal{M}^+$ is the set of non-degenerate models $M \in \mathcal{M}$ such that every algorithm* (1) *with sub-linearly many episodes and* (2) *which is no-regret on $\mathcal{M}$, has infinitely many exploration episodes almost surely. Non-explorative models are said **exploration-free**.*

Note that the explorative character of a Markov decision process depends on the ambient space. This makes explorative models slightly more difficult to describe than non-degenerate ones. That discussion is deferred to Appendix G.

3.3. The doubling trick leads to linear regret of exploration

In Figure 1, we observe that the regret at exploration times of `KLUCRL` that uses (`DT`) increases overall. This follows from a general principle that is quite intuitive: If a change of episode requires an increase of visits relatively to the initial visit count vector, and if deployed policies do not play actions with vanishing probabilities (see (8)), then the regret of exploration grows linearly on recurrent models at least. The following theorem is an alternative version of (Boone and Gaujal, 2023, Theorem 1), adapted to our definition of exploration times.

Theorem 6 *Fix a pair space $\mathcal{Z}$ and let $\mathcal{M}$ be the space of all recurrent models with pairs $\mathcal{Z}$. Let $f : \mathbf{N} \to (0, \infty)$ be such that $\lim f(n) = +\infty$. Any no-regret episodic learner (π_t), i.e., using fixed policies over episodes $\{t_k, \dots, t_{k+1} - 1\}$, satisfying*

$$\begin{aligned} \forall k \geq 1, \exists z \in \mathcal{Z}, \quad & N_z(t_{k+1}) \geq N_z(t_k) + f(N_z(t_k)) \\ \exists c > 0, \forall t \geq 0, \forall (s,a) \in \mathcal{Z}, \quad & \pi_t(a|z) \geq c \text{ or } \pi_t(a|z) = 0 \end{aligned} \tag{8}$$

has linear regret of exploration on the explorative sub-space of $\mathcal{M}$, i.e., for all $M \in \mathcal{M}^+$, we have $\mathrm{RegExp}(T) = \Omega(T)$ a.s. when $T \to \infty$.

This result applies to `KLUCRL` and more generally to all algorithms relying on the doubling trick (`DT`) to manage episodes, corresponding to $f(n) = n \vee 1$. This includes `UCRL2` Auer et al. (2009), `REGAL` Bartlett and Tewari (2009), `KLUCRL` Filippi et al. (2010), `UCRL2B` Fruit et al. (2020), `SCAL` Fruit et al. (2018), `UCRL3` Bourel et al. (2020), `EBF` Zhang and Ji (2019) and also `PMEVI` Boone and Zhang (2024) (up to mild modifications of (8) for a few of them). Therefore, Theorem 6 pinpoints an issue: The local regret of current optimistic algorithms is the worst possible, because the regret of exploration of these methods grows linearly. In this paper, we will fix this problem without too many side-effects. We alter these algorithms (focusing on `KLUCRL`) and achieve sub-linear regret of exploration without hurting the minimax regret guarantees. This is achieved by a small and cost-less modification of the episode stopping rule.

4. Logarithmic regret of exploration with the vanishing multiplicative condition

Our solution improves on Boone and Gaujal (2023), where the regret of exploration guarantees are only proved for deterministic transition MDPs. Their solution consists in stopping an episode if the current policy is no longer optimistically optimal *enough*. This is done by introducing a function $\psi(t)$ and ending episode k if $g^{\pi_{t_k}}(\mathcal{M}(t)) + \psi(t) < g^*(\mathcal{M}(t))$. This approach has the clear issue that one has to constantly monitor the values of $g^{\pi_{t_k}}(\mathcal{M}(t))$ and $g^*(\mathcal{M}(t))$. This has a high computational cost. Despite these limitations, the main observation of Boone and Gaujal (2023) is key: if π_{t_k} is sub-optimal, its optimistic gain $g^{\pi_{t_k}}(\mathcal{M}(t))$ decreases quickly over $\{t_k, \dots, t\}$ and otherwise, if π_{t_k} is optimal, then $g^{\pi_{t_k}}(\mathcal{M}(t))$ rather behaves like a random walk. Essentializing their argument, it actually boils down to show that the behaviors of confidence regions $\mathcal{R}_z(t)$ and $\mathcal{P}_z(t)$ are very different at high and low visit counts of z. At high visit counts, the evolution of confidence regions is slow and they mostly behave like random walks. At low visit counts, they *shrink* quickly as the number of visits increases. We call these two distinct behaviors the **shrinking-shaking effect**. It motivates a different and simpler approach than the one of Boone and Gaujal (2023): Renew the episode when there is an increase of information. This is actually the idea of the (DT), but instead of asking for a visit count to double, we suggest to wait for a multiplicative increase with respect to a vanishing time-dependent factor, i.e.,

$$N_t(S_t, \pi_{t_k}(S_t)) > (1 + f(t_k)) \max\{1, N_{t_k}(S_t, \pi_{t_k}(S_t))\} \tag{VM}$$

where $f : \mathbf{N} \to [0, 1]$ is a non-increasing vanishing function of t. The above condition will be referred to as the f-**Vanishing Multiplicative condition**, or f-(VM), or even more simply (VM). Remark that (DT) is also of the form (VM) with $f \equiv 1$, except that this function is not vanishing.

By changing (DT) for (VM), we get the following range of regret guarantees for `KLUCRL`.

Theorem 7 (Main result) *Let $f : \mathbf{N} \to [0, 1]$ and consider running `KLUCRL` with episodes managed by f-(VM) with prior information $\mathcal{M}^0$ and let $\mathcal{M}_D$ be the set of Markov decision processes with diameter less than D. We have:*

1. Minimax*: For $f = \Omega(t^{-1/2})$, $\sup_{M' \in \mathcal{M}_D} \mathrm{Reg}(T; M') = \mathrm{O}(DS\sqrt{AT \log(T)})$;*

Moreover, if M satisfies Assumption 4 and is explorative, we have:

2. Model dependent*: If $f > 0$, then $\mathrm{Reg}(T; M) = \mathrm{O}(\log(T) \log\log(T))$;*
3. Regret of exploration*: If $f(t) = \mathrm{o}\left(\frac{1}{\log(t)}\right)$, then $\mathrm{RegExp}(T; M) = \mathrm{O}(\log(T))$.*

Note that the model dependent regret guarantees and the regret of exploration guarantees only hold if M satisfies a structural assumption with respect to the ambient set of models, $\mathcal{M}^0$.

Assumption 4 (Interior assumption) *For $z \in \mathcal{Z}$, $r(z)$ and $p(z)$ are in the interior of $\mathcal{R}_z^0$ and $\mathcal{P}_z^0$ respectively, i.e., $\mathrm{supp}(r(z)) = \{0, 1\}$ and $\mathrm{supp}(p(z)) \supseteq \mathrm{supp}(p'(z))$ for all $p'(z) \in \mathcal{P}_z^0$.*

Comment 1 The minimax regret guarantees (*1.*) given in Theorem 7 are the same as for the original (DT) version of `KLUCRL`. The model dependent guarantees (*2.*) only suffer from an additional $\log\log(T)$ factor, while the regret of exploration guarantees are improved from $\Omega(T)$ to $\mathrm{O}(\log(T))$. Moreover, although Theorem 7 is stated specifically for `KLUCRL`, it can be generalized to other types of confidence regions such as ℓ_1, ℓ_2 or Bernstein-type., inducing similar results for other algorithms such as `UCRL`, `UCRL2`, `UCRL3`, `EBF` or `PMEVI`.

Comment 2 When $\mathcal{M}^0 = \prod_{z \in \mathcal{Z}}([0,1] \times \mathcal{P}(\mathcal{S}))$ (no prior information), M satisfies Assumption 4 if, and only if M is an ergodic model with fully-supported transition kernels. However, non-ergodic models can also be covered when prior information on the support of p is available.

Outline of the proof The minimax guarantees directly follow from a straight-forward upper bound of the number of episodes under f-(VM) and are detailed in Appendix B. The model dependent guarantees are proved in Appendix E. In the remaining of the paper, we focus on the analysis of the regret of exploration. The proof is challenging and requires an in-depth understanding of the behavior of optimistic algorithms in the long run.

First, we establish a noteworthy difference between the visit rates of optimal and non-optimal pairs in Section 4.1: the first are visited linearly and the others at most logarithmically. Following this observation, we explain in Section 4.2 how these distinct rates imply drastically different behaviors of the associated confidence regions, referred to as the **shrinking-shaking** dichotomy. In turn, it leads to a general conceptual property that we call **coherence** in Section 4.3. Lastly, we show that regret of exploration is logarithmic because of coherence, see Section 4.4.

4.1. Visit rates of optimal and non-optimal pairs

The result below describes the almost-sure asymptotic regime of versions of `KLUCRL` managing episodes with (VM). Up to the non-degeneracy of the underlying model, the visit counts can be split into two regimes: $N_z(t)$ grows linearly with t for $z \in \mathcal{Z}^{**}(M)$ while $N_z(t)$ grows sub-logarithmically for $z \notin \mathcal{Z}^{**}(M)$ (including $\mathcal{Z}^{**}(M) \setminus \mathcal{Z}^{*}(M)$ in particular).

Lemma 8 (Almost-sure asymptotic regime) *Let $M \in \mathcal{M}$ be a non-degenerate model satisfying Assumption 4. Assume that `KLUCRL` is run while managing episodes with f-(VM) with arbitrary $f > 0$. There exists $\lambda > 0$ such that:*

$$\begin{aligned} &\forall z \notin \mathcal{Z}^{**}(M), \quad \mathbf{P}^M(\exists T, \forall t \geq T : N_z(t) < \lambda \log(t)) = 1, \text{ and} \\ &\forall z \in \mathcal{Z}^{**}(M), \quad \mathbf{P}^M\left(\exists T, \forall t \geq T : N_z(t) > \tfrac{1}{\lambda} t\right) = 1. \end{aligned}$$

The result holds in particular for (DT). This is not much of a surprise, since `KLUCRL` is known to have logarithmic model dependent regret. In contrast, this result is remarkable for (VM) because the number of episodes can arbitrarily greater than logarithmic.

Note that (DT) and (VM) differ in the amount of time a sub-sampled pair can be visited during an episode. Indeed, for $z \in \mathcal{Z}$ with $\mathbf{P}^M(\exists T, \forall t \geq T : N_z(t) < \lambda \log(t)) = 1$, we have

$$N_z(t_{k+1}) \leq \lfloor (1 + f(t_k)) N_z(t_k) \rfloor + 1 = N_z(t_k) + 1 + \lfloor \lambda f(t_k) \log(t_k) \rfloor. \tag{9}$$

For $f(t) = \mathrm{o}\left(\frac{1}{\log(t)}\right)$, we have $\lfloor \lambda f(t_k) \log(t_k) \rfloor = 0$ provided that t_k is large enough. So, following (9), sub-sampled pairs are visited at most once per episode in the long run. So, under (VM), `KLUCRL` almost instantly refreshes its policy when sub-optimal pairs are visited.

4.2. The shrinking-shaking dichotomy in the behavior of confidence regions

The shrinking-shaking effect concerns the way confidence regions evolve over time under low (shrinking) and high (shaking) amounts of information. This is illustrated in Figure 3.

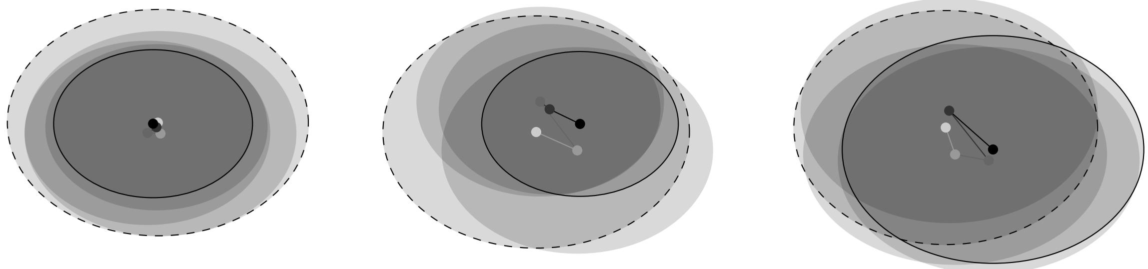


Figure 3: An artist view of the shrinking/shaking behavior of the $\mathcal{Q}_{s,a}(t)$ as the number of new samples $N_{s,a}(t') - N_{s,a}(t) \ll N_{s,a}(t)$ increases (from dashed to solid line).

Informal Property 9 (Shrinking-Shaking effect) *Let $(t_{k(i)})$ be the enumeration of exploration episodes. Fix $T \geq 1$ and denote $\mathcal{Q}_z(t) := \mathcal{R}_z(t) \times \mathcal{P}_z(t)$. With high probability and uniformly over $t \in \{t_{k(i)}, \ldots, t_{k(i)} + T\}$, we have*

$$N_z(t) > N_z(t_{k(i)}) + \mathbf{1}(z \notin \mathcal{Z}^{**}(M)) C \log(T) \implies \mathcal{Q}_z(t) \subseteq \mathcal{Q}_z(t_{k(i)-1}).$$

In Informal Property 9, which is informal and slightly wrong, $\mathbf{1}(z \notin \mathcal{Z}^{**}(M))$ incarnates the shrinking-shaking dichotomy. The rigorous treatment of this dichotomy is tedious and calculatory, while the idea behind the phenomenon is quite intuitive. We postpone the formal, precise and extensive description of the shrinking-shaking effect to Appendix D.

Instead, we provide here a heuristic derivation of what the shrinking-shaking phenomenon is and how it is proved. As a matter of fact, the phenomenon already appears in the simple setting of bandits with Gaussian rewards. Let us introduce specialized notations for that purpose. Let (X_i) be a sequence of i.i.d. random variables of distribution $\mathrm{N}(\mu, \sigma^2)$ and let $\widehat{\mu}(n) := \frac{1}{n} \sum_{k=1}^{n} X_k$ the empirical average after n samples. The typical way to construct a confidence region for μ follows from Azuma-Hoeffding's inequality, with $\mathcal{I}(n) := \{\widetilde{\mu} \in \mathbf{R} : |\widetilde{\mu} - \widehat{\mu}(n)| \leq \sigma\sqrt{\log(1/\delta)/n}\}$. The supremum of $\mathcal{I}(n)$ is the largest plausible value for μ, and is given by:

$$\widetilde{\mu}(n) := \widehat{\mu}(n) + \sigma\sqrt{\frac{\log(1/\delta)}{n}} \equiv \widehat{\mu}(n) + \sigma\sqrt{\frac{\log(t)}{n}}. \tag{10}$$

Regarding our setting, $\widetilde{\mu}(n)$ is the analogue of $\sup \mathcal{R}_z(t)$, i.e., the highest plausible reward for a given pair at a given time. The quantity $\log(1/\delta)$ can be seen as $\log(t)$ as our confidence regions are tuned for the confidence threshold at $\delta = \frac{1}{t}$. The **shrinking-shaking effect** is a general observation about the evolution of $\widetilde{\mu}(n + \mathrm{d}n)$ after a few samples $\mathrm{d}n$. It starts with the first order Taylor expansion of $\widetilde{\mu}(n + \mathrm{d}n)$, giving

$$\widetilde{\mu}(n + \mathrm{d}n) = \widetilde{\mu}(n) + \underbrace{\widehat{\mu}(n + \mathrm{d}n) - \widehat{\mu}(n)}_{\text{EMPIRICAL UPDATE}} - \underbrace{\frac{\sqrt{\log(t)}\mathrm{d}n}{2n\sqrt{n}}}_{\text{OPTIMISM DROP}}. \tag{11}$$

The update in the optimistic estimate is the sum of two quantities: the empirical update and the optimism drop. The empirical update is the change of $\widehat{\mu}(n)$ and is roughly noise. Indeed, from the law of large numbers, we have $\widehat{\mu}(n) \approx \mu$ and $\widehat{\mu}(n + \mathrm{d}n) - \widehat{\mu}(n) \approx \frac{1}{n} \sum_{k=n}^{n+\mathrm{d}n} (X_k - \mu)$, hence is the sum of $\mathrm{d}n$ i.i.d. centered random variables. By Azuma-Hoeffding's inequality, we find that $\widehat{\mu}(n + \mathrm{d}n) - \widehat{\mu}(n) \approx \frac{1}{n}\sigma\sqrt{\log(1/\delta')}\sqrt{\mathrm{d}n}$ with probability $1 - \delta'$.

Taking $n = \mathrm{O}(\log(t))$ leads to the **shrinking effect**: $\widetilde{\mu}(n + \mathrm{d}n)$ tends to decrease after a few additional samples $\mathrm{d}n$. Indeed, to have a decrease in the optimistic estimate in (11) from n to $n + \mathrm{d}n$, we need

$$\sigma\sqrt{\log\left(\frac{1}{\delta'}\right)}\sqrt{\mathrm{d}n} \leq \frac{1}{2}\sqrt{\frac{\log(t)}{n}}\mathrm{d}n. \tag{12}$$

When $n = \mathrm{O}(\log(t))$, (12) states that as soon as $\mathrm{d}n = \Omega(\sigma^2 \log(\frac{1}{\delta'}))$, the noise due to the empirical updates becomes negligible with respect to the optimism drop. Then, the optimistic estimate $\widetilde{\mu}(n + \mathrm{d}n)$ starts to decrease with quantifiable speed.

Taking $n = \Omega(t)$, we get the opposite; This is the **shaking effect**. More specifically, the optimism drop kills the noise of the empirical update. Indeed, by setting $n = t$ in (11), the optimism drop is of order $\frac{1}{n}(\frac{\log(t)}{t})^{1/2}\mathrm{d}n$. This quantity is eventually negligible in front of the noise, that is of order $\frac{1}{n}\sqrt{\mathrm{d}n}$ when $t \to \infty$. Hence the shaking effect.

What it means for `KLUCRL` When running `KLUCRL`, we have argued in Section 4.1 that optimal pairs satisfy $N_z(t) = \Omega(t)$ while non-optimal pairs satisfy $N_z(t) = \mathrm{O}(\log(t))$. When `KLUCRL` deploys a sub-optimal policy π, this policy uses non-optimal pairs, for which the confidence regions $\mathcal{R}_z(t)$ and $\mathcal{P}_z(t)$ tend to shrink while all the others are negligibly shaking. When iterating π, these non-optimal pairs are eventually visited enough, $\mathcal{R}_z(t)$ and $\mathcal{P}_z(t)$ eventually shrink, so the optimistic gain of π decreases. Hence, π won't be used as an optimistic policy anymore if the episode is updated—and it is updated quickly thanks to (`VM`), see (9).

4.3. A central conceptual property: coherence

The point of the shrinking-shaking effect is to establish a **coherence property** defined below.

Definition 10 (Coherence) *We say that an algorithm is* $(\boldsymbol{F}, \boldsymbol{\tau}, \boldsymbol{T}, \boldsymbol{\varphi})$***-coherent*** *if* $F \equiv (F_t : t \geq 1)$ *is an adapted sequence of events,* τ *a stopping time,* $T \geq 1$ *is a scalar and* $\varphi : \mathbf{N} \to [0, \infty)$ *is a function such that, for all* $t \in \{\tau, \ldots, \tau + T - 1\}$,

$$F_t \subseteq \left\{ g^{\pi_t}(S_t) < g^*(S_t) \Rightarrow \exists z \equiv (s, a) \in \mathrm{Reach}(\pi_t, S_t) : \begin{bmatrix} N_z(t) - N_z(\tau) \leq \varphi(\tau) \\ \textit{and } g^{\pi_t}(s) < g^*(s) \end{bmatrix} \right\}$$

where $z \equiv (s, a) \in \mathrm{Reach}(\pi_t, S_t)$ *stands for* $\pi(a|s) > 0$ *and* $\mathbf{P}^{\pi}_{S_t}(\tau_s < \infty) > 0$.

Roughly speaking, coherence states that the iteration of a sub-optimal policy is linked to a lack of information (quantified by a budget $\varphi(\tau)$) that has positive probability to be recovered by iterating that policy only. The purpose of the coherence property is its link with local regret guarantees, as shown by Lemma 12 below. However, the coherence property may only be conveniently used if the episodes of the algorithm are **weakly regenerative**, meaning that episodes may only end if the current state has already been visited during the episode. This property makes sure that the sub-sampled state-action pair, of which coherence ensures the existence, is reached and visited during the episode with positive probability.

Definition 11 (Weakly regenerative episodes) *We say that the episodes of an algorithm are **weakly regenerative** if, for all $k \geq 1$, there exists $t \in \{t_k, \ldots, t_{k+1} - 1\}$ such that $S_t = S_{t_{k+1}}$.*

Lemma 12 (Coherence and local regret) *Assume that M is non-degenerate (Definition 4). If the algorithm is (F, τ, T, φ)-coherent and has weakly regenerative episodes, then there exist model dependent constants $C_1, C_2, C_3, C_4 > 0$ such that:*

$$\forall x \geq 0, \quad \mathbf{P}\left(\Delta(\tau, \tau+T) \geq x + C_4\varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t\right) \leq C_1 T^{C_3} \exp\left(-\frac{x}{C_2}\right).$$

More specifically, C_1, C_2, C_3, C_4 only depend on M and are independent of F, τ, T and φ.

Using the shorthand $F_{\tau:\tau+T} := \bigcap_{t=\tau}^{\tau+T-1} F_t$, this means that on a good event $F_{\tau:\tau+T}$, the local regret $\Delta(\tau, \tau+T)$ has sub-exponential tails. The above result can also be written in the form $\mathbf{P}(\Delta(\tau, \tau+T) \geq C_1 + C_4\varphi(\tau) + (\eta C_2 + C_3)\log(T), F_{\tau:\tau+T}) \leq T^{-\eta}$ for all $\eta > 0$.

The proof of Lemma 12 is difficult and deferred to Appendix C.

4.4. Establishing regret of exploration guarantees via coherence

Based on the shrinking and shaking effects discussed upstream, we show that (VM) guarantees **local** coherence properties that, once combined with Lemma 8, become regret of exploration guarantees. The exact coherence property is detailed in Lemma 13 below. Once Lemma 13 is established (see Section D.4), Theorem 7 follows instantly.

Lemma 13 *Let $M \in \mathcal{M}^+$ be a non-degenerate explorative model. Consider running `KLUCRL` with model satisfying Assumption 4 and assume that episodes are managed with the f-(`VM`) with $f(t) = \mathrm{o}(\frac{1}{\log(t)})$. Let $(t_{k(i)})$ be the enumeration of exploration episodes. Then, there exists a constant $C(M) > 0$ such that, for all $T \geq 1$ and $\delta > 0$, there is an adapted sequence of events (E_t) and a function $\varphi : \mathbf{N} \to \mathbf{R}$ such that:*

1. *For all $i \geq 1$, the algorithm is $(E_t, t_{k(i)}, T, \varphi)$-coherent;*
2. $\mathbf{P}\big(\bigcup_{t=t_{k(i)}}^{t_{k(i)}+T-1} E_t^c\big) \leq \delta + \mathrm{o}(1)$ *when $i \to \infty$;*
3. $\varphi(t) \leq 1 + C\log(\frac{T}{\delta}) + \mathrm{o}(1)$ *when $t \to \infty$.*

Proof of Theorem 7, assertion 3 Use the coherence property of Lemma 13 with $\delta = \frac{1}{T}$ and apply Lemma 12. We obtain:

$$\begin{aligned}
&\limsup_{i\to\infty} \mathbf{P}\big(\Delta(t_{k(i)}, t_{k(i)}+T) \geq x + C_4\varphi(t_{k(i)})\big) \\
&\leq \limsup_{i\to\infty}\left\{\mathbf{P}\left(\Delta(t_{k(i)}, t_{k(i)}+T) \geq x + C_4\varphi(t_{k(i)}), \bigcap_{t=t_{k(i)}}^{t_{k(i)}+T-1} E_t\right) + \mathbf{P}\left(\bigcup_{t=t_{k(i)}}^{t_{k(i)}+T-1} E_t^c\right)\right\} \\
&\leq \exp\left(-\frac{x}{C_2} + C_3\log(T) + \log(C_1)\right) + \frac{1}{T}
\end{aligned}$$

which is bounded by $\frac{2}{T}$ for $x \geq C_2(1+C_3)\log(T) + C_2\log(C_1)$, where C_1, C_2, C_3, C_4 are the constants provided by Lemma 12. Using that $\limsup_{i\to\infty} \varphi(t_{k(i)}) \leq 1 + 2C\log(T)$ and setting $\psi(T) := (C_2(1+C_3) + 2C_4C)\log(T) + C_2\log(C_1) + C_4$, we obtain:

$$\mathrm{RegExp}(T) \leq \limsup_{i\to\infty} \big\{ \psi(T) + T \cdot \mathbf{P}\big(\mathrm{Reg}(t_{k(i)}, t_{k(i)} + T) \geq \psi(T)\big) \big\} \leq \psi(T) + 2. \quad (13)$$

This concludes the proof of Theorem 7. ■

5. Beyond asymptotic guarantees

Our theoretical results (Theorem 7) are only asymptotic. However, the behavior of `KLUCRL` with episodes managed by (`VM`) is remarkably better than its (`DT`) version over a single run and for reasonably small time horizons, as displayed in Figure 1. In Appendix A, we provide additional numerical insights with thorough evidence of the smoother behavior of (`VM`) over (`DT`) for `KLUCRL` as well as for other learning algorithms (`UCRL2`, `UCRL2B`, ...).

Acknowledgments

This research was partly supported by the PEPR IA FOUNDRY (ANR-23-PEIA-0003) of the French National Research Agency (ANR).

References

Rajeev Agrawal. Adaptive control of Markov chains under the weak accessibility. *29th IEEE Conference on Decision and Control*, pages 1426–1431 vol.3, 1990.

Peter Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *J. Mach. Learn. Res.*, 3:397–422, 2002.

Peter Auer and Ronald Ortner. Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning. *Proceedings of the 19th International Conference on Neural Information Processing Systems*, December 2006.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009.

Peter L. Bartlett and Ambuj Tewari. REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 35–42, Arlington, Virginia, USA, June 2009. AUAI Press. ISBN 978-0-9749039-5-8.

Victor Boone and Bruno Gaujal. The Regret of Exploration and the Control of Bad Episodes in Reinforcement Learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2824–2856. PMLR, July 2023.

Victor Boone and Odalric-Ambrym Maillard. The regret lower bound for communicating Markov Decision Processes, 2025. URL https://arxiv.org/abs/2501.13013. _eprint: 2501.13013.

Victor Boone and Zihan Zhang. Achieving Tractable Minimax Optimal Regret in Average Reward MDPs, 2024. URL https://arxiv.org/abs/2406.01234. _eprint: 2406.01234.

Hippolyte Bourel, Odalric Maillard, and Mohammad Sadegh Talebi. Tightening Exploration in Upper Confidence Reinforcement Learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1056–1066. PMLR, July 2020.

Apostolos Burnetas and Michael Katehakis. Optimal Adaptive Policies for Markov Decision Processes. *Mathematics of Operations Research - MOR*, 22:222–255, February 1997.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in Reinforcement Learning and Kullback-Leibler Divergence. *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122, September 2010. arXiv: 1004.5229.

Ronan Fruit. *Exploration-exploitation dilemma in Reinforcement Learning under various form of prior knowledge*. PhD Thesis, Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189, 2019.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning. *Proceedings of the 35 th International Conference on Machine Learning*, 2018.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved Analysis of UCRL2 with Empirical Bernstein Inequality. *ArXiv*, abs/2007.05456, 2020.

James Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3(2):97–139, 1957.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263, 2020.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, March 1985. ISSN 01968858.

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the Best Policy in Markov Decision Processes. *arXiv:2106.02847 [cs, stat]*, October 2021. arXiv: 2106.02847.

Ronald Ortner. Online regret bounds for Markov decision processes with deterministic transitions. *Theoretical Computer Science*, 411(29):2684–2695, 2010. ISSN 0304-3975.

Fabien Pesquerel and Odalric-Ambrym Maillard. IMED-RL: Regret optimal learning of ergodic Markov decision processes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26363–26374. Curran Associates, Inc., 2022.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1 edition, April 1994. ISBN 978-0-471-61977-2 978-0-470-31688-7.

Antoine Salomon, Jean-Yves Audibert, and Issam El Alaoui. Lower bounds and selectivity of weak-consistent policies in stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 14(1), 2013.

Paul J Schweitzer. On undiscounted Markovian decision processes with compact action spaces. *RAIRO-Operations Research*, 19(1):71–86, 1985. Publisher: EDP Sciences.

Paul J Schweitzer. A Brouwer fixed-point mapping approach to communicating Markov decision processes. *Journal of mathematical analysis and applications*, 123(1):117–130, 1987. Publisher: Elsevier.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-Aware Regret Bounds for Undiscounted Reinforcement Learning in MDPs. *Journal of Machine Learning Research*, pages 1–36, April 2018. Publisher: Microtome Publishing.

Ambuj Tewari and Peter Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. *Advances in Neural Information Processing Systems*, 20, 2007.

Aristide Tossou, Debabrota Basu, and Christos Dimitrakakis. Near-optimal Optimistic Reinforcement Learning using Empirical Bernstein Inequalities. *arXiv:1905.12425 [cs, stat]*, December 2019. arXiv: 1905.12425.

Damianos Tranos and Alexandre Proutiere. Regret Analysis in Deterministic Reinforcement Learning. *arXiv:2106.14338 [cs, stat]*, June 2021. arXiv: 2106.14338.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Zihan Zhang and Xiangyang Ji. Regret Minimization for Reinforcement Learning by Evaluating the Optimal Bias Function. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Appendix A. Experiments: The Vanishing Multiplicative condition in practice

In this appendix, we provide a few numerical insights in `KLUCRL` and the differences between (`DT`) and (`VM`). In Section A.1, we justify the choice of `KLUCRL` both from a theoretical and experimental viewpoint. In Section A.2, we show that in practice, the regret of `KLUCRL` is slightly better with (`VM`) than with (`DT`) in expectation, in distribution and sometimes in variance. In Section A.3, we provide an experimental proxy for the regret of exploration, that displays a clear advantage in using (`VM`) over (`DT`).

A.1. On the choice of KLUCRL as a reference algorithm

The confidence region $\mathcal{M}(t)$ is built in product form, i.e., $\mathcal{M}(t) \equiv \prod_{z \in \mathcal{Z}} (\mathcal{R}_z(t) \times \mathcal{P}_z(t))$ where $\mathcal{R}_z(t) \subseteq [0, 1]$ and $\mathcal{P}_z(t) \subseteq \mathcal{P}(\mathcal{S})$. Both are confidence regions for categorical distributions (of dimension $d = 2$ for rewards and $d = |\mathcal{S}|$ for kernels), in which case the constructions of $\mathcal{R}_z(t)$ and $\mathcal{P}_z(t)$ are traditionally relying on concentration inequalities. These relate how far is the empirical estimate ($\hat{r}_t(z)$ and $\hat{p}_t(z)$) from the true expected value ($r(z)$ and $p(z)$). In the literature of model-based optimistic algorithms for Markov decision processes, three main concentration inequalities are being used, taking the form below:

$$
\begin{aligned}
\text{Weissman's inequality:} \quad & N_z(t) \|\hat{p}_t(z) - p(z)\|_1^2 \leq f(t); \\
\text{Bernstein's inequality:} \quad & |\hat{p}_t(s|z) - p(s|z)| \leq \sqrt{\frac{2\hat{p}_t(s|z)(1 - \hat{p}_t(s|z)) f(t)}{N_z(t)}} + \frac{7 f(t)}{3 N_z(t)}; \\
\text{Empirical likelihoods:} \quad & N_z(t) \mathrm{KL}(\hat{p}_t(z) || p(z)) \leq f(t).
\end{aligned}
$$

Choosing one among these three respectively provides (in order) `UCRL2` Auer et al. (2009), `UCRL2B` Fruit et al. (2020) and `UCRL3` Bourel et al. (2020), and `KLUCRL` Filippi et al. (2010).

Best confidence region in theory On the theoretical side, asymptotically tight concentration inequalities are based on empirical likelihoods. The reason for that is Sanov's theorem. Let $p \in \mathcal{P}(\mathcal{S})$ be a categorical distribution and (X_k) i.i.d. random variables of distribution p. Let $\hat{p}_n := \frac{1}{n} \sum_{k=1}^n e_{X_k}$ denote the empirical estimation of p after n independent samples of it. Sanov's theorem states that for all $\mathcal{U} \subseteq \mathcal{P}(\mathcal{S})$, we have $\mathbf{P}(\hat{p}_n \in \mathcal{U}) = \exp\{-n \inf_{p' \in \mathcal{U}} \mathrm{KL}(p'||p) + \mathrm{o}(n)\}$. This justifies that $\mathrm{KL}(-||-)$ is a natural object to measure where $\hat{p}_n$ concentrates. We also find:

$$
\begin{aligned}
\mathbf{P}(\mathrm{KL}(\hat{p}_n||p) > x) &= \mathbf{P}(\hat{p}_n \in \{p' : \mathrm{KL}(p'||p) > x\}) \\
&= \exp\left\{ -n \inf_{p' \in \{p'' : \mathrm{KL}(p''||p) > x\}} \mathrm{KL}(p'||p) + \mathrm{o}(n) \right\} \\
&= \exp\{-nx + \mathrm{o}(n)\}.
\end{aligned}
$$

Therefore, KL-semi-balls naturally provide tight confidence regions. Such confidence regions are also variance aware (see Talebi and Maillard (2018)), which is crucial to provide minimax regret bounds that are better than the usual $\mathrm{O}(DS\sqrt{AT \log(T)})$.

Best confidence region in practice In practice, KLUCRL is known to perform well. This is displayed in Figure 4, where a selection of algorithms is run. The regret is averaged over 100 runs and at each run, the environment is picked uniformly at random among ergodic Markov decision processes—the environment is re-rolled every time to mitigate the possible over-specialization of some algorithms for some environments. We compare the performance of UCRL2 Auer et al. (2009), UCRL2B Fruit et al. (2020) and KLUCRL Filippi et al. (2010) that are all optimistic algorithms relying on EVI (Section B.1) to compute optimistic policies from their confidence regions. This is the framework described in details in Appendix B, for which our proof techniques for regret of exploration guarantees ((VM), Appendices C and D) are applicable. For fairness, these algorithms are reworked with state-of-the-art confidence regions in ℓ_1-norm, Bernstein's style and in empirical likelihoods.

In Figure 4, we observe that KLUCRL has better expected regret than the other algorithms.

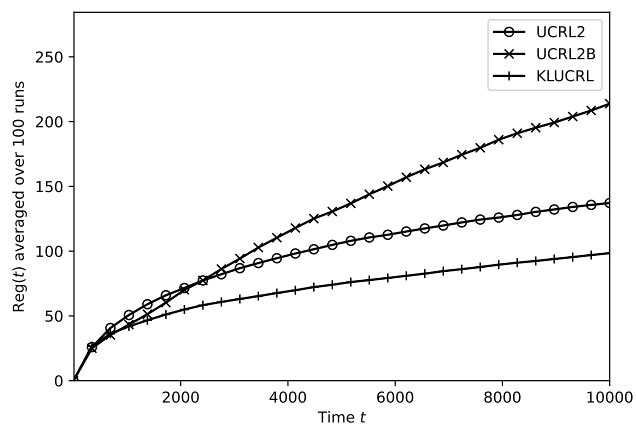


Figure 4: Bayesian regret of UCRL2, UCRL2B and KLUCRL. Each algorithm is run 100 times on an ergodic environment with 5 states and 2 actions, picked at random and renewed for every run.

A.2. regret guarantees of (VM) on experiments

The model independent regret of an algorithm is quite difficult to measure experimentally, because it is found as the expected regret on the worst environment $M \in \mathcal{M}$, that depends on the algorithm and is hard to find. Instead, we focus on the model dependent regret.

In Figure 5, we compare the behavior of KLUCRL when managing episodes with (DT) and with f-(VM) for $f(t) = \sqrt{\log(1+t)/t}$. The chosen environment is a small ergodic Markov decision process and a huge number of runs is done to accurately determine the distribution of the regret for both algorithms. We observe that both are bimodal with a concentration around the expected value, with KLUCRL−(VM) being better than KLUCRL−(DT) overall, both in expectation and in distribution (stochastic dominance).

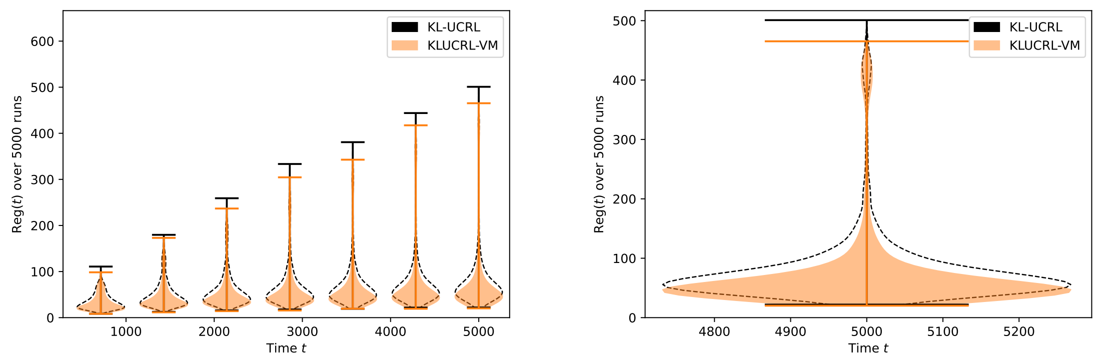


Figure 5: Violin plots of the regret of `KLUCRL` with episodes managed by ([DT](#)) (in **black** with dashed lines) and by ([VM](#)) (in **orange** with solid lines) on a small ergodic environment. By changing ([DT](#)) to ([VM](#)), we observe a slight improvement of the expected regret with an overall shift of its distribution to smaller values. These observations are uniform over the time horizon.

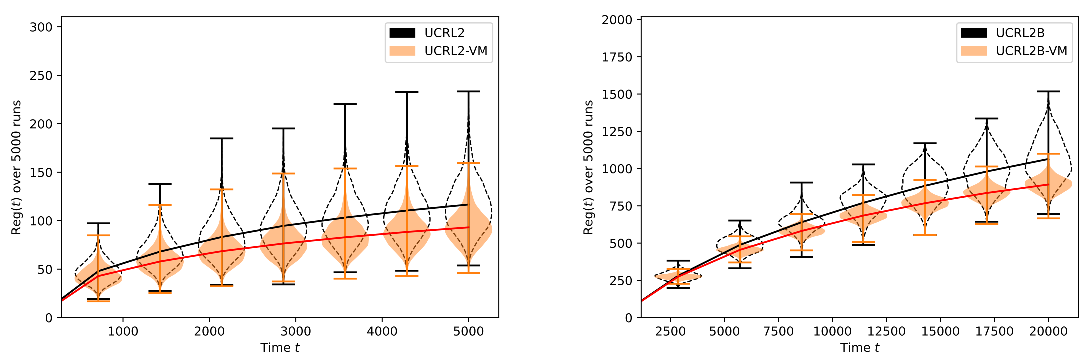


Figure 6: Violin plots of the regret of `UCRL2` and `UCRL2B` with episodes managed by ([DT](#)) (in **black** with dashed lines) and by ([VM](#)) (in **orange** with solid lines) on a small ergodic environment. For these algorithms, we further observe a reduction of the variance.

In Figure 6, we run the same experiments as in Figure 5 but with `UCRL2` Auer et al. (2009) and `UCRL2B` Fruit et al. (2020). The same observation than with `KLUCRL` can be made: the version relying on (VM) stochastically dominates the version relying on (DT). A phenomenon that is hard to see for `KLUCRL` but that is striking for `UCRL2` and `UCRL2B` is the reduction of the variance. Indeed, we can see that the distributions is much more concentrated around its mean with (VM) than with (DT).

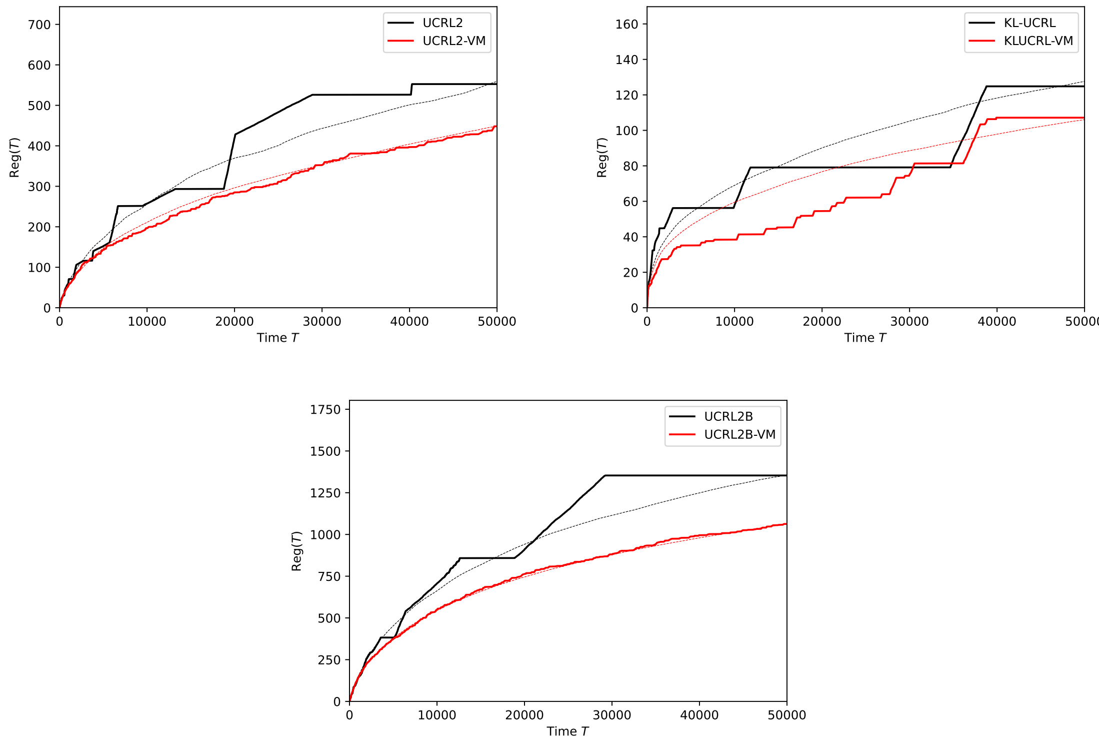


Figure 7: Pseudo-regret of a selection of algorithms on a fixed ergodic environment with 5 states and 2 actions picked at random. The dashed line is the average over 256 runs, while the solid line displays the pseudo-regret over a single trajectory.

In Figure 7, we run `UCRL2` Auer et al. (2009), `UCRL2B` Fruit et al. (2020) and `KLUCRL` Filippi et al. (2010) in their vanilla (with episodes managed by (`DT`)) and their reworked versions (with episodes managed with f-(`VM`) for $f(t) = \sqrt{\log(t)/t}$) tagged with (`VM`) in the legend. The environment is a fixed ergodic Markov decision process with 5 states and 2 actions per state, picked at random. We display the regret averaged over 256 runs with a dashed line, and the solid line is the pseudo-regret over a single trajectory, picked among those that minimize $\Delta(50000; M) - \mathrm{Reg}(T; M)$ for readability. The plotted average pseudo-regrets show that using (`VM`) rather than (`DT`) has a real advantage regarding regret minimization already. Looking at the single trajectory curves, we observe that the duration of periods of sub-optimal play is much shorter under f-(`VM`) than under (`DT`), for all three algorithms. Note that not all bad episodes are guaranteed to be small (see for e.g. the plot of `KLUCRL`). This is consistent with theory: A bound on the regret of exploration guarantees that periods of sub-optimal play are short in average, but does not rule out the existence of long periods of sub-optimal play.

A.3. The regret of exploration under (VM)

As the regret of exploration is a $\limsup$, it is impossible to measure it experimentally. We approximate it in finite time by looking at the quantity:

$$T \mapsto \max\{\mathrm{Reg}(t_{k(i)}, t_{k(i)} + T) : t_{k(i)} \in \{\psi(T_{\max}), \ldots, T_{\max}\}\} \tag{14}$$

where $T_{\max} \geq 1$ is the number of learning steps in the experiment and $\psi : \mathbf{N} \to \mathbf{N}$ is a threshold function. The threshold function satisfies $\psi(t) < t$. First, we want $\psi(t) \to \infty$ to remove the burn-in phase of the learning algorithm. Second, we want $\psi(t) = \mathrm{o}(t)$ to make sure that $\{\psi(t), \ldots, t\}$ contains many episodes of exploration so that the regret of exploration is estimated correctly.

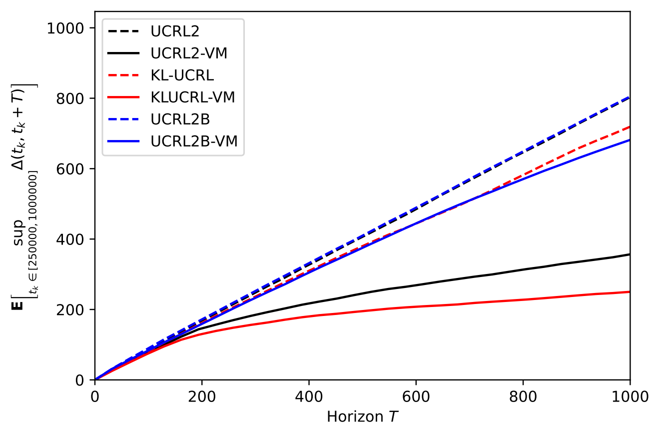


Figure 8: Estimation of the regret of exploration of several algorithms, following the proxy (14).

In Figure 8, we plot a proxy for the regret of exploration in the form of (14), with $T_{\max} = 10^6$ and $\psi(T_{\max}) = 10^5$. The environment is a small River-Swim with 3 states, known to be a hard-to-learn environment (see (Bourel et al., 2020, Figure 4)). We run UCRL2 Auer et al. (2009), UCRL2B Fruit et al. (2020) and KLUCRL Filippi et al. (2010) in their vanilla (with episodes managed by (DT)) and their reworked versions (with episodes managed with f-(VM) for $f(t) = \sqrt{\log(t)/t}$ tagged with (VM) in the legend.

We observe that the regret of exploration indeeds is a sub-linear function of T for all three algorithms under (VM), while their (DT) versions display a linear regret of exploration.

Appendix B. Minimax regret guarantees under (VM)

In this appendix, we establish the minimax regret guarantees as given by Theorem 7, assertion 1. We further provide a large range of general[1] results on EVI (Section B.1) and confidence regions (Section B.2) that will be used in other sections. In Section B.4, we provide a regret bound for instances of KLUCRL running with the f-(VM) rule for general non-increasing $f : \mathbf{N} \to [0, 1]$. This bound is to be combined with the bound on the number of episodes provided in Section B.5 to obtain Theorem 14.

Theorem 14 *Let $f : \mathbf{N} \to [0, 1]$ and consider running KLUCRL with episodes managed by f-(VM) and let $\mathcal{M}_D$ be the set of Markov decision processes with diameter less than D. If $f(t) = \Omega(t^{-1/2})$, then:*

$$\sup_{M' \in \mathcal{M}_D} \operatorname{Reg}(T; M') = \mathrm{O}\Big(DS\sqrt{AT\log(T)}\Big)$$

Notations. The empirical transition kernel and mean reward vector at learning step t are denoted $\hat{p}_t$ and $\hat{r}_t$. The policy played at time t is π_t. By design of EVI, the policy π that it returns at time t satisfies a Poisson equation (Corollary 16) of the form $\tilde{g}_t + \tilde{h}_t = \tilde{r}_t + \tilde{p}_t\tilde{h}_t$ where $\operatorname{sp}(\tilde{g}_t) = 0$, $\tilde{r}_t(s, \pi(s)) \in \mathcal{R}_{s,\pi(s)}(t)$ and $\tilde{p}_t(s, \pi(s)) \in \mathcal{P}_{s,\pi(s)}(t)$ for all $s \in \mathcal{S}$; and $\tilde{h}_t$ is the bias function of the Markov reward process $(\tilde{r}_t, \tilde{p}_t)$. As shown formally in Section B.1 thereafter, $\tilde{g}_t = g^*(\mathcal{M}(t))$ is the optimal gain of $\mathcal{M}(t)$ and

$$\operatorname{sp}(\tilde{h}_t) \leq D(\mathcal{M}(t)) \tag{15}$$

which is bounded by $D(M)$ as soon as $M \in \mathcal{M}(t)$. We further introduce $\mathcal{K}(T) := \{k \in \mathbf{N} : t_k \leq T\}$ the set of episodes starting prior to $T \geq 1$.

B.1. Properties of Extended Value Iteration (EVI) and extended MDPs

When the confidence region $\mathcal{M}(t)$ is in product form $\mathcal{M}(t) \equiv \prod_{z \in \mathcal{Z}} \mathcal{R}_z(t) \times \mathcal{P}_z(t)$, such as in our case (see Section 2.3 and Equation (5)), it can be seen as a single Markov decision process with compact action space by *extending* actions. This extended formulation of $\mathcal{M}(t)$ goes back to Auer et al. (2009) and is what allows to interpret the optimistic gain (4) as the optimal gain function of $\mathcal{M}(t)$ seen as a Markov decision process. Specifically, the extended action space of $\mathcal{M}(t)$ from $s \in \mathcal{S}$ is:

$$\tilde{\mathcal{A}}(s; t) := \prod_{a \in \mathcal{A}(s)} (\mathcal{R}_{s,a}(t) \times \mathcal{P}_{s,a}(t))$$

that we may more simply write $\tilde{\mathcal{A}}(s)$. Accordingly, a choice of action in $\mathcal{M}(t)$ consists in a choice of a vanilla action from s (i.e., $a \in \mathcal{A}(s)$) as well as a plausible reward and transition kernel for that choice of action. Policies of $\mathcal{M}(t)$ are **extended policies**, and take the form of (π, r', p') where π is a policy of M and r', p' are plausible choices of reward function and

1. With the exception of the pioneer work of Auer et al. (2009), previous works tend to overlook the well-behavior of EVI from a theoretical perspective. Section B.1 provides a more rigorous treatment.

transition kernel for that policy. Extended Value Iteration (`EVI`) consists in iterating the **Bellman operator** (Puterman, 1994, §8.5) of $\mathcal{M}(t)$ seen as an extended MDP. In the case of $\mathcal{M}(t)$, its Bellman operator is given by $\mathcal{L}(t) \equiv L(\mathcal{M}(t)) : \mathbf{R}^{\mathcal{S}} \to \mathbf{R}^{\mathcal{S}}$,

$$(\mathcal{L}(t)u)(s) = \max_{a \in \mathcal{A}(s)} \max_{r \in \mathcal{R}_{s,a}(t)} \max_{p \in \mathcal{P}_{s,a}(t)} \left\{ r(s,a) + p(s,a)u \right\}. \tag{16}$$

`EVI` consists in iterating $\mathcal{L}(t)$ until convergence to a near span-fixpoint, i.e., in computing $u_{n+1} = \mathcal{L}(t)u_n$ until $\mathrm{sp}(u_{n+1} - u_n) < \epsilon$ where ϵ is the desired numerical precision. Once the condition "$\mathrm{sp}(u_{n+1} - u_n) < \epsilon$" is reached, the algorithm returns the policy $\pi : \mathcal{S} \to \mathcal{A}$ such that $\pi(s)$ is a choice of action achieving the maximum in (16) for $u = u_n$. This is the algorithm Value Iteration (Puterman, 1994, §8.5) applied to $\mathcal{M}(t)$.

While `EVI` performs very well in practice and rarely struggles to converge, there is actually no existing theoretical guarantees regarding its convergence.

MDPs with compact action spaces are to be treated with care, especially because the existence of solutions to the Bellman equations is not always guaranteed. As a consequence to this, the convergence of the iterates of Bellman operators is not guaranteed in general. This issue has been largely overlooked in the reinforcement learning literature and curious readers can take a look at Schweitzer (1985) for that matter. In Auer et al. (2009) for `UCRL2`, the authors do address this issue and argue that the maximum in (16) must be achieved at some vertex of the polytope given by the ℓ_1-ball spawn by the confidence region. Therefore, the maximum is always a maximum over finitely many elements, so $\mathcal{M}(t)$ can be reduced to an extended MDP with *finite* action space. This argument can be replicated for confidence regions based on empirical Bernstein inequalities such as for `UCRL2B` Fruit et al. (2020), although not explicitly mentioned. It fails completely when $\mathcal{P}_z(t)$ has smooth boundary, such as for `KLUCRL` Filippi et al. (2010) and confidence regions used here. Thankfully and in general, MDPs with compact action spaces are much better behaved when they are **communicating** (see Assumption 2). Thankfully again, this is the case of $\mathcal{M}(t)$.

Proposition 15 (Schweitzer (1987)) *Let $\mathcal{M}$ be a communicating Markov decision process with finite state space $\mathcal{S}$ and compact action space $\mathcal{A}$. Assume that $r(z) \in [0, 1]$ and that $a \in \mathcal{A}(s) \mapsto r(s, a)$ and $a \in \mathcal{A}(s) \mapsto p(s, s)$ are continuous functions. Then:*

1. *Its Bellman operator $L : \mathbf{R}^s \to \mathbf{R}^{\mathcal{S}}$ given by $(Lu)(s) = \max_{a \in \mathcal{A}(s)}\{r(s, a) + p(s, a)u\}$ admits a span-fixpoint, i.e., $\exists u \in \mathbf{R}^{\mathcal{S}}$ such that $\mathrm{sp}(Lu - u) = 0$;*

2. *If $p(s|s, a) > 0$ for all $(s, a) \in \mathcal{Z}$, then the iterates of the Bellman operator converge to a span-fixpoint with linear convergence speed, i.e., there is $\gamma < 1$ such that for all $u \in \mathbf{R}^{\mathcal{S}}$, $\mathrm{sp}(L^{n+1}u - L^n u) = \mathrm{O}(\gamma^n)$ when $n \to \infty$.*

The span-fixpoint to which L converges is denoted h^*, is the optimal bias function of $\mathcal{M}$, and satisfies a Bellman equation $g^* + h^* = \max_{a \in \mathcal{A}(s)}\{r(s, a) + p(s, a)h^*\}$. The technical condition "$p(s|s, a) > 0$ for all $(s, a) \in \mathcal{Z}$" can always be guaranteed under an aperiodicity transform of $\mathcal{M}$, see (Puterman, 1994, §8.5.4) and (Bartlett and Tewari, 2009, §4), that consists in iterating $\frac{1}{2}(L + \mathrm{Id})$ instead of L. This aperiodicity transform indeed improves the convergence speed of `EVI` in practice, without modification of the quality of its output policy. The communicativity assumption is always satisfied, because by design of the confidence region, $\mathcal{P}_z(t)$ contains fully-supported transition kernel for all $z \in \mathcal{Z}$ and $t \geq 1$.[2] In the end, we can provide generic guarantees for the convergence of `EVI`.

Corollary 16 *Let $M = (\mathcal{Z}, r, p)$ be a communicating Markov decision process. Let $\mathcal{M}(t) \equiv \prod_{z \in \mathcal{Z}} \mathcal{R}_z(t) \times \mathcal{P}_z(t)$ be a compact confidence region for M. Assume that $\mathcal{R}_z(t) \subseteq [0, 1]$ and that, for all $z \in \mathcal{Z}$, $\mathcal{P}_z(t)$ contains some $p'(z)$ with $\mathrm{supp}(p'(z)) \supseteq \mathrm{supp}(p(z))$. Then:*

1. *The extended Bellman operator $\mathcal{L}(t)$, see (16), admits a span-fixpoint and the optimistic gain $g^*(\mathcal{M}(t))$ of (4) is the optimal gain of the extended MDP $\mathcal{M}(t)$;*
2. *The iterates of $\frac{1}{2}(\mathcal{L}(t) + \mathrm{Id})$ converge linearly fast to a span-fixpoint of $\mathcal{L}(t)$, $h^*(\mathcal{M}(t))$, that satisfies the Bellman equation*

$$g^*(s; \mathcal{M}(t)) + h^*(s; \mathcal{M}(t)) = \max_{r'(s,a) \in \mathcal{R}_{s,a}(t)} \max_{p'(s,a) \in \mathcal{P}_{s,a}(t)} \left\{ r'(s, a) + p'(s, a) h^*(\mathcal{M}(t)) \right\}.$$

Therefore, the extended policy (π, r', p') achieving $\mathcal{L}(t)h^*(\mathcal{M}(t))$ satisfies the Poisson equation $g^*(s; \mathcal{M}(t)) + h^*(s; \mathcal{M}(t)) = r'(s) + p'(s)h^*(\mathcal{M}(t))$.

A last property that is crucial in the regret analysis of `EVI`-based algorithms is that their optimal bias h^* given by Corollary 16 have small span. This result is well-known, see for example (Fruit, 2019, Proposition 3.6). We provide a short proof for self-containedness.

Lemma 17 *Let $\mathcal{M}$ be a communicating Markov decision process with finite state space $\mathcal{S}$ and compact action space $\mathcal{A}$. Assume that $r(z) \in [0, 1]$ and that $a \in \mathcal{A}(s) \mapsto r(s, a)$ and $a \in \mathcal{A}(s) \mapsto p(s, s)$ are continuous functions. Let h^* be a span-fixpoint of its Bellman operator. Then:*

$$\mathrm{sp}(h^*) \leq D(\mathcal{M})$$

where $D(\mathcal{M})$ is the diameter of $\mathcal{M}$, as given by (2).

Proof Fix two states $s, s' \in \mathcal{S}$ and let π such that $\mathbf{E}_s^\pi[\tau_{s'}] < \infty$ where $\tau_{s'} := \inf\{t > 1 : S_t = s'\}$ is the reaching time to s'. Since $g^*(s) + h^*(s) \geq r(s, \pi(s)) + p(s, \pi(s))h^*$, we have:

$$\begin{aligned} 0 &\leq \mathbf{E}_s^\pi \left[\sum_{t=1}^{\tau_{s'}-1} (g^*(S_t) - r(Z_t) + (e_{S_t} - p(Z_t))h^*) \right] \\ &\overset{(\dagger)}{\leq} \mathrm{sp}(r)\mathbf{E}_s^\pi[\tau_{s'}] + h^*(s) - h^*(s') \end{aligned}$$

where $(\dagger)$ follows from Doob's optional stopping theorem and that $\mathrm{sp}(g^* - r) \leq \mathrm{sp}(r)$. By taking the policy minimizing $\mathbf{E}_s^\pi[\tau_{s'}]$, we conclude that $h^*(s') - h^*(s) \leq D(\mathcal{M})$. Because this holds for arbitrary $s, s' \in \mathcal{S}$, we conclude that $\mathrm{sp}(h^*) = \max(h^*) - \min(h^*) \leq D(\mathcal{M})$. ■

2. If `KLUCRL` is ran with prior information on the support of $p(z)$, then $\mathcal{P}_z(t)$ always contains elements with the same support than $p(z)$—and the communicativity assumption only depends on the support of transition kernels, independently of how small the transition probabilities can be. So $\mathcal{M}(t)$ is communicating when M is communicating, which is the case in this work.

B.2. The confidence region of KLUCRL

The confidence region of KLUCRL is designed to hold with high probability (Lemma 18).

Lemma 18 *The confidence region holds with high probability*

$$\mathbf{P}(\exists t \geq T : M \notin \mathcal{M}(t)) \leq 2|\mathcal{Z}|T^{-1}.$$

Proof This result follows by a time-uniform concentration inequality for empirical likelihoods of Jonsson et al. (2020), see their Proposition 1. (Jonsson et al., 2020, Proposition 1) states the following: Given $d \geq 2$, and $p \in \mathcal{P}[d]$ a probability distribution over $\{1, \ldots, d\}$, if $\hat{p}_n \in \mathcal{P}[d]$ denotes the empirical average of n i.i.d. samples of p, then for all $\delta \geq 0$,

$$\mathbf{P}\left(\exists n \geq 1 : n\mathrm{KL}(\hat{p}_n||p) > \log\left(\frac{1}{\delta}\right) + (d-1)\log\left(e\left(1 + \frac{n}{d-1}\right)\right)\right) \leq \delta.$$

In our case, we readily obtain that for all $z \in \mathcal{Z}$ and $t \geq 1$,

$$\mathbf{P}\left(N_z(t)\mathrm{KL}(\hat{p}_t(z)||p(z)) > \log(t) + (|\mathcal{S}|-1)\log\left(e\left(1 + \frac{N_z(t)}{|\mathcal{S}|-1}\right)\right)\right) \leq \frac{1}{t}. \tag{17}$$

Since $N_z(t) \leq t - 1$, we have in particular that for all $z \in \mathcal{Z}$ and $t \geq 1$,

$$\mathbf{P}\Big(N_z(t)\mathrm{KL}(\hat{p}_t(z)||p(z)) > |\mathcal{S}|\log(2et)\Big) \leq \frac{1}{t} \tag{18}$$

where we recognize the definition of $\mathcal{P}_z(t)$ in (5). Rewards are done similarly, applying (Jonsson et al., 2020, Proposition 1) for $d = 2$. Conclude by union bound over $q \in \{r, p\}$ and $z \in \mathcal{Z}$. ■

Note that there is a significant loss of information when going from (17) to (18), in the sense that (17) is much more precise than (18). It means that we could take a much more precise confidence region than the one used in (5). The confidence region has been simplified to ease the calculations in the proof of the shrinking effect (Lemma 26), see Section D.2.

B.3. Bounds of classical error terms

The maximal version of Hoeffding's inequality below (Lemma 19) is a standard result from Hoeffding Hoeffding (1963). It is used in the proofs of Lemmas 20 and 22 in integrated form to bound the error due to optimism.

Lemma 19 (Hoeffding (1963)) *Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. random variables in $[0, 1]$ and let $\hat{\mu}_n$ their empirical mean after n samples. Let $\mu := \mathbf{E}[X_1]$ be the true mean. Then, for all $x \geq 0$ and $m \geq 1$, we have:*

$$\mathbf{P}\left(\max_{n\geq m}\{\hat{\mu}_n - \mu\} \geq x\right) \leq \exp\{-2mx^2\}.$$

Lemma 20 *The expected cumulative optimistic reward error is bounded as follows:*

$$\mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\left[\tilde{r}_{t_k}(Z_t) - r(Z_t)\right]_+\right] = \mathrm{O}\Big(\sqrt{|\mathcal{Z}|T\log(T)}\Big).$$

Proof We write:

$$
\begin{aligned}
&\mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\left[\tilde{r}_{t_k}(Z_t)-r(Z_t)\right]_+\right] \\
&\overset{(\dagger)}{\leq} \mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\left[\tilde{r}_{t_k}(Z_t)-\hat{r}_{t_k}(Z_t)\right]_+\right] + \mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\left[\hat{r}_{t_k}(Z_t)-r(Z_t)\right]_+\right] \\
&\overset{(\ddagger)}{\leq} \mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{2\mathrm{kl}(\hat{r}_{t_k}(Z_t)||\tilde{r}_{t_k}(Z_t))}\right] + \mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\left[\hat{r}_{t_k}(Z_t)-r(Z_t)\right]_+\right],
\end{aligned}
$$

where $(\dagger)$ follows by sub-additivity of $[-]_+$ and $(\ddagger)$ by Pinsker's inequality.

The first term is bounded as follows. By construction of the confidence region (5), we have $N_{t_k}(z_t)\mathrm{kl}(\tilde{r}_{t_k}(Z_t)||\tilde{r}_{t_k}(Z_t)) \leq \log(Te(1+N_z(t_k))) \leq 2\log(T)+1$. We obtain:

$$
\begin{aligned}
\mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{2\mathrm{kl}(\hat{r}_{t_k}(Z_t)||\tilde{r}_{t_k}(Z_t))}\right] &\leq \mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{2(2\log(T)+1)}{N_{Z_t}(t_k)}}\right] \\
&\overset{(\dagger)}{\leq} \mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{4(2\log(T)+1)}{N_{Z_t}(t)}}\right] \\
&\leq 2\sqrt{2\log(T)+1}\cdot\mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sum_{n=1}^{N_z(T)}\frac{1}{\sqrt{n}}\right] \\
&\leq 4\sqrt{2\log(T)+1}\cdot\mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sqrt{N_z(T)}\right] \\
&\overset{(\ddagger)}{\leq} 4\sqrt{2\log(T)+1}\cdot\sqrt{|\mathcal{Z}|T}
\end{aligned}
$$

where $(\dagger)$ follows from the observation that, under f-(VM), we have $N_z(t_k) \leq 2N_z(t)$ and $(\ddagger)$ by Cauchy-Schwartz' inequality.

We continue by bounding the second term. In the computation below, we denote $\hat{r}_{(n)}(z)$ the empirical reward at $z \in \mathcal{Z}$ after exactly n samples of it. In particular, note that $r_{(N_z(t))}(z) = r_t(z)$. We have:

$$
\begin{aligned}
&\mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\left[\hat{r}_{t_k}(Z_t)-r(Z_t)\right]_+\right] \\
&= \mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}(Z_t=z)\left[\hat{r}_{(N_z(t_k))}(z)-r(z)\right]_+\right] \\
&\overset{(\dagger)}{\leq} \mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}(Z_t=z)\max_{n\geq\lfloor\frac{1}{2}N_z(t)\rfloor}\left[\hat{r}_{(n)}(z)-r(z)\right]_+\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sum_{m=1}^{N_z(T)} \max_{n\geq\lfloor\frac{1}{2}m\rfloor}\big[\hat{r}_{(n)}(z) - r(z)\big]_+\right] \\
&\leq 2\cdot\mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sum_{m=1}^{N_z(T)} \max_{n\geq m}\big[\hat{r}_{(n)}(z) - r(z)\big]_+\right] \\
&\overset{(\ddagger)}{=} 2\cdot\mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sum_{m=1}^{N_z(T)} \int_0^\infty \mathbf{P}\left(\max_{n\geq m}\big[\hat{r}_{(n)}(z) - r(z)\big]_+ \geq x\right)\mathrm{d}x\right] \\
&\overset{(\S)}{\leq} 2\cdot\mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sum_{m=1}^{N_z(T)} \int_0^\infty \exp\{-2mx^2\}\mathrm{d}x\right] \\
&= \mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sum_{m=1}^{N_z(T)}\sqrt{\frac{\pi}{2m}}\right] \leq \sqrt{2\pi}\mathbf{E}\left[\sum_{z\in\mathcal{Z}}\sqrt{N_z(T)}\right] \overset{(\$)}{\leq} \sqrt{2\pi|\mathcal{Z}|T}
\end{aligned}
$$

where $(\dagger)$ follows from the observation that $N_t(z) \leq 2N_{t_k}(z)$ for $t \in \{t_k, \ldots, t_{k+1}-1\}$; $(\ddagger)$ follows from Doob's optional stopping theorem; $(\S)$ follows from Lemma 19 and $(\$)$ is obtained with Cauchy-Schwartz' inequality. ■

We obtain a similar result for transition kernels, by changing $[-]_+$ to $\|-\|_1$, invoking the time-uniform concentration result of (Jonsson et al., 2020, Proposition 1) of empirical likelihoods in dimension $d = |\mathcal{S}|$ instead of $d = 2$. Lemma 19 has to be modified to take into account these modifications, see Lemma 21 below, which is a maximal version of Weissman's inequality Weissman et al. (2003).

Lemma 21 *Let $p \in \mathcal{P}[d]$ for $d \geq 2$ and let (X_k) be a sequence of i.i.d. samples of p. Denote $\hat{p}_n := \frac{1}{n}(e_{X_1} + \ldots + e_{X_k})$ the empirical distribution after n samples. Then, for all $x \geq 0$ and $m \geq 1$, we have:*

$$
\mathbf{P}\left(\max_{n\geq m}\|\hat{p}_n - p\|_1 \geq x\right) \leq \exp\big\{-2mx^2 + |\mathcal{S}|\log(2)\big\}.
$$

Proof Note that $\|\hat{p}_n - p\|_1 = \max_{u\in\{-1,1\}^\mathcal{S}}(\hat{p}_n - p)\cdot u$. So, introducing the notations $X_k^u := e_{X_k}\cdot u$ for $u \in \{-1,1\}^\mathcal{S}$ together with $\hat{\mu}_n^u := \frac{1}{n}(X_1^u + \ldots + X_n^u)$ and $\mu^u = \mathbf{E}[X_1^u]$, we have $\|\hat{p}_n - p\|_1 = \max_{u\in\{-1,1\}^\mathcal{S}}(\hat{\mu}_n^u - \mu^u)$. So,

$$
\begin{aligned}
\mathbf{P}\left(\max_{n\geq m}\|\hat{p}_n - p\|_1 \geq x\right) &= \mathbf{P}\left(\exists u \in \{-1,1\}^\mathcal{S} : \max_{n\geq m}\{\hat{\mu}_n^u - \mu^u\} \geq x\right) \\
&\leq \sum_{u\in\{-1,1\}^\mathcal{S}} \mathbf{P}\left(\max_{n\geq m}\{\hat{\mu}_n^u - \mu^u\} \geq x\right) \overset{(\dagger)}{\leq} 2^{|\mathcal{S}|}\exp\{-2mx^2\}
\end{aligned}
$$

where $(\dagger)$ follows from Lemma 19. ■

Lemma 22 *The expected cumulative optimistic error on kernels is bounded as follows:*

$$\mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}\|\tilde{p}_{t_k}(Z_t)-p(Z_t)\|_1\right]=\mathrm{O}\Big(\sqrt{|\mathcal{S}||\mathcal{Z}|T\log(T)}\Big).$$

Proof Same proof as Lemma 20, changing $[-]_+$ for $\|-\|_1$, taking care of the extra $|\mathcal{S}|$ in the confidence region for kernels, and invoking Lemma 21 instead of Lemma 19. ■

B.4. Bounding the regret relatively to the number of episodes

In this section, we prove in Lemma 23 that the regret under (VM) can be decoupled as the classical term $SD\sqrt{AT\log(T)}$ and another which is proportional to the number of episodes. The regret analysis is classical and inspired from Auer et al. (2009) for UCRL2, excepted that the analysis is written in expectation rather than in probability. The analysis could be adapted to obtain a result in probability as well.

Lemma 23 *Consider running* KLUCRL *while managing episodes with the* f*-*(VM) *rule for some non-increasing* $f:\mathbf{N}\to[0,1]$*. For all* M *with diameter less than* $D>0$*, we have:*

$$\mathrm{Reg}(T;M)=\mathrm{O}\Big(D\sqrt{|\mathcal{S}||\mathcal{Z}|T\log(T)}+D\mathbf{E}^M\big|\mathcal{K}(T)\big|\Big)$$

Proof We have $\mathrm{Reg}(T;M)=\mathbf{E}[\sum_{t=1}^T\Delta^*(Z_t;M)]$, so

$$\begin{aligned}
&\mathrm{Reg}(T;M)\\
&\le\mathbf{E}\left[\sum_{t=1}^T(g^*(M)-R_t)\right]+\mathrm{sp}(h^*(M))\\
&\overset{(\dagger)}{=}\mathbf{E}\left[\sum_{k\in\mathcal{K}(T)}\sum_{t=t_k}^{t_{k+1}-1}(g^*(M)-r(Z_t))\right]+\mathrm{sp}(h^*(M))\\
&\le\sqrt{|\mathcal{Z}|T}+\mathbf{E}\left[\sum_{k\in K(T)}\mathbf{1}\Big(t_k\ge\sqrt{|\mathcal{Z}|T},M\in\mathcal{M}(t_k)\Big)\sum_{t=t_k}^{t_{k+1}-1}(g^*(M)-r(Z_t))\right]\\
&\qquad\qquad+T\cdot\mathbf{P}\Big(\exists t\ge\sqrt{|\mathcal{Z}|T}:M\notin\mathcal{M}(t)\Big)+\mathrm{sp}(h^*(M))\\
&\overset{(\ddagger)}{=}\underbrace{\mathbf{E}\left[\sum_{k\in K(T)}\mathbf{1}\Big(t_k\ge\sqrt{|\mathcal{Z}|T},M\in\mathcal{M}(t_k)\Big)\sum_{t=t_k}^{t_{k+1}-1}(g^*(M)-r(Z_t))\right]}_{\mathrm{A}}+\mathrm{O}\Big(\sqrt{|\mathcal{Z}|T}\Big)
\end{aligned}$$

where $(\dagger)$ follows from the tower property and $(\ddagger)$ follows from Lemma 18, stating that $\mathbf{P}(\exists t\ge T:M\notin\mathcal{M}(T))\le2|\mathcal{Z}|T^{-1}$. We focus on the first expectation. Further introduce the good event $\mathcal{E}_t := \{t \geq \sqrt{|\mathcal{Z}|T}, M \in \mathcal{M}(t)\}$. It is $\sigma(O_t)$-measurable. At time t_k and under $\mathcal{E}_{t_k}$, we have $\tilde{g}_{t_k} = g^*(\tilde{\mathcal{M}}(t_k)) \geq g^*(M)$ and $\mathrm{sp}(\tilde{h}_{t_k}) \leq D(M)$ (see Section B.1). So:

$$
\begin{aligned}
\mathrm{A} := \mathbf{E}\left[\sum_{k \in K(T)} \mathbf{1}(\mathcal{E}_{t_k}) \sum_{t=t_k}^{t_{k+1}-1} (g^*(M) - r(Z_t))\right]
&\leq \mathbf{E}\left[\sum_{k \in K(T)} \mathbf{1}(\mathcal{E}_{t_k}) \sum_{t=t_k}^{t_{k+1}-1} (\tilde{g}_{t_k} - r(Z_t))\right] \\
&= \mathbf{E}\left[\sum_{k \in K(T)} \mathbf{1}(\mathcal{E}_{t_k}) \left(\sum_{t=t_k}^{t_{k+1}-1} (\tilde{g}_{t_k} - \tilde{r}_{t_k}(Z_t)) + \sum_{t=t_k}^{t_{k+1}-1} (\tilde{r}_{t_k}(Z_t) - r(Z_t)) \right)\right] \\
&\leq \mathbf{E}\left[\sum_{k \in K(T)} \mathbf{1}(\mathcal{E}_{t_k}) \sum_{t=t_k}^{t_{k+1}-1} (\tilde{g}_{t_k} - \tilde{r}_{t_k}(Z_t))\right] + \mathbf{E}\left[\sum_{k \in K(T)} \sum_{t=t_k}^{t_{k+1}-1} [\tilde{r}_{t_k}(Z_t) - r(Z_t)]_+\right] \\
&\overset{(\dagger)}{=} \mathbf{E}\left[\sum_{k \in K(T)} \mathbf{1}(\mathcal{E}_{t_k}) \sum_{t=t_k}^{t_{k+1}-1} (\tilde{g}_{t_k} - \tilde{r}_{t_k}(Z_t))\right] + \mathrm{O}\left(\sqrt{|\mathcal{Z}|T \log(T)}\right)
\end{aligned}
$$

where $(\dagger)$ follows from Lemma 20. We proceed as follows:

$$
\begin{aligned}
&\mathbf{E}\left[\sum_{k \in K(T)} \mathbf{1}(\mathcal{E}_{t_k}) \sum_{t=t_k}^{t_{k+1}-1} (\tilde{g}_{t_k} - \tilde{r}_{t_k}(Z_t))\right] \\
&\overset{(\dagger)}{=} \mathbf{E}\left[\sum_{k \in K(T)} \mathbf{1}(\mathcal{E}_{t_k}) \sum_{t=t_k}^{t_{k+1}-1} (e_{S_t} - \tilde{p}_{t_k}(Z_t))\tilde{h}_{t_k}\right] \\
&= \mathbf{E}\left[\sum_{k \in K(T)} \mathbf{1}(\mathcal{E}_{t_k}) \left(\sum_{t=t_k}^{t_{k+1}-1} (e_{S_t} - p(Z_t))\tilde{h}_{t_k} + \sum_{t=t_k}^{t_{k+1}-1} (p(Z_t) - \tilde{p}_{t_k}(Z_t))\tilde{h}_{t_k} \right)\right] \\
&\overset{(\ddagger)}{\leq} \mathbf{E}\left[\sum_{k \in K(T)} \mathbf{1}(\mathcal{E}_{t_k}) \left(\tilde{h}_{t_k}(S_{t_k}) - \tilde{h}_{t_k}(S_{t_{k+1}}) \right)\right] + \frac{D}{2}\mathbf{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \sum_{k \in K(T)} \|\tilde{p}_{t_k}(Z_t) - p(Z_t)\|_1\right] \\
&\overset{(\S)}{\leq} \mathbf{E}\left[\sum_{k \in K(T)} D\right] + \mathrm{O}\left(D\sqrt{|\mathcal{S}||\mathcal{Z}|T \log(T)}\right) = \mathrm{O}\left(D\mathbf{E}|\mathcal{K}(T)| + D\sqrt{|\mathcal{S}||\mathcal{Z}|T \log(T)}\right)
\end{aligned}
$$

where $(\dagger)$ follows from the Poisson equation $\tilde{g}_{t_k} - \tilde{r}_{t_k}(s, a) = \tilde{h}_{t_k}(s) - \tilde{p}_{t_k}(s, a)\tilde{h}_{t_k}$; $(\ddagger)$ is obtained using the telescopic nature of the first term and by using that $(p(z) - \tilde{p}_{t_k}(z))\tilde{h}_{t_k} \leq \frac{1}{2}\mathrm{sp}(\tilde{h}_{t_k})\|\tilde{p}_{t_k}(z) - p(z)\|_1$ to bound the second, and further using that $\mathrm{sp}(\tilde{h}_{t_k}) \leq D(\mathcal{M}(t_k)) \leq D(M)$ on $\mathcal{E}_{t_k}$ by Lemma 17; and $(\S)$ follows by bounding the first term using that $\mathbf{1}(\mathcal{E}_{t_k})\mathrm{sp}(\tilde{h}_{t_k}) \leq D(M)$ (Lemma 17) and by bounding the second using Lemma 22. We conclude accordingly. ■

B.5. Bounding the number episodes under f-(VM)

The episodes under f-(VM) are bounded in a similar than for (DT). The technique that we provide below provides a result that ends up being asymptotically better than (Auer et al., 2009, Proposition 18) for the doubling trick, $|\mathcal{K}(T)| \leq |\mathcal{Z}| \log_2(\frac{8T}{|\mathcal{Z}|})$.

Lemma 24 *Assume that episodes are managed with f-(VM) where $f : \mathbf{N} \to (0, 1]$ is non-increasing. Whatever $M \in \mathcal{M}$, we have*

$$|\mathcal{K}(T)| \leq \frac{|\mathcal{Z}| \log\left(\frac{2T+\mathrm{O}(1)}{|\mathcal{Z}|}\right)}{\log(1 + f(T))} \quad \text{a.s.}$$

Proof Given $z \in \mathcal{Z}$, let $\mathcal{K}_z(T) := \left\{k : t_k \leq T \text{ and } N_z(t_{k+1}) > (1+f(t_k)) \max\{1, N_z(t_k)\}\right\}$ the set of episodes that are ended by visiting $z \in \mathcal{Z}$. Remark that:

$$N_z(t_{k+1}) \geq \prod_{\ell \in \mathcal{K}_z(t_k)} (1 + f(t_\ell)) \geq (1 + f(T))^{|\mathcal{K}_z(t_k)|}.$$

We have $t_{k+1} \leq 2t_k + \mathrm{O}(1)$ when $k \to \infty$, hence summing the above over $z \in \mathcal{Z}$, we obtain:

$$2T \geq \sum_{z \in \mathcal{Z}} (1 + f(T))^{|\mathcal{K}_z(T)|} \geq \inf_{\omega \in \mathcal{P}(\mathcal{Z})} \left\{ \sum_{z \in \mathcal{Z}} (1 + f(T))^{\omega_z |\mathcal{K}(T)|} \right\}$$

where the second inequality follows from the observation that the union $\mathcal{K}(T) = \bigcup_{z \in \mathcal{Z}} \mathcal{K}_z(T)$ is disjoint. The RHS of the above is the infemum of a convex function $\psi(\omega)$. The KKT conditions show that this infemum is reached when $\omega_z = |\mathcal{Z}|^{-1}$ for all $z \in \mathcal{Z}$. Plugging these values in the above and solving in $|\mathcal{K}(T)|$, we obtain the desired result. ■

Appendix C. The coherence lemma: Proof of Lemma 12

In this appendix, we provide a proof of the coherence lemma (Lemma 12). Stated in its general form, this lemma can be instantiated in various forms to obtain a large variety of results. It is used *twice* to provide the regret of exploration guarantees of KLUCRL, first in a macroscopic (or global) way to provide the asymptotic regime (see Section 4.1 and Lemma 8) and lastly in a microscopic (or local) way to finally provide regret of exploration guarantees (see Section 4.4 and Lemma 13). It is also used to obtain instance dependent regret guarantees (Appendix E and Theorem 31), showing that every instance KLUCRL managing episodes with f-(VM) is consistent on the sub-space of non-degenerate Markov decision processes.

We recall the statement of Lemma 12 below.

Lemma 12 (Coherence and local regret) *Assume that M is non-degenerate (Definition 4). If the algorithm is (F, τ, T, φ)-coherent and has weakly regenerative episodes, then there exist model dependent constants $C_1, C_2, C_3, C_4 > 0$ such that:*

$$\forall x \geq 0, \quad \mathbf{P}\left(\Delta(\tau, \tau + T) \geq x + C_4\varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t\right) \leq C_1 T^{C_3} \exp\left(-\frac{x}{C_2}\right).$$

More specifically, C_1, C_2, C_3, C_4 only depend on M and are independent of F, τ, T and φ.

Outline of the proof The whole appendix is dedicated to a proof of Lemma 12. In Section C.1, the time-range $\{\tau, \ldots, \tau + T - 1\}$ is partioned into segments $\{\tau_i, \ldots, \tau_{i+1} - 1\}$ alternating between periods of sub-optimal and optimal play. We start by bounding the regret due to sub-optimal segments in Section C.2. In (**STEP 1**), we relate the total duration and the number of sub-optimal segments to the potential $\varphi(\tau)$, to show in (**STEP 2**) that the number of sub-optimal time-segments has sub-exponential tails under the good event $\bigcap_{t=\tau}^{\tau+T-1} F_t$. This leads to sub-exponential tails for the total duration of sub-optimal segments in (**STEP 3**) under the same good event. It provides an immediate regret bound for the regret induced by sub-optimal periods of play in (**STEP 4**). In Section C.3, we move to the bound of the regret on optimal segments, where the algorithm plays gain optimal policies. However, even if the algorithm plays a gain optimal policy on $\{\tau_i, \ldots, \tau_{i+1} - 1\}$, it may play a few sub-optimal actions before the recurrent class of that policy is reached: This is the well-known "cost" induced by switching policies. Therefore, we motivate in Section C.3 that we need to bound the time that the algorithm takes to reach the optimal class on *all* optimal segments. This is related to the number of sub-optimal segments in (**STEP 1**), and as optimal and sub-optimal segments alternate by construction, the work done in Section C.2 provides a bound on that number. This leads to a sub-exponential tails for the regret induced by optimal segments in (**STEP 3**). Everything is combined in Section C.4 to conclude the proof of Lemma 12.

Notations Given a policy $\pi \in \Pi$ and a state $s \in \mathcal{S}$, we write $\mathrm{Reach}(s, \pi)$ the set of reachable pairs under π from s, i.e., the set of $z \in \mathcal{Z}$ such that $\mathbf{P}_s^\pi(\exists t \geq 1 : Z_t = z) > 0$.

C.1. Partioning of $\{\tau, \ldots, \tau + T - 1\}$ into optimal and sub-optimal segments

The time segment of interest $[\tau, \tau + T)$ is partioned into sub-segments $\biguplus_{i=1}^{I} [\tau_i, \tau_{i+1})$ as follows:

$$\tau_1 := \tau,$$
$$\tau_{i+1} := (\tau + T) \wedge \begin{cases} \inf\{t_k : t_k > \tau_i\} \\ \inf\{t > \tau_i : \mathbf{1}(g^{\pi_t}(S_t, M) = g^*(M)) \neq \mathbf{1}(g^{\pi_{\tau_i}}(S_{\tau_i}, M) = g^*(M))\} \end{cases}$$

and we write $i \in \mathcal{I}_{\text{opt}}$ if $g^{\pi_{\tau_i}}(S_{\tau_i}; M) = g^*(M)$ and $i \in \mathcal{I}_{\text{sub}}$ if $g^{\pi_{\tau_i}}(S_{\tau_i}; M) < g^*(M)$, that we refer to as **optimal** and **sub-optimal** segments. By design, every segment $[\tau_i, \tau_{i+1})$ is a subset of an episode and the sequence (τ_i) is a increasing sequence of stopping times. The regret is decomposed according to this partition:

$$\mathrm{Reg}(\tau, \tau + T) = \sum_{i \in \mathcal{I}_{\text{sub}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \Delta^*(Z_t) + \sum_{i \in \mathcal{I}_{\text{opt}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \Delta^*(Z_t). \tag{19}$$

Both terms are bounded separately. The first corresponds to the regret on segments where the current policy is sub-optimal, while the second corresponds to the regret on segments where the current policy is asymptotically optimal.

C.2. Upper bounding the regret on sub-optimal segments

We have:

$$\sum_{i \in \mathcal{I}_{\text{sub}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \Delta^*(Z_t) \leq \left(\max_{z \in \mathcal{Z}} \Delta^*(z) \right) \sum_{i \in \mathcal{I}_{\text{sub}}} (\tau_{i+1} - \tau_i). \tag{20}$$

We bound $\sum_{i \in \mathcal{I}_{\text{sub}}} (\tau_{i+1} - \tau_i)$ directly.

(STEP 1) *There exists a constant $\epsilon > 0$ such that, on $\bigcap_{t=\tau}^{\tau+T-1} F_t$, we have:*

$$|\mathcal{Z}|(\varphi(\tau) + 1) \geq \epsilon \sum_{i \in \mathcal{I}_{\text{sub}}} (\tau_{i+1} - \tau_i) + \sum_{i \in \mathcal{I}_{\text{sub}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \left(e_{S_{t+1}} - p(Z_t) \right) \sum_{z \in \mathcal{Z}} h^{\pi_{\tau_i}}(e_z, p) - \frac{|\mathcal{I}_{\text{sub}}|}{\epsilon} \tag{21}$$

with $\mathrm{sp}(\sum_{z \in \mathcal{Z}} h^{\pi_{\tau_i}}(e_z, p)) \leq \frac{1}{\epsilon}$, where $h^\pi(e_z, p)$ is the bias function of the policy π under the reward function e_z and kernel p. Moreover, ϵ can be chosen independently of F, τ, T and φ.

Proof Let $i \in \mathcal{I}_{\text{sub}}$ and fix $z \in \mathcal{Z}$. Because the segment $[\tau_i, \tau_{i+1})$ is a piece of episode, π_{τ_i} is used all throughout the segment. The gain and bias functions of π_{τ_i} on the model with reward function e_z (equal to one at z and null elsewhere) and kernel p are respectively denoted $g^{\pi_{\tau_i}}(-; e_z, p)$ and $h^{\pi_{\tau_i}}(-; e_z, p)$. Using the Poisson equation, we obtain:

$$\begin{aligned} N_z(\tau_{i+1}) - N_z(\tau_i) &= \sum_{t=\tau_i}^{\tau_{i+1}-1} g^{\pi_{\tau_i}}(S_t; e_z, p) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \left(e_{S_{t+1}} - p(Z_t) \right) h^{\pi_{\tau_i}}(e_z, p) \\ &\quad + \left(h^{\pi_{\tau_i}}(S_{\tau_i}; e_z, p) - h^{\pi_{\tau_i}}(S_{\tau_{i+1}}; e_z, p) \right) \\ &\geq \sum_{t=\tau_i}^{\tau_{i+1}-1} g^{\pi_{\tau_i}}(S_t; e_z, p) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \left(e_{S_{t+1}} - p(Z_t) \right) h^{\pi_{\tau_i}}(e_z, p) - \frac{1}{\epsilon} \end{aligned}$$

where ϵ is any positive quantity smaller than $(\max_\pi \max_z \mathrm{sp}(h^\pi(e_z, p)))^{-1} > 0$.

Let $\mathcal{I}^z_{\text{sub}} := \{i \in \mathcal{I}_{\text{sub}} : z \in \text{Reach}(\pi_{\tau_i}, S_{\tau_{i+1}-1})\}$.

Because the segment $[\tau_i, \tau_{i+1})$ is a piece of episode, π_{τ_i} is used all throughout the segment hence a pair that is reachable at time $\tau_{i+1} - 1$ is necessarily reachable during the entire segment. Therefore, if $i \in \mathcal{I}^z_{\text{sub}}$, then $g^{\pi_{\tau_i}}(S_t; e_z, p) > 0$ for all $t \in [\tau_i, \tau_{i+1} - 1)$. Further assume that ϵ is smaller than $\min\{g^\pi(s; e_z, p) : z \in \text{Reach}(\pi, s), s \in \mathcal{S}, \pi \in \Pi\} > 0$. We obtain:

$$N_z(\tau_{i+1}) - N_z(\tau_i) \geq \epsilon(\tau_{i+1} - \tau_i) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(e_{S_{t+1}} - p(Z_t)\big) h^{\pi_{\tau_i}}(e_z, p) - \frac{1}{\epsilon}.$$

Summing for i provides

$$\max_{i \in \mathcal{I}_{\text{sub}}} N_z(\tau_{i+1}) - N_z(\tau) \geq \epsilon \sum_{i \in \mathcal{I}^z_{\text{sub}}} (\tau_{i+1} - \tau_i) + \sum_{i \in \mathcal{I}^z_{\text{sub}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(e_{S_{t+1}} - p(Z_t)\big) h^{\pi_{\tau_i}}(e_z, p) - \frac{|\mathcal{I}^z_{\text{sub}}|}{\epsilon}.$$

Recall that for $i \in \mathcal{I}_{\text{sub}}$, the segment last until the next episode and $g^{\pi_{\tau_i}}(S_t, M) < g^*(M)$ holds for all $t \in [\tau_i, \tau_{i+1})$. Meanwhile, coherence guarantees that, on $\bigcap_{t=\tau}^{\tau+T-1} F_t$, we have $N_z(\tau_{i+1}) \leq N_z(\tau) + \varphi(\tau) + 1$ for all $i \in \mathcal{I}_{\text{sub}}$ and $z \notin \mathcal{Z}^*(M)$. So, for all $z \notin \mathcal{Z}^*(M)$ and on $\bigcap_{t=\tau}^{\tau+T-1} F_t$, we have

$$\varphi(\tau) + 1 \geq \epsilon \sum_{i \in \mathcal{I}^z_{\text{sub}}} (\tau_{i+1} - \tau_i) + \sum_{i \in \mathcal{I}^z_{\text{sub}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(e_{S_{t+1}} - p(Z_t)\big) h^{\pi_{\tau_i}}(e_z, p) - \frac{|\mathcal{I}^z_{\text{sub}}|}{\epsilon}.$$

By coherence and on $\bigcap_{t=\tau}^{\tau+T-1} F_t$ again, we see that $i \in \mathcal{I}_{\text{sub}}$ belongs to one $\mathcal{I}^z_{\text{sub}}$ for some $z \notin \mathcal{Z}^*(M)$ at least. Summing for $z \notin \mathcal{Z}^*(M)$, we obtain the claim. ■

(**STEP 2**) *There exists a constant $\eta > 0$ such that*

$$\forall x \geq 0, \quad \mathbf{P}\left(|\mathcal{I}_{\text{sub}}| \geq x + \frac{1}{\eta}\varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t \right) \leq \exp(-\eta x). \tag{22}$$

Moreover, η can be chosen independently of F, τ, T and φ.

Proof Denote $\mathcal{T}_{\text{sub}}(\tau, \tau+T) := \bigcup_{i \in \mathcal{I}_{\text{sub}}} [\tau_i, \tau_{i+1})$ the time instants when $g^{\pi_t}(S_t, M) < g^*(M)$.

Introduce the quantity $\phi(t) := \sum_z [\varphi(\tau) + N_z(\tau) - N_z(t)]_+$ for $t \in [\tau, \tau + T)$, which is non-increasing by construction. By coherence and on F_t, if $t \in \mathcal{T}_{\text{sub}}(\tau, \tau + T)$ then there exists a reachable z such that $\varphi(\tau) + N_z(\tau) - N_z(t) > 0$. The crucial remark is that for $i \in \mathcal{I}_{\text{sub}}$ with $[\tau_i, \tau_{i+1}) \subseteq [t_k, t_{k+1})$, two things may hold at time τ_{i+1}: (1) Either $i + 1 \in \mathcal{I}_{\text{opt}}$, meaning that a state from which π_{τ_i} is optimal has been reached; (2) Or $i + 1 \notin \mathcal{I}_{\text{sub}}$ and $\tau_{i+1} = t_{k+1}$, in which case $S_{\tau_{i+1}}$ has been already visited since τ_i. For (2), remark indeed that $S_{\tau_{i+1}}$ has been visited already since t_k by regenerativity of episodes (Definition 11), but if $t_k \neq \tau_i$ then $g^{\pi_{\tau_i}}(S_t, M) = g^*(M)$ for all $t \in [t_k, \tau_i)$ hence $S_{t_{k+1}}$ cannot appear within the collection of states visited in the time-range $[t_k, \tau_i)$. Combining (1) and (2), we conclude that conditionally on the history O_{τ_i}, every reachable pair $z \in \text{Reach}(\pi_{\tau_i}, S_{\tau_i})$ from which

π_{τ_i} is sub-optimal has positive probability $\epsilon(S_{\tau_i}, \pi_{\tau_i}, z, M)$ to be visited until τ_{i+1}. Letting $\epsilon := \min_{s,\pi,z} \epsilon(s, \pi, z, M) > 0$, we get:

$$\begin{aligned}
&\mathbf{P}(\phi(\tau_{i+1}) < \phi(\tau_i) | O_{\tau_i}, i \in \mathcal{I}_{\text{sub}}, F_{\tau_i}) \\
&\geq \min_{\substack{z \equiv (s,a) \in \text{Reach}(S_{\tau_i}, \pi_{\tau_i}) \\ g^{\pi_{\tau_i}}(s,M) < g^*(M)}} \mathbf{P}(N_z(\tau_{i+1}) > N_z(\tau_i) | O_{\tau_i}, i \in \mathcal{I}_{\text{sub}}, F_{\tau_i}) \\
&\geq \epsilon.
\end{aligned}$$

Let $\phi_0(\tau) := SA\varphi(\tau)$ and denote $F_{\tau:\tau+T} := \bigcap_{t=\tau}^{\tau+T-1} F_t$. On $F_{\tau:\tau+T}$, ϕ can only decrease up to $\phi_0(\tau)$ times before reaching zero, and once it has reached zero, we cannot have $t \in \mathcal{T}_{\text{sub}}(\tau, \tau + T)$ anymore. Accordingly, for all $m \geq 1$, $|\mathcal{I}_{\text{sub}}| \geq m + \phi_0(\tau)$ implies on $F_{\tau:\tau+T}$ that the first in the first $m + \phi_0(\tau)$ elements of $\mathcal{I}_{\text{sub}}$, at least m of them are such that $\phi(\tau_{i+1}) = \phi(\tau_i)$. Introduce the short-hand $U_{\tau_i} := \mathbf{1}(\phi(\tau_{i+1}) = \phi(\tau_i))$. For $\lambda > 0$ and $m \geq 1$, we have:

$$\begin{aligned}
\psi(m) &:= \mathbf{P}(|\mathcal{I}_{\text{sub}}| \geq m + \phi_0(\tau) \text{ and } F_{\tau:\tau+T}) \\
&= \mathbf{P}\left(\sum_{j=1}^{m+\phi_0(\tau)} U_{\tau_j} \geq m \text{ and } F_{\tau:\tau+T} \right) \\
&= \mathbf{E}\left[\mathbf{1}\left(\exp\left(\lambda \sum_{j=1}^{m+\phi_0(\tau)} U_{\tau_j} \right) \geq \exp(\lambda m) \right) \mathbf{1}(F_{\tau:\tau+T}) \right] \\
&\leq \exp(-\lambda m) \mathbf{E}\left[\exp\left(\lambda \sum_{j=1}^{m+\phi_0(\tau)} U_{\tau_j} \right) \mathbf{1}(F_{\tau:\tau+T}) \right] \\
&\overset{(\dagger)}{\leq} \exp(-\lambda m) \mathbf{E}\left[\exp\left(\lambda \sum_{j=1}^{m+\phi_0(\tau)-1} U_{\tau_j} \right) \cdot \begin{matrix} \mathbf{1}(F_{\tau:\tau_{m+\phi_0(\tau)}}) \cdot \mathbf{1}(F_{\tau_{m+\phi_0(\tau)}}) \\ \mathbf{E}\left[\exp(\lambda U_{\tau_{m+\phi_0(\tau)}}) \middle| F_{\tau_{m+\phi_0(\tau)}} \right] \end{matrix} \right] \\
&\overset{(\ddagger)}{\leq} \exp(-\lambda m) \mathbf{E}\left[\exp\left(\lambda \sum_{j=1}^{m+\phi_0(\tau)-1} U_{\tau_j} \right) \mathbf{1}(F_{\tau:\tau_{m+\phi_0(\tau)}}) \cdot \exp\left(\lambda(1-\epsilon) + \frac{\lambda^2}{8} \right) \right] \\
&\vdots \\
&\leq \exp\left(-\lambda m + \lambda(1-\epsilon)(m + \phi_0(\tau)) + (m + \phi_0(\tau)) \frac{\lambda^2}{8} \right).
\end{aligned}$$

In the above, $(\dagger)$ use that $\mathbf{1}(F_{\tau:\tau_{m+\phi_0}}) \cdot \mathbf{1}(F_{\tau_{m+\phi_0}}) \leq \mathbf{1}(F_{\tau:\tau+T})$ and $(\ddagger)$ is an application of Hoeffding's Lemma together with the fact that $\mathbf{E}[U_{\tau_i} | F_{\tau_i}] \mathbf{1}(F_{\tau_i}) \leq 1 - \epsilon$. Assume that m is large enough so that $\epsilon m > (1-\epsilon)\phi_0(\tau)$. Then we continue by factorizing the polynomial within the exponential and minimizing in λ, straight forward algebra shows that for $m \geq \frac{2\phi_0(\tau)}{\epsilon}$, we have:

$$\mathbf{P}(|\mathcal{I}_{\text{sub}}| \geq m + \phi_0(\tau) \text{ and } F_{\tau:\tau+T}) \leq \exp\left(-\frac{3\epsilon^2 m}{4} \right). \tag{23}$$

We conclude accordingly by choosing $\eta = \Theta(1 + \frac{2}{\epsilon})$. ■

(**STEP 3**) *There exists constants* $C_0, C_1, C_2, C_3 > 0$ *such that*

$$\forall x \geq 0, \quad \mathbf{P}\left(\sum_{i \in \mathcal{I}_{\text{sub}}} (\tau_{i+1} - \tau_i) > x + C_3 \varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t \right) \leq C_1 T^{C_2} \exp(-C_0 x). \tag{24}$$

Moreover, C_0, C_1, C_2, C_3 *can be chosen independently of* F, τ, T *and* φ.

Proof Using a time-uniform Azuma-Hoeffding's inequality (see (Bourel et al., 2020, Lemma 5)), we have for all $\delta > 0$,

$$\mathbf{P}\left(\sum_{i \in \mathcal{I}_{\text{sub}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \left(e_{S_{t+1}} - p(Z_t) \right) \sum_{z \in \mathcal{Z}} h^{\pi_{\tau_i}}(e_z, p) < -\frac{1}{\epsilon} \sqrt{\sum_{i \in \mathcal{I}_{\text{sub}}} (\tau_{i+1} - \tau_i) \log\left(\tfrac{T}{\delta}\right)} \right) \leq \delta.$$

Combined with (21) from (**STEP 1**), we obtain an equation of the form $x \leq \alpha + \beta\sqrt{x}$ with $x = \sum_{i \in \mathcal{I}_{\text{sub}}} (\tau_{i+1} - \tau_i)$, $\alpha = \frac{1}{\epsilon}(|\mathcal{Z}|(\varphi(\tau) + 1) + \frac{1}{\epsilon}|\mathcal{I}_{\text{sub}}|)$ and $\beta = \frac{1}{\epsilon}\sqrt{\log(T/\delta)}$. Simple algebra shows that $x \leq 2\alpha + 2\beta^2$. In other words, we have shown that:

$$\forall \delta > 0, \quad \mathbf{P}\left(\sum_{i \in \mathcal{I}_{\text{sub}}} (\tau_{i+1} - \tau_i) > C_0 \log\left(\tfrac{T}{\delta}\right) + C_1 \varphi(\tau) + C_2 |\mathcal{I}_{\text{sub}}| \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t \right) \leq \delta$$

for some model dependent constants $C_0, C_1, C_2 > 0$. Use the sub-exponential tail property of $|\mathcal{I}_{\text{sub}}|$ (22) from (**STEP 2**) to obtain a sub-exponential tail for $\sum_{i \in \mathcal{I}_{\text{sub}}} (\tau_{i+1} - \tau_i)$. ■

(**STEP 4**) *There exist constants* $C_0, C_1, C_2, C_3 > 0$ *such that, for all* $\eta > 0$*,*

$$\mathbf{P}\left(\sum_{j \in \mathcal{J}_{\text{sub}}^+} \sum_{t=\tau_j^+}^{\tau_{j+1}^+ - 1} \Delta^*(Z_t) > x + C_3 \varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T} F_t \right) \leq C_1 T^{C_2} \exp(-C_0 x). \tag{25}$$

Moreover, C_0, C_1, C_2, C_3 *can be chosen independently of* F, τ, T *and* φ.

Proof Combine (20) with the result of (**STEP 3**). ■

C.3. Upper bounding the regret on optimal segments

We start by merging consecutive optimal segments. This is done by setting:

$$\begin{aligned} \tau_1^+ &:= \inf\{\tau_i : i \in \mathcal{I}_{\text{opt}}\} \\ \tau_{2j}^+ &:= \inf\{\tau_i > \tau_{2j-1}^+ : i \in \mathcal{I}_{\text{sub}}\} \\ \tau_{2j+1}^+ &:= \inf\{\tau_i > \tau_{2j}^+ : i \in \mathcal{I}_{\text{opt}}\} \end{aligned} \tag{26}$$

that design a macroscopic decomposition of $[\tau, \tau + T - 1)$ into time-segments, of which (τ_i) is a refinement. Remark that if j is even, then $[\tau_j^+, \tau_{j+1}^+) \subseteq \bigcup_{i \in \mathcal{I}_{\text{sub}}} [\tau_i, \tau_{i+1})$ and conversely, if j is odd, then $[\tau_j^+, \tau_{j+1}^+) \setminus \bigcup_{i \in \mathcal{I}_{\text{sub}}} [\tau_i, \tau_{i+1}) = \emptyset$. We write $j \in \mathcal{J}_{\text{sub}}^+$ and $j \in \mathcal{J}_{\text{opt}}^+$ respectively.

By non-degeneracy of the model M, all asymptotically optimal policies of M have the same (unique) invariant probability measure $\mu^* \in \mathcal{P}(\mathcal{Z})$. On segments $[\tau_j^+, \tau_{j+1}^+)$ with $j \in \mathcal{J}_{\text{opt}}^+$, $\Delta^*(Z_t)$ can only be positive if the optimal recurrent states $\mathcal{S}(\text{supp}(\mu^*))$ have not been reached yet. The proof consists in showing that when $j \in \mathcal{J}_{\text{opt}}^+$, the optimal recurrent class is quickly reached on $[\tau_j^+, \tau_{j+1}^+)$. Indeed, setting $\tau_{j+1}^* := \tau_{j+1}^+ \wedge \inf\{t > \tau_j^+ : \mu^*(S_t) > 0\}$ the reaching time to $\text{supp}(\mu^*)$ after τ_j^+, we have:[3]

$$\begin{aligned}\sum_{j\in\mathcal{J}_{\text{opt}}^+} \sum_{t=\tau_j^+}^{\tau_{j+1}^+-1} \Delta^*(Z_t) &\le \left(\max_{z\in\mathcal{Z}} \Delta^*(z)\right) \sum_{j\in\mathcal{J}_{\text{opt}}^+} \left(N_{\mathcal{Z}^-(M)}(\tau_{j+1}^+) - N_{\mathcal{Z}^-(M)}(\tau_j^+)\right) \\ &= \left(\max_{z\in\mathcal{Z}} \Delta^*(z)\right) \sum_{j\in\mathcal{J}_{\text{opt}}^+} \left(\tau_{j+1}^* - \tau_j^+\right).\end{aligned} \tag{27}$$

We now upper bound the RHS.

(**STEP 1**) *There exists a constant $D_* > 0$ as well as an adapted sequence (h_t) with $\text{sp}(h_t) \le D^*$ s.t.:*

$$\begin{aligned}\sum_{j\in\mathcal{J}_{\text{opt}}^+} \left(\tau_{j+1}^* - \tau_j^+\right) \le 2D^* &\left(\left|\mathcal{J}_{\text{opt}}^+\right| + \sum_{j\in\mathcal{J}_{\text{opt}}^+} \left|\left\{t_\ell \in (\tau_j^+, \tau_{j+1}^+) : \mu^*(S_{t_\ell}) = 0\right\}\right| \right) \\ &+ \sum_{j\in\mathcal{J}_{\text{opt}}^+} \sum_{t=\tau_j^+}^{\tau_j^*-1} (e_{S_{t+1}} - p_{Z_t}) h_t.\end{aligned}$$

Moreover, D_* is independent of F, τ, T and φ.

Proof Notice that $[\tau_j^+, \tau_{j+1}^+)$ is of the form $[t'_k, t_{k+1}) \uplus \biguplus_\ell [t_\ell, t_{\ell+1})$ where $[t'_k \in [t_k, t_{k+1}]$ is a time such that $g^{\pi_{t-1}}(S_{t-1}; M) < g^{\pi_t}(S_t; M) = g^*(S_t; M)$. Consider the reward function $f(z) := \mathbf{1}(z \notin \mathcal{Z}^*(M))$. Over an episode $[t_\ell, t_{\ell+1}) \subseteq [\tau_j^+, \tau_{j+1}^+)$, the gain and the bias of π^ℓ associated to this reward function are respectively denoted $g^{(\ell)}$ and $h^{(\ell)}$. Because the recurrent pairs under π^ℓ from S_{t_ℓ} are $\text{supp}(\mu^*)$, we have $g^{(\ell)}(s) = 0$ for all $(s,a) \in \text{Reach}(S_{t_\ell}, \pi^\ell, M)$ and $h^{(\ell)}(s) = 0$ for all $(s,a) \in \text{supp}(\mu^*)$. Let $D^* < \infty$ be the maximum $\text{sp}(h^{(\ell)})$ possible over all $\pi^\ell \in \Pi$. Using Poisson's equation $g^{(\ell)}(s) + h^{(\ell)}(s) = f(s, \pi^\ell(s)) + p(s, \pi^\ell(s))h^{(\ell)}$, we obtain:

$$\begin{aligned}(-) &:= \tau_{j+1}^* - \tau_j^+ \\ &= N_{\mathcal{Z}^-(M)}(\tau_{j+1}^+) - N_{\mathcal{Z}^-}(\tau_j^+) \\ &= \sum_\ell \left(h^{(\ell)}(S_{t_\ell}) - h^{(\ell)}(S_{t_{\ell+1}})\right) + \sum_\ell \sum_{t=t_\ell}^{t_{\ell+1}-1} \left(e_{S_{t+1}} - p_{S_t,A_t}\right) h^{(\ell)}\end{aligned}$$

3. μ is a measure on $\mathcal{Z}$. For $s \in \mathcal{S}$, we write $\mu(s) := \sum_{a\in\mathcal{A}(s)} \mu(s,a)$.

$$
\begin{aligned}
&\leq 2D^* + \sum_{\ell : t_\ell \in (\tau_j^+, \tau_{j+1}^+)} \Big(h^{(\ell)}(S_{t_\ell}) - h^{(\ell)}(S_{t_{\ell+1}}) \Big) + \sum_{\ell} \sum_{t=t_\ell}^{t_{\ell+1}-1} \big(e_{S_{t+1}} - p_{S_t, A_t} \big) h^{(\ell)} \\
&\overset{(\dagger)}{=} 2D^* + \sum_{\ell > k} \mathbf{1}(t_\ell < \tau_j^*) \Big(h^{(\ell)}(S_{t_\ell}) - h^{(\ell)}(S_{t_{\ell+1}}) \Big) \\
&\qquad + \sum_{\ell} \sum_{t=t_\ell}^{t_{\ell+1}-1} \mathbf{1}(t < \tau_j^*) \big(e_{S_{t+1}} - p_{S_t, A_t} \big) h^{(\ell)} \\
&\overset{(\ddagger)}{\leq} 2D^* \big(1 + \big| \{ t_\ell \in (\tau_j^+, \tau_{j+1}^+) : \mu^*(S_{t_\ell}) = 0 \} \big| \big) + \sum_{t=\tau_j}^{\tau_j^*-1} \big(e_{S_{t+1}} - p_{S_t, A_t} \big) h_t
\end{aligned}
$$

where $(\dagger)$ follows from $h^{(\ell)} = 0$ on the support of μ^*, and $(\dagger)$ introduces h_t as the unique $h^{(\ell)}$ such that $t \in [t_\ell, t_{\ell+1})$. Conclude by summing over $i \in \mathcal{J}_{\mathrm{opt}}^+$. ■

(STEP 2) *There exist constants $C_1, C_2, C_3, C_4 > 0$ such that, for all $\eta > 0$,*

$$
\forall x \geq 0, \quad \mathbf{P}\left(\sum_{j \in \mathcal{J}_{\mathrm{opt}}^+} \big(\tau_{j+1}^* - \tau_j^+ \big) > x + C_4 \varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T} F_t \right) \leq C_1 T^{C_2} \exp(-C_3 x). \tag{28}
$$

Moreover, C_1, C_2, C_3, C_4 can be chosen independently of F, τ, T and φ.

Proof We bound every term appearing in (**STEP 1**).

The **first term** involves $|\mathcal{J}_{\mathrm{opt}}^+|$. Because elements of $\mathcal{J}_{\mathrm{opt}}^+$ and $\mathcal{J}_{\mathrm{sub}}^+$ are intertwined, we $|\mathcal{J}_{\mathrm{opt}}^+| \leq 1 + |\mathcal{J}_{\mathrm{sub}}^+|$. Moreover, since macroscopic segments $[\tau_j^+, \tau_{j+1}^+)$ are unions of segments $[\tau_i, \tau_{i+1})$, we have $|\mathcal{J}_{\mathrm{sub}}^+| \leq |\mathcal{I}_{\mathrm{sub}}|$ that has been bounded in (22) already. Accordingly, $|\mathcal{J}_{\mathrm{sub}}^+|$ has sub-exponential tails on the good event $\bigcap_{t=\tau}^{\tau+T-1} F_t$:

$$
\forall x \geq 0, \quad \mathbf{P}\left(|\mathcal{J}_{\mathrm{sub}}^+| \geq x + \frac{1}{c} \varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t \right) \leq \exp(-cx) \tag{29}
$$

where $c > 0$ is a model dependent constant.

For the **second term**, remark that for each $t_\ell \in [\tau_j, \tau_{j+1})$ with $j \in \mathcal{J}_{\mathrm{opt}}^+$, the probability that the episode ends with $S_{t_{\ell+1}} \in \mathrm{supp}(\mu^*)$ is positive because of the regenerativity property (Definition 11) of (VM). This is also true for the first (possibly) truncated episode $[t'_k, t_{k+1})$ that starts the macroscopic segment $[\tau_j^+, \tau_{j+1}^+)$ because as the gain $g^{\pi_t}(S_t; M)$ increases from $t'_k - 1$ to t'_k to the optimal $g^{\pi_t}(S_t, M) = g^*(S_t; M)$, all states that are reachable from S_t under π_t cannot have been visited yet during the episode. In the end, the probability of reaching $\mathrm{supp}(\mu^*)$ by the end of the episode is lower bounded by some $\epsilon'(\pi_{t_\ell}, S_{t_\ell}, M) > 0$ and we denote $\epsilon' > 0$ the minimum for all possible values of π^ℓ and S_{t_ℓ}. We conclude that $\mathbf{P}(\mu^*(S_{t_{\ell+1}}) > 0 \mid O_{t_\ell}) > \epsilon'$. Accordingly,

$$
U_{\tau_j^+} := \big| \{ t_\ell \in (\tau_j^+, \tau_{j+1}^+) : \mu^*(X_{t_\ell}) = 0 \} \big|
$$

is stochastically dominated by a geometric distribution $\mathrm{G}(\epsilon')$. Using bounds on tails of geometric random variables (Lemma 25), we obtain:

$$\mathbf{P}\left(\sum_{j\in\mathcal{J}^+_{\text{opt}}} \left|\left\{t_\ell \in (\tau_j^+, \tau_{j+1}^+) : \mu^*(S_{t_\ell}) = 0\right\}\right| > \left|\mathcal{J}^+_{\text{opt}}\right|\left(1+\tfrac{2}{\epsilon'}\right) + \frac{2\eta\log(T)}{\log\left(\frac{1}{1-\epsilon'}\right)}\right) \leq T^{-\eta}. \quad (30)$$

The **third term** $\sum_{j\in\mathcal{J}^+_{\text{opt}}}\sum_{t=\tau_j}^{\tau^*_{j+1}-1}(e_{S_{t+1}} - p_{Z_t})h_t$ is the sum of a martingale difference sequence, each term having span at most D^* by (**STEP 1**). By applying a time-uniform Azuma-Hoeffding's inequality (see (Bourel et al., 2020, Lemma 5)), we obtain:

$$\mathbf{P}\left(\sum_{j\in\mathcal{J}^+_{\text{opt}}}\sum_{t=\tau_j}^{\tau^*_{j+1}-1}(e_{S_{t+1}} - p_{Z_t})h_t > D^*\sqrt{\sum_{j\in\mathcal{J}^+_{\text{opt}}}\left(\tau^*_{j+1}-\tau^+_j\right)\left(\tfrac{1}{2}+\eta\right)\log(1+T)}\right) \leq T^{-\eta}. \quad (31)$$

Combining the bound of the first term, (30) and (31), we see that there exists C_1, C_2, C_3, C_4 such that for all $\eta > 0$, with probability $1 - 3T^{-\eta}$,

$$\begin{aligned}&\sum_{j\in\mathcal{J}^+_{\text{opt}}}\left(\tau^*_{j+1}-\tau^+_j\right)\\ &\leq C_1 + (C_2+\eta C_3)(\log(T)+\varphi(\tau)) + C_4\sqrt{\sum_{j\in\mathcal{J}^+_{\text{opt}}}\left(\tau^*_{j+1}-\tau^+_j\right)\left(\tfrac{1}{2}+\eta\right)\log(T)}.\end{aligned}$$

This is an equation of the form $x \leq \alpha + \beta\sqrt{x}$ that implies in particular $x \leq 2(\alpha+\beta^2)$. We conclude by rearranging terms of the equation. ■

(**STEP 3**) *There exist constants $C_1, C_2, C_3, C_4 > 0$ such that, for all $\eta > 0$,*

$$\mathbf{P}\left(\sum_{j\in\mathcal{J}^+_{\text{opt}}}\sum_{t=\tau_j^+}^{\tau^+_{j+1}-1}\Delta^*(Z_t) > x + C_4\varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T} F_t\right) \leq C_1 T^{C_2}\exp(-C_3 x). \quad (32)$$

Moreover, C_1, C_2, C_3, C_4 can be chosen independently of F, τ, T and φ.

Proof Invoke (27) and apply the result of (**STEP 2**). ■

C.4. Combining everything

Conclude by combining (19) with Section C.2 (**STEP 4**) and Section C.3 (**STEP 3**).

C.5. Technical result: Tails of geometric random variables

In this section, we provide a result on the tails of sums of geometric random variables.

Lemma 25 (Tails of Geometric Random Variables) *Let (X_i) be a sequence of i.i.d. random variable of distribution $\mathrm{G}(p)$, and let $S_n := X_1 + \ldots + X_n$ be their sum. Then, for all $c \geq 2$ and $t \geq 0$,*

$$\mathbf{P}(S_n \geq c(\mathbf{E}[S_n] + t)) \leq (1-p)^t \exp\Big(-\tfrac{(2c-3)n}{4}\Big).$$

Proof This proof is standard and was found on `math.stackexchange.com`[4]. We rely on Chernoff's method as usual by using the Laplace transform $\mathbf{E}[e^{sX_i}]$. We have:

$$\mathbf{P}(S_n \geq c(\mathbf{E}[S_n] + t)) \leq e^{-sct} e^{-scn/p} \prod_{i=1}^{n} \mathbf{E}[e^{sX_i}].$$

We compute the Laplace transform of X_i: $\mathbf{E}[e^{sX_i}] = (1 - \frac{1-e^s}{p})^{-1}$. Setting $s = -\frac{1}{c}\log(1-p)$, we have $\exp(-sc) = 1-p$. In the above formula, we readily obtain:

$$\mathbf{P}(S_n \geq c(\mathbf{E}[S_n] + t)) \leq (1-p)^t \exp\Big(n\Big(\tfrac{a}{p} - \log\Big(1 - \tfrac{b}{p}\Big)\Big)\Big)$$

where $a := \log(1-p)$ and $b = 1 - (1-p)^{1/c}$. We want that exponential to decrease quickly to 0 with n, i.e., we want $a/p - \log(1 - b/p) < 0$. By Bernoulli's inequality, we have $b = 1 - (1-p)^{1/c} \leq p/c \leq p/2$, hence $b/p \leq \frac{1}{2}$. Moreover, for $z \in (0, \frac{1}{2}]$, $\log(1-z) \geq -z - z^2$, hence

$$\tfrac{a}{p} - \log\Big(1 - \tfrac{b}{p}\Big) \leq \tfrac{a}{p}\tfrac{b}{p} - \log(\tfrac{1}{2}) \leq \tfrac{1}{2}\big(\tfrac{a}{b} + \tfrac{3}{2}\big).$$

Finally, since $a = \log(1-p)$ and $b \leq p/c$, it follows that $\frac{a}{b} \leq \frac{c\log(1-p)}{p} \leq -c$, so we get $\frac{a}{b} - \log(1 - \frac{b}{p}) \leq \frac{1}{2}(-c + \frac{3}{2})$. This concludes the proof. ■

4. https://math.stackexchange.com/questions/110691/tail-bound-on-the-sum-of-independent-non-identical-geometric-random-variables

Appendix D. Regret of exploration guarantees of (VM)

In this appendix, we provide the details of Section 4 and behind the proof of Theorem 7, assertion 3. Below is a map of the proof. It will be reported throughout the proof to keep track of where the current lemma of interest in used in the proof's architecture.

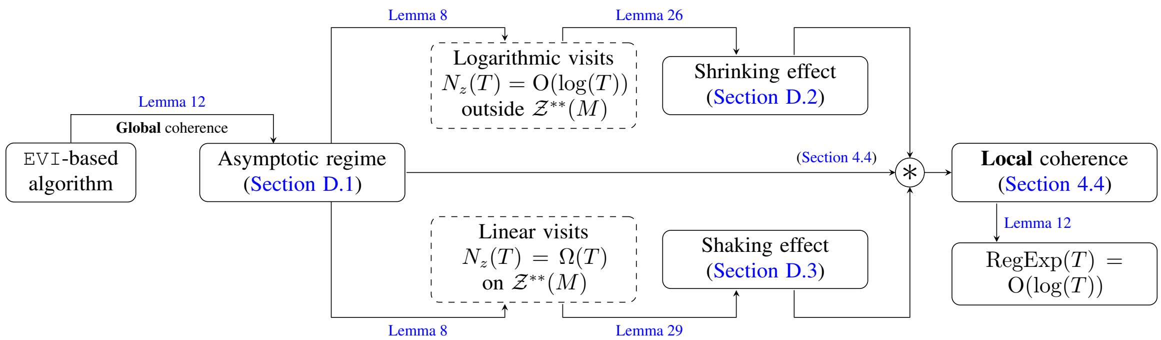


Figure 9: Proof map of regret of exploration guarantees.

Outline The appendix is organized as such. In Section D.1 we prove Lemma 8, describing the asymptotic visit rates. We continue by establishing formal version of the shrinking-shaking effect, discussed informally in Section 4.2, beginning with the shrinking effect in Section D.2 and continuing with the shaking effect in Section D.3. We conclude by linking the shrinking-shaking effect and the asymptotic visit rates to a local coherence property of Lemma 13 in Section D.4, which is the last step of the proof of the assertion 3 of Theorem 7, see Section 4.4.

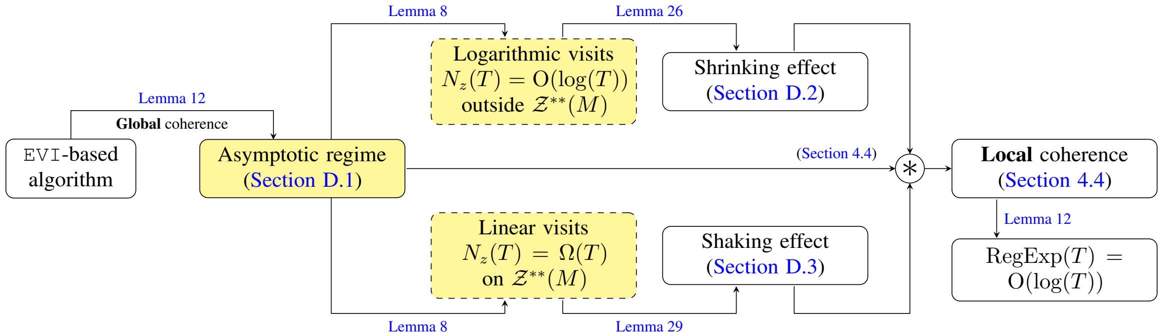
D.1. The asymptotic regime of (VM): Proof of Lemma 8

In this section, we provide a proof of Lemma 8:

Lemma 8 *Let $M \in \mathcal{M}$ be non-degenerate satisfying Assumption 4. Assume that* KLUCRL *is run while managing episodes with f-*(VM) *with arbitrary $f > 0$. There exists $\lambda > 0$ such that:*

$$\begin{aligned} &\forall z \notin \mathcal{Z}^{**}(M), \quad \mathbf{P}^M(\exists T, \forall t \geq T : N_z(t) < \lambda \log(t)) = 1, \textit{ and} \\ &\forall z \in \mathcal{Z}^{**}(M), \quad \mathbf{P}^M\left(\exists T, \forall t \geq T : N_z(t) > \tfrac{1}{\lambda} t\right) = 1. \end{aligned}$$

Outline of the proof The proof relies on the coherence lemma (Lemma 12). We show that the confidence regions are such that, if a sub-optimal policy is played, one of the pairs responsible for its optimistic gain must be sub-sampled. This provides a "global" coherence property, see (**STEP 1**), this is used show that sub-optimal pairs are visited at most logarithmically often in the asymptotic regime, see (**STEP 2**) and (36). However, the coherence lemma (Lemma 18) cannot be invoked unless Assumption 4 holds and M is non-degenerate. We deduce in parallel that non optimal pairs, i.e., $\mathcal{Z} \setminus \mathcal{Z}^{**}(M)$, are visited at most logarithmically often in (**STEP 3**) with (38), and that optimal pairs, i.e., $\mathcal{Z}^{**}(M)$, are visited at least linearly often in (**STEP 4**) with (39).



(STEP 1) *There exists a sequence of adapted events (F_t) satisfying $\mathbf{P}(\exists T, \forall t \geq T : F_t) = 1$ and a function $\varphi : \mathbf{N} \to \mathbf{R}_+$ with $\varphi(t) = \mathrm{O}(t)$ s.t. the algorithm is $((F_t), \lfloor \log(T) \rfloor, T, \varphi)$-coherent.*

Proof Introduce the good event $E_t := \{M \in \mathcal{M}(t)\}$. By design of the confidence region (see Lemma 18), we know that $\mathbf{P}(\exists t \geq T : M \notin \mathcal{M}(t)) = \mathrm{O}(T^{-1})$, so $\mathbf{P}(\exists T, \forall t \geq T : E_t) = 1$. Let $T \geq 1$ and set $T_0 := \lfloor \log(T) \rfloor$. Pick $t \in \{T_0, \ldots, T\}$ and let $\{t_k, \ldots, t_{k+1} - 1\}$ be the unique episode it falls in. We denote $\pi \equiv \pi_{t_k}$ for short and assume that π is sub-optimal from S_t, i.e.,

$$g^*(S_t; M) > g^\pi(S_t; M). \tag{33}$$

So there exists a class of pairs $\mathcal{Z}'$ which is recurrent under π, with $\mathcal{Z}' \subseteq \mathrm{Reach}(\pi, S_t)$ and such that $g^\pi(s; M) < g^*(s; M)$ for every $s \in \mathcal{S}(\mathcal{Z}')$. Let $s_0 \in \mathcal{S}(\mathcal{Z}')$.

Denote $\Delta_g := \min\{g^*(s; M) - g^\pi(s; M) : \pi \in \Pi, s \in \mathcal{S}, g^*(s; , M) > g^\pi(s; M)\} > 0$ the minimal gain gap in M. Because π is output by EVI (Section B.1) at time t_k, it is optimistically optimal at time t_k and $g^*(S_t; \mathcal{M}(t_k)) = g(S_t; \tilde{r}_\pi, \tilde{p}_\pi)$ for some $\tilde{r}_\pi \in \prod_s \mathcal{R}_{s,\pi(s)}(t_k)$ and $\tilde{p}_\pi \in \prod_s \mathcal{P}_{s,\pi(s)}(t_k)$. Furthermore, on E_t, we have $D(\mathcal{M}(t)) \leq D(M)$ hence every policy returned by EVI (Section B.1) has optimistic bias span at most $D(M)$ and its optimistic gain has span equal to 0. We have, on E_{t_k},

$$\begin{aligned}
\Delta_g &\leq g^*(s_0; M) - g^\pi(s_0; M) \\
&\overset{(\dagger)}{\leq} g^{\pi_{t_k}}(s_0; \mathcal{M}(t_k)) - g^\pi(s_0; M) \\
&\overset{(\ddagger)}{\leq} \|\tilde{r} - r\|_{\infty, \mathrm{Reach}(\pi, s_0)} + \tfrac{1}{2} D(M) \|\tilde{p} - p\|_{1, \mathrm{Reach}(\pi, s_0)}.
\end{aligned}$$

In $(\dagger)$, we have used that, $g^*(s_0; M) \leq g^*(s_0; \mathcal{M}(t_k)) = g^*(S_{t_k}; \mathcal{M}(t_k)) = g^\pi(S_{t_k}; \mathcal{M}(t_k))$ on E_{t_k}. In $(\ddagger)$, we first invoke a gain deviation inequality (Lemma 32), then rely on the fact that by Assumption 4, the optimistic gain of π computed by EVI only depends on pairs that are reachable from s_0 under π on M. One of the two terms of the RHS of the above equation must be at least $\frac{1}{2}\Delta_g$. For instance, $D(M)\|\tilde{p} - p\|_{1, \mathrm{Reach}(\pi, s_0)} \geq \Delta_g$. We have:

$$\begin{aligned}
\Delta_g &\leq D(M) \Big(\|\tilde{p} - \hat{p}_{t_k}\|_{1, \mathrm{Reach}(\pi, s_0)} + \|\hat{p}_{t_k} - p\|_{1, \mathrm{Reach}(\pi, s_0)} \Big) \\
&= D(M) \bigg(\min_{z \in \mathrm{Reach}(\pi, s_0)} \|\tilde{p}(z) - \hat{p}_{t_k}(z)\|_1 + \min_{z \in \mathrm{Reach}(\pi, s_0)} \|\hat{p}_{t_k}(z) - p(z)\|_1 \bigg).
\end{aligned} \tag{34}$$

Now, given $\tilde{p} \in \mathcal{P}_z(t)$, we have $N_z(t)\mathrm{KL}(\hat{p}_t(z)||\tilde{p}(z)) \leq |\mathcal{S}| \log(2et)$, see (5). By Pinsker's inequality, we deduce that there are constants $\alpha, \beta > 0$ (independent of $t \geq 1$, $z \in \mathcal{Z}$ and $\tilde{p} \in \mathcal{P}_z(t)$) such that $N_z(t)\|\tilde{p}(z) - \hat{p}_t(z)\|_1^2 \leq \alpha \log(\beta t)$. Accordingly,

$$\mathcal{P}_z(t) \subseteq \left\{ \tilde{p}_z \in \mathcal{P}(\mathcal{S}) : N_z(t)\|\tilde{p}_z - \hat{p}_z(t)\|_1^2 \leq \alpha \log(\beta t) \right\} =: \mathcal{P}'_z(t).$$

Since $\mathbf{P}(\exists T \geq 1 : p(z) \in \mathcal{P}_z(t)) = 1$, we deduce that $\mathbf{P}(\exists T \geq 1 : p(z) \in \mathcal{P}'_z(t)) = 1$. Injecting this in (34), we see that on the asymptotically almost sure event $F^p_{t_k} := \{\forall z, p_z \in \mathcal{P}'_z(t)\}$, we have:

$$\Delta_g \leq 2D(M) \min_{z \in \mathrm{Reach}(\pi, s_0)} \sqrt{\frac{\alpha \log(\beta t_k)}{N_z(t_k)}} \overset{(\dagger)}{\leq} 2D(M) \min_{z \in \mathrm{Reach}(\pi, s_0)} \sqrt{\frac{2\alpha \log(2\beta t)}{N_z(t)}} \tag{35}$$

where $(\dagger)$ uses that the (VM) guarantees $N_z(t_{k+1}) \leq 2N_z(t_k)$ and $t_{k+1} \leq 2t_k$. Solving (35) in $N_z(t)$, we find a condition of the form $N_z(t) \leq \alpha' \log(\beta' t)$.

The same rationale can be used to handle the case where $\|\tilde{r} - r\|_{\infty, \mathrm{Reach}(\pi, s_0)} \geq \frac{1}{2}\Delta_g$, dealing with the design of another asymptotically almost sure event $F^r_t := \{\forall z, r_z \in \mathcal{R}'_z(t)\}$ and ending with the same kind of upper-bound on $N_z(t)$. In the end, setting $F_t := \bigcap_{t'=\lfloor t/2 \rfloor}^{t} F^r_{t'} \cap F^p_{t'} \cap E_{t'}$ and $\varphi(T_0) = \alpha' \log(\beta' t)$, we see that the algorithm is $((F_t), T_0, T, \varphi)$-coherent. ■

(**STEP 2**) *There exists* $C > 0$ *such that:*

$$\mathbf{P}\big(\exists T, \forall t \geq T, \forall z \in \mathcal{Z}^-(M) : N_z(t) \leq C \log(t)\big) = 1. \tag{36}$$

Proof Since M is non-degenerate, coherence can be converted to regret guarantees (Lemma 12): Applying Lemma 12 following (**STEP 1**), there exist constants $C_1, C_2 > 0$ such that:

$$\forall T \geq 1, \quad \mathbf{P}\left(\mathrm{Reg}(\log(T), T) \geq C_1 + C_2 \log(T) \text{ and } \bigcap_{t=\lfloor \log(T) \rfloor}^{T} F_t \right) \leq T^{-2}. \tag{37}$$

Since $N_z(T) \leq N_z(T_0) + \Delta^*(z)^{-1}\mathrm{Reg}(T_0, T)$, the condition $\mathrm{Reg}(\log(T), T) \leq C_1 + C_2 \log(T)$ is converted to $N_z(T) \leq C'_1 + C'_2 \log(T)$ for all $z \in \mathcal{Z}^-(M)$. We have:

$$\begin{aligned}
&\mathbf{P}\big(\forall T, \exists t \geq T, \exists z \in \mathcal{Z}^-(M) : N_z(t) > C'_1 + C'_2 \log(t)\big) \\
&\overset{(\dagger)}{=} \mathbf{P}\left(\forall T, \exists t \geq T, \exists z \in \mathcal{Z}^-(M) : N_z(t) > C'_1 + C'_2 \log(t) \text{ and } \bigcap_{t=\lfloor \log(T) \rfloor}^{T} F_t \right) \\
&\leq \mathbf{P}\left(\forall T, \exists t \geq T, \exists z \in \mathcal{Z}^-(M) : \mathrm{Reg}(\log(T), T) > C_1 + C_2 \log(T) \text{ and } \bigcap_{t=\lfloor \log(T) \rfloor}^{T} F_t \right) \\
&= \lim_{T \to \infty} \sum_{t \geq T} \sum_{z \in \mathcal{Z}^-(M)} \mathbf{P}\left(\mathrm{Reg}(\log(T), T) > C_1 + C_2 \log(T) \text{ and } \bigcap_{t=\lfloor \log(T) \rfloor}^{T} F_t \right)
\end{aligned}$$

$$\overset{(\ddagger)}{\leq} SA \lim_{T\to\infty} \tfrac{1}{T} = 0.$$

In the above, $(\dagger)$ follows by $\mathbf{P}(\limsup F_t) = 1$ and $(\ddagger)$ by (37). Up to assuming t large enough, we eventually have $C_2' \log(T) \geq C_1'$ hence the constant term can be ignored. ■

(STEP 3) *There exists* $C > 0$ *such that:*

$$\mathbf{P}(\exists T, \forall t \geq T, \forall z \notin \mathcal{Z}^{**}(M) : N_z(t) \leq C\log(t)). \tag{38}$$

Proof Because M is non-degenerate, $\mathcal{Z}^*(M)$ defines a unique policy that we denote π^*, given by $\pi(s) = a$ where $a \in \mathcal{A}(s)$ is the unique action such that $(s, a) \in \mathcal{Z}^*(M)$.

Introduce the reward function $f(z) := \mathbf{1}(z \notin \mathcal{Z}^{**}(M))$. Let g^f, h^f and Δ^f be the gain, bias and gap functions of π^* in M endowed with the reward function f. Remark that $g^f(s) = 0$, that $h^f(s) = 0$ for $(s, \pi^*(s)) \in \mathcal{Z}^{**}(M)$ and that $\Delta^f(z) = 0$ for $z \in \mathcal{Z}^*(M)$. Denote $H_f := \max\{\mathrm{sp}(h^f), \max_z |\Delta^f(z)|\}$. Therefore:

$$\begin{aligned}
\sum_{z\notin\mathcal{Z}^{**}(M)} N_z(T) &= \sum_{t=1}^{T} f(Z_t) \\
&= \sum_{t=1}^{T} \big((e_{S_t} - p(Z_t))h^f - \Delta_f(Z_t)\big) \\
&\leq H^f + \sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathcal{Z}^{**}(M))\big(e_{S_{t+1}} - p(Z_t)\big)h^f + H^f \sum_{z\in\mathcal{Z}^-(M)} N_z(T) \\
&\overset{(\dagger)}{\leq} H^f \left(1 + 2\sqrt{\sum_{z\notin\mathcal{Z}^{**}(M)} N_z(T)\log(T)} + \sum_{z\in\mathcal{Z}^-(M)} N_z(T)\right) \\
&\overset{(\ddagger)}{\leq} H^f \left(1 + 2\sqrt{\sum_{z\notin\mathcal{Z}^{**}(M)} N_z(T)\log(T)} + SAC\log(T)\right)
\end{aligned}$$

where $(\dagger)$ holds with probability $1 - T^{-2}$ by Azuma-Hoeffding's inequality (see (Bourel et al., 2020, Lemma 5)), and $(\dagger)$ holds on the asymptotically almost sure event $(\forall z \in \mathcal{Z}^-(M), N_z(T) \leq C\log(T))$ (see (36)). This is an equation of the form $n \leq \alpha + \beta\sqrt{n}$ that implies in particular $n \leq 2(\alpha + \beta^2)$. In the end, we get:

$$\mathbf{P}\left(\forall T, \exists t \geq T : \sum_{z\notin\mathcal{Z}^{**}(M)} N_z(t) \leq 2H^f(1 + SAC\log(T) + 4\log(T))\right) = 1.$$

This concludes the proof. ■

(STEP 4) *There exists* $c > 0$ *such that:*

$$\mathbf{P}(\exists T, \forall t \geq T, \forall z \in \mathcal{Z}^{**}(M) : N_z(t) \geq ct) = 1. \tag{39}$$

Proof This is established with a similar technique than (38) in (**STEP 3**). By non-degeneracy of M, $\mathcal{Z}^*(M)$ defines a unique policy that we denote π^*. Fix $z_0 \in \mathcal{Z}^{**}(M)$ and introduce the reward function $f(z) = \mathbf{1}(z = z_0)$. Remark that $g^f(s) = c > 0$ for all $s \in \mathcal{S}$ and that $\Delta^f(z) = 0$ for all $z \in \mathcal{Z}^*(M)$. Let $H^f := \mathrm{sp}(h^f) \vee \max_z |\Delta^f(z)|$. We have:

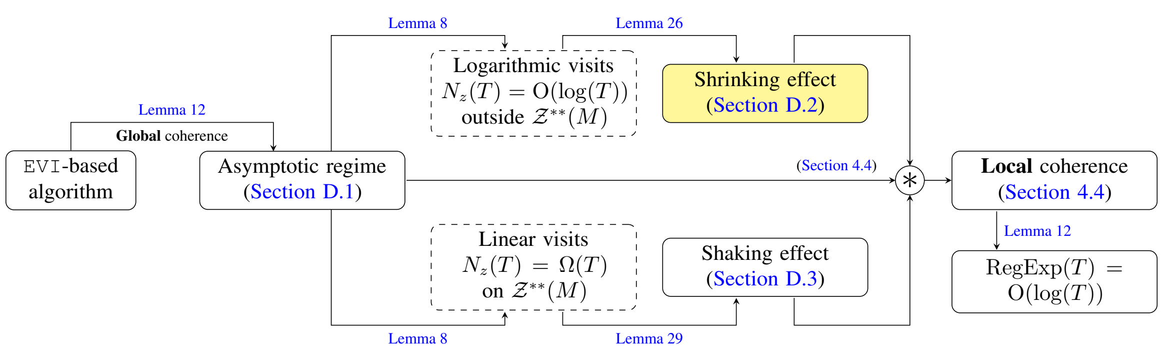
$$
\begin{aligned}
N_{z_0}(T) &:= \sum_{t=1}^{T} f(Z_t) \\
&= cT + \sum_{t=1}^{T} \big((e_{S_t} - p(Z_t)) h^f - \Delta_f(Z_t) \big) \\
&\geq cT - \sum_{t=1}^{T} \mathbf{1}(Z_t \in \mathcal{Z}^-(M)) \big(e_{S_{t+1}} - p(Z_t) \big) h^f - H^f \sum_{z \in \mathcal{Z}^-(M)} N_z(T) \\
&\geq cT - 2\sqrt{H^f SAC} \cdot \log(T) - H^f SAC \log(T) \sim cT
\end{aligned}
$$

where the last inequality holds with probability $1 - T^{-2}$ on the asymptotically almost sure event $(\forall z \in \mathcal{Z}^-(M) : N_z(T) \leq C \log(T))$ given by (36). We conclude accordingly. ■

About Assumption 4 In the coherence property, the first statement, which is about the reachability of sub-sampled pairs, is not guaranteed to hold if we run `KLUCRL` on an arbitrary model. The issue lies in the fact that the high optimistic gain of a policy may be due states that are unreachable under the optimistically optimal policy. This is because in the confidence region $\mathcal{M}(t)$, there may be models with a richer transition structure than the true hidden model M. This is where Assumption 4 seems necessary. Assumption 4 is roughly equivalent to stating that the support of the transitions of M are known in advance. We conjecture that this assumption cannot be removed without a significant rework of `EVI`. Under Assumption 4, the optimistic gain of a policy π from a fixed state s only depends on $\mathcal{R}_z(t), \mathcal{P}_z(t)$ for pairs z that are reachable from s under π on M. This echoes the reachability requirement of sub-sampled pairs.

D.2. The shrinking effect: Formal version of Informal Property 9

In this section, we provide a proof of a formalized version of the **shrinking effect** part of Informal Property 9.



In Lemma 26 below, we show that if $N_z(t) = \mathrm{O}(\log(t))$ and under a good event, the kernel confidence region $\mathcal{P}_z(t)$ remains confined in the confidence region $\mathcal{P}_z(t_{k(i)-1})$ at time $t_{k(i)-1}$, the beginning of the previous exploitation episode (when the current policy is gain optimal). For rewards, the shrinking effect is shown strict by quantifying its speed. The shrinking speed is shown to be faster than any $(\frac{1}{t})^\eta$ for $\eta > 0$. This will be essential later, so that the shrinking effect on non-optimal pairs completely dominates the shaking effect on optimal pairs.

Lemma 26 *Let $(t_{k(i)})$ be the enumeration of exploration episodes, and let $T \geq 1$. Fix $\lambda > 0$ and $z \in \mathcal{Z}$. For all $\delta, \eta > 0$, we can find $\epsilon, m, C > 0$ such that:*

$$\mathbf{P}\left(\exists t \in \left\{ t_{k(i)}, \dots t_{k(i)} + T \right\} : \begin{array}{c} \mathcal{P}_z(t) \not\subseteq \mathcal{P}_z(t_{k(i)-1}) \\ \text{and } F_t \text{ and } N_z(t) > N_z(t_{k(i)}) + C\log\left(\frac{T}{\delta}\right) \end{array} \right) \leq \delta,$$

$$\mathbf{P}\left(\exists t \in \left\{ t_{k(i)}, \dots, t_{k(i)} + T \right\} : \begin{array}{c} \sup \mathcal{R}_z(t) > \sup \mathcal{R}_z(t_{k(i)-1}) - \frac{N_z(t) - N_z(t_{k(i)})}{C \cdot (t_{k(i)})^\eta} \\ \text{and } F_t \text{ and } N_z(t) > N_z(t_{k(i)}) + C\log\left(\frac{T}{\delta}\right) \end{array} \right) \leq \delta$$

with $F_t := (N_z(t) < \frac{1}{\lambda}\log(t), \mathrm{KL}(\hat{q}_t(z)||q(z)) < \epsilon, t > m)$ where $q(z) \equiv (r(z), p(z))$.

D.2.1. A "LARGE" SHRINKING EFFECT FOR KERNELS

We beginning with a proof of the shrinking effect for the confidence regions of kernels. The shrinking is shown large, in the sense that we show a property of the form "$\mathcal{P}_z(t) \subseteq \mathcal{P}_z(t_{k(i)-1})$" but do not quantify how smaller than $\mathcal{P}_z(t_{k(i)-1})$ the region $p_z(t)$ is subjected to be.

Lemma 27 (Shrinking effect, kernels) *Let $(t_{k(i)})$ be the enumeration of exploration episodes, and let $T \geq 1$. Fix $\lambda > 0$ and $z \in \mathcal{Z}$. For all $\delta > 0$, we can find $\epsilon, m, C > 0$ such that:*

$$\mathbf{P}\left(\exists t \in \left\{ t_{k(i)}, \dots t_{k(i)} + T \right\} : \begin{array}{c} \mathcal{P}_z(t) \not\subseteq \mathcal{P}_z(t_{k(i)-1}) \\ \text{and } F_{t_{k(i)-1}} \cap F_t \text{ and } N_z(t) > N_z(t_{k(i)}) + C\log\left(\frac{T}{\delta}\right) \end{array} \right) \leq \delta$$

with $F_t := (N_z(t) < \frac{1}{\lambda}\log(t), \mathrm{KL}(\hat{q}_t(z)||q(z)) < \epsilon, t > m)$ where $q(z) \equiv (r(z), p(z))$.

Proof We write $N_z(t, t') := N_z(t') - N_z(t)$ the number of times $z \in \mathcal{Z}$ is visited between the times t and t'. We write $w_{t_{k(i)-1},t}(z) := \hat{p}_t(z) - \hat{p}_{t_{k(i)-1}}(z)$ the change of kernel from time $t_{k(i)-1}$ to t for the pair $z \in \mathcal{Z}$. Fix $\epsilon, \lambda, m > 0$, $z \in \mathcal{Z}$ and introduce the event:

$$F_t \equiv F_t^{(\epsilon,\lambda,m)} := \left(N_z(t) < \frac{1}{\lambda}\log(t), \mathrm{KL}(\hat{q}_t(z)||q(z)) < \epsilon, t > m \right) \tag{40}$$

(STEP 1) *There exists a function $\lambda \mapsto m_\lambda \in \mathbf{N}$ such that, for $m \geq m_\lambda$, we have:*

$$\mathbf{P}\left(\exists t \in [t_{k(i)}, t_{k(i)} + T] : F_{t_{k(i)-1}}, \|w_{t_{k(i)-1},t}(z)\|_1 > \frac{2 + \epsilon^2 N_z(t_{k(i)}, t) + \sqrt{|\mathcal{S}| N_z(t_{k(i)}, t)\log\left(\frac{T}{\delta}\right)}}{N_z(t)} \right) \leq \delta$$

Proof With straight-forward algebra, we check that $w_{t_{k(i)-1},t}(z)$ is equal to

$$\frac{1}{N_z(t)}\left(N_z(t_{k(i)-1},t)\Big(p(z)-\hat{p}_{t_{k(i)-1}}(z)\Big)+\sum_{i=t_{k(i)-1}}^{t-1}\mathbf{1}(Z_i=z)\big(e_{S_{i+1}}-p(z)\big)\right). \tag{41}$$

On the $F_{t_{k(i)-1}}$, we know that $\mathrm{KL}(\hat{p}_{t_{k(i)-1}}(z)||p(z))<\epsilon$, so by Pinsker's inequality, follows $\|p(z)-\hat{p}_{t_{k(i)-1}}(z)\|_1 \leq \epsilon^2$. So $\|w_{t_{k(i)-1},t}(z)\|_1 \leq \frac{1}{N_z(t)}(N_z(t_{k(i)-1},t)\epsilon^2+\|\sum_i \mathbf{1}(Z_i = z)(e_{S_{i+1}}-p(z))\|_1)$, consisting in two terms. The first term is an error a priori, while the second is the norm of a martingale which is the sum of $N_z(t_{k(i)-1},t)$ terms. On $F_{t_{k(i)-1}}$, we have:

$$\begin{aligned}
N_z(t_{k(i)}) &\leq \lfloor\big(1+f(t_{k(i)-1})\big)N_z(t_{k(i)-1}))\rfloor+1\\
&\overset{(\dagger)}{\leq} N_z(t_{k(i)-1})+1+\left\lfloor\tfrac{1}{\lambda}f(t_{k(i)-1})\log(t_{k(i)-1})\right\rfloor\\
&\overset{(\ddagger)}{=} N_z(t_{k(i)-1})+1
\end{aligned}$$

where $(\dagger)$ is by definition on $F_{t_{k(i)-1}}$ and $(\ddagger)$ holds for $t\to\infty$ since $f(t)=\mathrm{o}(\log(t)^{-1})$, hence provided that $t_{k(i)-1}\geq \frac{1}{2}t_{k(i)}\geq\frac{1}{2}m$ is large enough with respect to λ, e.g., $m\geq m_\lambda\in\mathbf{N}$. Accordingly, we have $N_z(t_{k(i)-1},t_{k(i)})\leq 1$ on $F_{t_{k(i)-1}}$. So, on $F_{t_{k(i)-1}}$, we have:

$$\left\|w_z(t_{k(i)-1},t)\right\|_1 \leq \frac{1}{N_z(t)}\left(2+N_z(t_{k(i)},t)\epsilon^2+\left\|\sum_{i=t_{k(i)}}^{t-1}\mathbf{1}(Z_i=z)\big(e_{S_{i+1}}-p(z)\big)\right\|_1\right).$$

Applying Weissman's inequality (see Weissman et al. (2003) or (Auer et al., 2009, Equation (44)) or Lemma 21), the martingale can then be bounded as follows:

$$\mathbf{P}\left(\exists t\in[t_{k(i)},t_{k(i)}+T],\ \left\|\sum_{i=t_{k(i)}}^{t-1}\mathbf{1}(Z_i=z)\big(e_{S_{i+1}}-p(z)\big)\right\|_1\geq\sqrt{|\mathcal{S}|N_z(t_{k(i)},t)\log\left(\tfrac{T}{\delta}\right)}\right)$$
$$\leq\delta.$$

We conclude accordingly. ■

(STEP 2) *Assume that* $\epsilon<(\frac{|\mathcal{S}|\log(T)}{T})^{1/4}$ *and* $m\geq m_\lambda$. *Then, for all* $\delta>0$, *we have:*

$$\mathbf{P}\left(\exists t\in[t_{k(i)},t_{k(i)}+T]: F_{t_{k(i)-1}},\|w_{t_{k(i)-1},t}(z)\|_1>\frac{2\Big(1+\sqrt{|\mathcal{S}|N_z(t_{k(i)},t)\log\left(\frac{T}{\delta}\right)}\Big)}{N_z(t)}\right)\leq\delta$$

Proof We know that for $t\in\{t_{k(i},\ldots,t_{k(i)}+T\}$, we have $N_z(t_{k(i)},t)\leq T$. Solve $\epsilon^2 T<\sqrt{|\mathcal{S}|T\log(T)}$ in ϵ and invoke **(STEP 1)**. ■

(STEP 3) *There exists* $\epsilon_z>0$ *such that, for all* $p'(z)$ *satisfying* $\mathrm{KL}(p'(z)||p(z))<\epsilon_z$, *we have* $\mathrm{supp}(p')\supseteq\mathrm{supp}(p)$ *and* $p'(s|z)\geq\frac{1}{2}p(s|z)$ *for all* $s\in\mathcal{S}$.

Proof Denote $x := \mathrm{KL}(p'(z)||p(z))$. By Pinkser's inequality, we have $\|p'(z)-p(z)\|_1 \leq \sqrt{2x}$, so

$$\forall s \in \mathcal{S}, \quad |p'(s|z) - p(s|z)| \leq \sqrt{2x}.$$

Assume that $\sqrt{2x} \leq \frac{1}{2}\min_{s\in\mathrm{supp}(p(z))} p(s|z)$. Then $p'(s|z) \geq \frac{1}{2}p(s|z)$ for all $s \in \mathcal{S}$ and in particular, $p'(z) \gg p(z)$. Hence the result. ■

(STEP 4) *For $\epsilon < (\frac{|\mathcal{S}|\log(T)}{T})^{1/4}$ and for $m \geq m_\lambda$, for all $\delta > 0$ and $m \geq t_\delta \in \mathbf{N}$, we have:*

$$\mathbf{P}\left(\exists t \in \{t_{k(i)}, \ldots, t_{k(i)} + T\} : \begin{array}{c} N_z(t_{k(i)}, t) \geq \dfrac{T\lambda^2|\mathcal{S}|\log\left(\frac{T}{\delta}\right)}{c^2} + 4\log\left(\dfrac{e}{c}\right)^2 \\ \textit{and } F_t \textit{ and } F_{t_{k(i)-1}} \textit{ and } \mathcal{P}_z(t) \not\subseteq \mathcal{P}_z(t_{k(i)-1}) \end{array}\right) \leq \delta.$$

Proof Let $\tilde{p}(z) \in \mathcal{P}_z(t)$. We derive conditions on $N_z(t_{k(i)}, t)$ such that $\tilde{p}(z) \in \mathcal{P}_z(t_{k(i)-1})$ with high probability, by looking at when $N_z(t_{k(i)-1})\mathrm{KL}(\hat{p}_{t_{k(i)-1}}(z)||\tilde{p}(z)) \leq \alpha\log(\beta t_{k(i)-1})$ where $\alpha = |\mathcal{S}|$ and $\beta = 2e$. Let $\mathcal{S}(z) := p(z)$, which is the same as the support of $\hat{p}_t(z)$ on F_t by **(STEP 3)**. We have:

$$\begin{aligned}
&N_z(t_{k(i)-1})\mathrm{KL}(\hat{p}_{t_{k(i)-1}}(z)||\tilde{p}(z)) \\
&= N_z(t_{k(i)-1})\mathrm{KL}(\hat{p}_t(z) - w_{t_{k(i)-1},t}(z)||\tilde{p}(z)) \\
&= N_z(t_{k(i)-1})\sum_{s\in\mathcal{S}(z)}\left(\hat{p}_t(s|z) - w_{t_{k(i)-1},t}(s|z)\right)\log\left(\frac{\hat{p}_t(s|z) - w_{t_{k(i)-1},t}(s|z)}{\tilde{p}(s|z)}\right) \\
&= N_z(t_{k(i)-1})\sum_{s\in\mathcal{S}(z)}\left(\hat{p}_t(s|z) - w_{t_{k(i)-1},t}(s|z)\right)\left(\log\left(\frac{\hat{p}_t(s|z)}{\tilde{p}(s|z)}\right) + \log\left(1 - \frac{w_{t_{k(i)-1},t}(s|z)}{\hat{p}_t(s|z)}\right)\right) \\
&= N_z(t_{k(i)-1})\left(\mathrm{KL}(\hat{p}_t(z)||\tilde{p}(z)) - \sum_{s\in\mathcal{S}(z)} w_{t_{k(i)-1},t}(s|z)\left(\log(\hat{p}_t(s|z)) + \log\left(\frac{1}{\tilde{p}(s|z)}\right)\right)\right) \\
&\quad + N_z(t_{k(i)-1})\sum_{s\in\mathcal{S}(z)}\left(\hat{p}_t(s|z) - w_{t_{k(i)-1},t}(s|z)\right)\log\left(1 - \frac{w_{t_{k(i)-1},t}(s|z)}{\hat{p}_t(s|z)}\right).
\end{aligned} \tag{42}$$

Let $c := 2\min_{s\in\mathcal{S}(z)} p(s|z)$. By **(STEP 3)**, $\min_{s\in\mathcal{S}(z)}\hat{p}_t(s|z) \geq c$ on F_t. So $|\log(\hat{p}_t(s|z))| \leq \log(\frac{1}{c})$ for all $s \in \mathcal{S}(z)$.

Furthermore, as $\tilde{p}(z) \in \mathcal{P}_z(t)$, we have $N_z(t)\mathrm{KL}(\hat{p}_t(z)||\tilde{p}(z)) \leq \alpha\log(\beta t)$ where $\alpha = |\mathcal{S}|$ and $\beta = 2e$ by construction of $\mathcal{P}_z(t)$, see (5). Writing $\mathrm{Ent}(\hat{p}_t(z)) := -\sum_s \hat{p}_t(s|z)\log(\hat{p}_t(s|z))$ the Shannon entropy of $\hat{p}_t(z)$, we have

$$\begin{aligned}
\alpha\log(\beta t) &\geq N_z(t)\sum_{s\in\mathcal{S}(z)}\hat{p}_t(s|z)\log\left(\frac{\hat{p}_t(s|z)}{\tilde{p}(s|z)}\right) \\
&\geq N_z(t)\left(\sum_{s\in\mathcal{S}(z)}\hat{p}_t(s|z)\log\left(\frac{1}{\tilde{p}(s|z)}\right) - \mathrm{Ent}(\hat{p}_t(z))\right)
\end{aligned}$$

$$\geq N_z(t)\left(\sum_{s\in\mathcal{S}(z)}\hat{p}_t(s|z)\log\left(\frac{1}{\tilde{p}(s|z)}\right)-\log|\mathcal{S}|\right)$$

so we find that $\log(\frac{1}{\tilde{p}(s|z)}) \leq \frac{1}{cN_z(t)}(\alpha\log(\beta t)+\log|\mathcal{S}|) \leq \frac{\alpha\log(\beta' t)}{cN_z(t)}$ for some $\beta' > 0$. Using this to continue the computations from (42) and further using $\log(1+x)\leq x$, we have:

$$\begin{aligned}
&N_z(t_{k(i)-1})\mathrm{KL}(\hat{p}_{t_{k(i)-1}}(z)||\tilde{p}(z))\\
&\leq N_z(t_{k(i)-1})\left(\mathrm{KL}(\hat{p}_t(z)||\tilde{p}(z))+\|w_{t_{k(i)-1},t}(z)\|_1\left(\frac{\alpha\log(\beta' t)}{cN_z(t)}+\log\left(\frac{1}{c}\right)\right)\right)\\
&\quad-N_z(t_{k(i)-1})\sum_{s\in\mathcal{S}(z)}\left(\hat{p}_t(s|z)-w_{t_{k(i)-1},t}(s|z)\right)\frac{w_{t_{k(i)-1},t}(s|z)}{\hat{p}_t(s|z)}.\\
&\leq N_z(t_{k(i)-1})\left(\mathrm{KL}(\hat{p}_t(z)||\tilde{p}(z))+\|w_{t_{k(i)-1},t}(z)\|_1\left(\frac{\alpha\log(\beta' t)}{cN_z(t)}+\log\left(\frac{1}{c}\right)+1\right)\right)\\
&\quad+N_z(t_{k(i)-1})\frac{\|w_{t_{k(i)-1},t}(z)\|_2^2}{c}\\
&\overset{(\dagger)}{\leq}\alpha\log(\beta t)-\tfrac{N_z(t_{k(i)-1},t)\alpha\log(\beta t)}{N_z(t)}+\tfrac{\sqrt{|\mathcal{S}|N_z(t_{k(i)},t)\log(T/\delta)}}{N_z(t)}\left(\tfrac{\alpha\log(\beta' t)}{cN_z(t)}+\log(\tfrac{1}{c})+1\right)\\
&\quad+\mathrm{O}\left(\tfrac{1}{N_z(t)}\left(\tfrac{\alpha\log(\beta' t)}{cN_z(t)}+\log(\tfrac{1}{c})+1\right)+\tfrac{1+|\mathcal{S}|N_z(t_{k(i)},t)\log(T/\delta)}{cN_z(t)}\right)\\
&\leq\alpha\log(\beta t)+\tfrac{\alpha\log(\beta t)}{N_z(t)}\left(-N_z(t_{k(i)},t)+\tfrac{\log(\beta t)}{\log(\beta' t)}\left(\tfrac{\log(\beta' t)}{cN_z(t)}\sqrt{|\mathcal{S}|\log(\tfrac{T}{\delta})}+\log(\tfrac{e}{c})\right)\sqrt{N_z(t_{k(i)},t)}\right)\\
&\quad+\mathrm{O}\left(\tfrac{T\log(T/\delta)}{N_z(t)}\right)\\
&\overset{(\ddagger)}{\leq}\alpha\log(\beta t_{k(i)-1})+\tfrac{\alpha\log(\beta t)}{N_z(t)}\left(-N_z(t_{k(i)},t)+2\left(\tfrac{\lambda}{c}\sqrt{|\mathcal{S}|\log(\tfrac{T}{\delta})}+\log(\tfrac{e}{c})\right)\sqrt{N_z(t_{k(i)},t)}\right)\\
&\quad+\mathrm{O}\left(\tfrac{N_z(t_{k(i)},t)\log(T/\delta)}{N_z(t)}\right)
\end{aligned}$$

where $(\dagger)$ follows from (**STEP 2**) and holds with probability $1-\delta$ on $F_{t_{k(i)-1}}$, and $(\ddagger)$ follows by using that (1) $t_{k(i)-1}\leq 3t$ if t is large enough, (2) that $N_z(t)<\frac{1}{\lambda}\log(t)$ on F_t and (3) that $\log(\beta' t)/\log(\beta t)\leq 2$ for t large enough. We want the RHS to be smaller than $\alpha\log(\beta t_{k(i)-1})$. For large t, we can neglect the second order term in $N_z(t_{k(i)},t)\log(T/\delta)/N_z(t)$ when $t\gg T/\delta$, because $\log(\beta t)\gg\log(T/\delta)$. This leads to a condition of the form:

$$N_z(t_{k(i)},t)\geq 2\left(\tfrac{\lambda}{c}\sqrt{|\mathcal{S}|\log(\tfrac{T}{\delta})}+\log(\tfrac{e}{c})\right)\sqrt{N_z(t_{k(i)},t)}$$

that leads immediately to the claimed result by using $(a+b)^2\leq 2a^2+2b^2$. ■

(**STEP 4**) concludes the proof. ■

D.2.2. A "STRICT" SHRINKING EFFECT FOR REWARDS

We continue with the shrinking effect for rewards. The proof is essentially similar to the shrinking effect for kernels (Lemma 27) but the result is more precise, because we quantify the speed of the shrinking phenomenon. Therefore, the proof requires an extra step.

Lemma 28 (Shrinking effect, rewards) *Let $(t_{k(i)})$ be the enumeration of exploration episodes, and let $T \geq 1$. Fix $\lambda > 0$ and $z \in \mathcal{Z}$. For all $\delta, \eta > 0$, we can find $\epsilon, m, C > 0$ such that:*

$$\mathbf{P}\left(\exists t \in \{t_{k(i)}, \ldots, t_{k(i)} + T\} : \begin{array}{c} \max \mathcal{R}_z(t) > \max \mathcal{R}_z(t_{k(i)-1}) - \frac{N_z(t) - N_z(t_{k(i)})}{C \cdot (t_{k(i)})^\eta} \\ \text{and } F_{t_{k(i)-1}} \cap F_t \text{ and } N_z(t) > N_z(t_{k(i)}) + C \log\left(\frac{T}{\delta}\right) \end{array} \right) \leq \delta$$

with $F_t := (N_z(t) < \frac{1}{\lambda} \log(t), \mathrm{KL}(\hat{q}_t(z)||q(z)) < \epsilon, N_z(t) > m)$ where $q(z) \equiv (r(z), p(z))$.

Proof The proof is essentially similar to Lemma 27. For rewards however, Lemma 28 quantifies the shrinking speed, hence we need to refine what is being said at the end of the proof of Lemma 27. Following (**STEP 4**) of the previous proof, for

$$\frac{1}{2} N_z(t_{k(i)}, t) \geq 2\left(\tfrac{\lambda}{c} \sqrt{|\mathcal{S}| \log\left(\tfrac{T}{\delta}\right)} + \log\left(\tfrac{e}{c}\right) \right) \sqrt{N_z(t_{k(i)}, t)},$$

we essentially have, on $F_t \cap F_{t_{k(i)-1}}$, that

$$N_z(t_{k(i)-1}) \mathrm{KL}(\hat{r}_{t_{k(i)-1}}(z)||\tilde{r}(z)) \leq \left(1 - \frac{N_z(t_{k(i)}, t)}{2 N_z(t)}\right) \alpha \log(\beta t) \tag{43}$$

for all $\tilde{r}(z) \in \mathcal{R}_t(z)$. Introduce the optimistic rewards $r_t^+(z) := \max \mathcal{R}_z(t)$ and $r_{t_{k(i)-1}}^+(z) := \max \mathcal{R}_z(t_{k(i)-1})$, and let $\omega_{t_{k(i)-1},t}^+(z) := r_t^+(z) - r_{t_{k(i)-1}}^+(z)$ be their difference. Following (43), we have

$$\mathrm{KL}(\hat{r}_{t_{k(i)-1}}(z)||r_t^+(z)) \leq \left(1 - \frac{N_z(t_{k(i)}, t)}{2 N_z(t)}\right) \cdot \mathrm{KL}(\hat{r}_{t_{k(i)-1}}(z)||r_{t_{k(i)-1}}^+(z)).$$

Approximating $\mathrm{KL}(\hat{r}_{t_{k(i)-1}}(z)||r_{t_{k(i)-1}}^+(z) + w_{t_{k(i)-1},t}^+(z))$ by its Taylor expansion at first order, we find:

$$\mathrm{KL}(\hat{r}_{t_{k(i)-1}}(z)||r_t^+(z)) \approx \mathrm{KL}(\hat{r}_{t_{k(i)-1}}(z)||r_{t_{k(i)-1}}^+(z)) + \frac{r_{t_{k(i)-1}}^+(z) - \hat{r}_{t_{k(i)-1}}(z)}{r_{t_{k(i)-1}}^+(z)(1 - r_{t_{k(i)-1}}^+(z))} w_{t_{k(i)-1},t}^+(z)$$

so that, at first order, we obtain the equation:

$$\frac{r_{t_{k(i)-1}}^+(z) - \hat{r}_{t_{k(i)-1}}(z)}{r_{t_{k(i)-1}}^+(z)(1 - r_{t_{k(i)-1}}^+(z))} w_{t_{k(i)-1},t}^+(z) \approx -\frac{N_{t_{k(i)-1},t}(z) \alpha \log(\beta t_{k(i)-1})}{N_z(t_{k(i)-1})^2}$$

and solving in $w_{t_{k(i)-1},t}^+(z)$ provides:

$$w_{t_{k(i)-1},t}^+(z) \approx -\frac{N_{t_{k(i)-1},t}(z) \alpha \log(\beta t_{k(i)-1})}{N_z(t_{k(i)-1})^2} \cdot \frac{r_{t_{k(i)-1}}^+(z)(1 - r_{t_{k(i)-1}}^+(z))}{r_{t_{k(i)-1}}^+(z) - \hat{r}_{t_{k(i)-1}}(z)}. \tag{44}$$

The question is how close to the boundary $r^+_{t_{k(i)-1}}(z)(1-r^+_{t_{k(i)-1}}(z))$ can be. Thanks to (**STEP 3**) of the proof of Lemma 27, on $F_{t_{k(i)-1}}$, $\hat{r}_{t_{k(i)-1}}(z)$ and $r(z)$ have the same support with $\hat{r}_{t_{k(i)-1}}(z) \leq 2r(z) - 1 < 1$. By writing $\mathrm{KL}(x||y) = -\mathrm{Ent}(x) + x\log(\frac{1}{y}) + (1-x)\log(\frac{1}{1-y})$, the inequality $N_z(t_{k(i)-1})\mathrm{KL}(\hat{r}_{t_{k(i)-1}}(z)||r^+_{t_{k(i)-1}}(z)) = \alpha\log(\beta t_{k(i)-1})$ leads to:

$$
\begin{aligned}
1 - r^+_{t_{k(i)-1}}(z) &\geq \exp\left(- \frac{\frac{\alpha \log(\beta t_{k(i)-1})}{N_z(t_{k(i)-1})} - \mathrm{Ent}(\hat{r}_{t_{k(i)-1}}(z))}{1 - \hat{r}_{t_{k(i)-1}}(z)} \right) \\
&\geq 2^{\frac{2}{1-r(z)}} \left(\beta t_{k(i)-1}\right)^{-\frac{2\alpha}{(1-r(z))N_z(t_{k(i)-1})}} = \Omega\big((t_{k(i)-1})^{-\eta}\big) \qquad (45)
\end{aligned}
$$

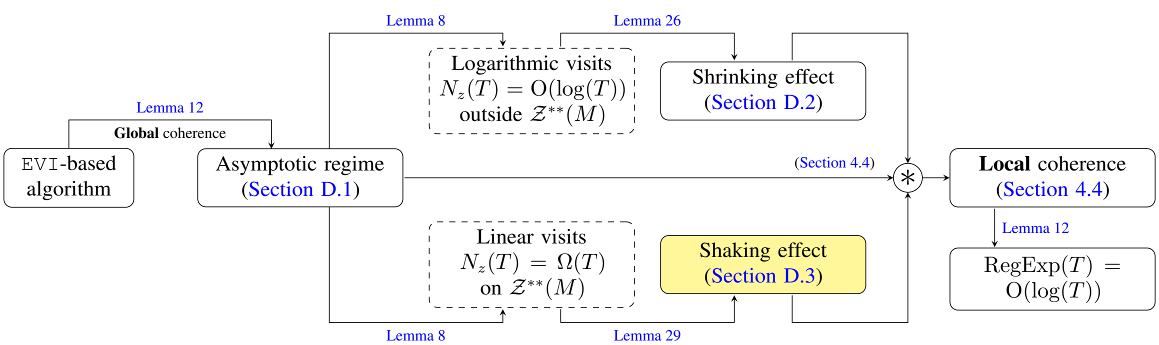
provided that $N_z(t_{k(i)-1}) \geq \frac{2\alpha}{\eta(1-r(z))}$. To conclude, we inject (45) into (44) together with the fact that, on $F_{t_{k(i)-1}}$, we have $N_z(t_{k(i)-1}) < \frac{1}{\lambda}\log(t_{k(i)-1})$, to get:

$$
w^+_{t_{k(i)-1},t}(z) \lesssim -\frac{N_z(t_{k(i)}, t)}{N_z(t_{k(i)-1})} \cdot \Omega\big((t_{k(i)-1})^{-\eta}\big) = -\Omega\left(\frac{N_z(t_{k(i)}, t)}{(t_{k(i)-1})^{\eta} \log(t_{k(i)-1})} \right).
$$

This concludes the proof. ■

D.3. The shaking effect: Proof of Lemma 29

In this section, we provide a proof of a formalized version of the **shaking effect** part of Informal Property 9.



In Lemma 29 below, we show that if $N_z(t) = \Omega(t)$ and under a good event, the reward-kernel confidence region $\mathcal{Q}_z(t) := \mathcal{R}_z(t) \times \mathcal{P}_z(t)$ barely changes compared to its state $\mathcal{Q}_z(t_{k(i)-1})$ at time $t_{k(i)-1}$, the beginning of the previous exploitation episode. The amount of displacement is quantified in Hausdorff distance and is shown of order $\sqrt{\log(t)/t}$. This will be negligible with respect to the displacements of the confidence region due to the shrinking effect, of which the order of magnitude is $\Omega((\frac{1}{t})^{\eta})$ for all $\eta > 0$.

Lemma 29 *Let $(t_{k(i)})$ be the enumeration of exploration episodes, and let $T \geq 1$. Fix $\lambda, z \in \mathcal{Z}$ and for two sets $\mathcal{U}, \mathcal{V} \subseteq \mathbf{R}^n$, denote $d_{\mathrm{H}}(\mathcal{U}, \mathcal{V})$ the Hausdorff distance induced by the one-norm. We can find $c, m > 0$ such that:*

$$
(\textit{kernels}) \quad F_{t_{k(i)}} \supseteq \left(\forall t \in [t_{k(i)}, t_{k(i)} + T] : d_{\mathrm{H}}(\mathcal{P}_z(t), \mathcal{P}_z(t_{k(i)-1})) \leq \sqrt{\frac{c\log(t)}{t}} \right),
$$

$$(rewards) \quad F_{t_{k(i)}} \supseteq \left(\forall t \in [t_{k(i)}, t_{k(i)} + T] : d_{\mathrm{H}}(\mathcal{R}_z(t), \mathcal{R}_z(t_{k(i)-1})) \leq \sqrt{\tfrac{c \log(t)}{t}} \right)$$

where $F_{t_{k(i)}} := (N_z(t_{k(i)-1}) > \lambda t_{k(i)-1}, t_{k(i)} > m) \cap (\forall t \in [t_{k(i)-1}, t_{k(i)}], M \in \mathcal{M}(t))$.

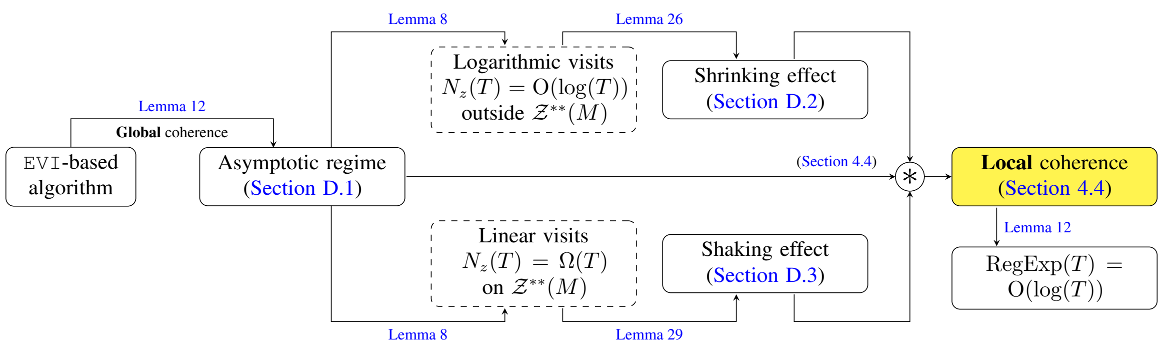
Proof We provide the argument for kernels, as the argument for rewards is the same in a smaller dimension. By Pinsker's inequality, $\|\hat{p}_t(z) - p'(z)\|_1 \leq 2\mathrm{KL}(\hat{p}_t(z)||p'(z))$, so on F_t and for all $p'(z) \in \mathcal{P}_z(t)$, we have $\|\hat{p}_t(z) - p'(z)\|_1 \leq (\lambda t)^{-1} 2\alpha \log(\beta t)$. On F_t, we further have $p(z) \in \mathcal{P}_z(t)$ as well, so $\|\hat{p}_t(z) - p(z)\|_1 \leq (\lambda t)^{-1} \cdot 2\alpha \log(\beta t)$. We deduce that, on F_t:

$$\mathcal{P}_z(t) \subseteq \left\{ p'(z) \in \mathcal{P}(\mathcal{S}) : \|p'(z) - p(z)\|_1 \leq 2\sqrt{\frac{2\alpha \log(\beta t)}{\lambda t}} \right\}.$$

The result is therefore obtained by estimating the Hausdorff distance between ℓ_1-ball of radius $\Theta(\sqrt{\log(t)/t})$ centered at $p(z)$. ■

D.4. Combining everything together: Proof of Lemma 13

Combining the shrinking-shaking effect and the asymptotic visit rates of optimal and non-optimal pairs, we establish the local coherence property of Lemma 13.



This is the last step in the proof of Theorem 7, assertion 3.

Lemma 13 (Local coherence) *Let $M \in \mathcal{M}^+$ be a non-degenerate explorative model. Consider running* `KLUCRL` *with model satisfying Assumption 4 and assume that episodes are managed with the f-(*VM*) with $f(t) = \mathrm{o}(\frac{1}{\log(t)})$. Let $(t_{k(i)})$ be the enumeration of exploration episodes. Then, there exists a constant $C(M) > 0$ such that, for all $T \geq 1$ and $\delta > 0$, there is an adapted sequence of events (E_t) and a function $\varphi : \mathbf{N} \to \mathbf{R}$ such that:*

1. *For all $i \geq 1$, the algorithm is $(E_t, t_{k(i)}, T, \varphi)$-coherent;*
2. $\mathbf{P}(\bigcup_{t=t_{k(i)}}^{t_{k(i)}+T-1} E_t^c) \leq \delta + \mathrm{o}(1)$ *when $i \to \infty$;*
3. $\varphi(t) \leq 1 + C \log(\frac{T}{\delta}) + \mathrm{o}(1)$ *when $t \to \infty$.*

Proof By correctness of the confidence region, $\mathbf{P}(\exists T, \forall t \geq T : \forall \pi, g^\pi(\mathcal{M}(t)) \geq g^\pi(M)) = 1$, hence a policy with optimistic gain less than $g^*(M)$ won't be optimistically optimal on this event, so won't be the result of `EVI`. Considering an exploration time $t_{k(i)}$, we know that the policy of the previous episode was optimal in M, hence $g^*(\mathcal{M}(t_{k(i)-1})) = g^{\pi^*}(\mathcal{M}(t_{k(i)-1}))$ where $\pi^* \in \Pi^*(M)$. By Assumption 4, we know that $g^*(\mathcal{M}(t_{k(i)-1}))$ only depends on $\mathcal{R}_z(t_{k(i)-1})$ and $\mathcal{P}_z(t_{k(i)-1})$ for $z \in \mathcal{Z}^{**}(M)$ where $N_z(t_{k(i)-1}) \geq \lambda t_{k(i)-1}$ by Lemma 8. Using Lemma 32 to quantify the sensibility of the gain to kernel and reward perturbations, we get that

$$g^*(M) \leq g^*(\mathcal{M}(t_{k(i)-1})) \leq g^*(M) + \mathrm{O}\left(\sqrt{\frac{\log(t_{k(i)-1})}{t_{k(i)-1}}}\right) \tag{46}$$

holds with probability one when $i \to \infty$.

Fix $t \in \{t_{k(i)}, \dots, t_{k(i)} + T - 1\}$. Recall that a policy that `EVI` outputs must have optimistic gain with span zero. Let π be the output of `EVI` at time $t' \in [t_{k(i)}, t]$, and assume that (1) π is sub-optimal in M from S_t, so that there exists $s \in \mathcal{S}$ such that $g^\pi(s; M) < g^*(s; M)$ and s is reachable from S_t under π; and (2) that $N_z(t) > N_z(t_{k(i)}) + C \log(T/\delta)$ for all $z \in \mathcal{Z}$, where C is given by the shrinking-shaking Lemmas 26 and 29. Without loss of generality, we can assume that s is recurrent under π on M and let $\mathcal{Z}' \subseteq \mathcal{Z}$ be the associated recurrent component of pairs. By Assumption 4, we see that $g^\pi(s; \mathcal{M}(t))$ only depends on data on $\mathcal{Z}'$. Since π was output by `EVI`, $g^\pi(\mathcal{M}(t))$ only depends on data on $\mathcal{Z}'$. Let $\mathcal{Z}'_- := \mathcal{Z}' \setminus \mathcal{Z}^{**}(M)$. This set is non-empty because $g^\pi(s; M) < g^*(s; M)$. Let $\mathcal{Z}'_+ := \mathcal{Z}' \cap \mathcal{Z}^{**}(M)$. We have:

$$\begin{aligned}
g^\pi(\mathcal{M}(t)) &= \sup_{\tilde{r} \in \mathcal{R}_\pi(t)} \sup_{\tilde{p} \in \mathcal{P}_\pi(t)} g(r, p) = \sup_{\tilde{r} \in \mathcal{R}_{\mathcal{Z}'}(t)} \sup_{\tilde{p} \in \mathcal{P}_{\mathcal{Z}'}(t)} g(\tilde{r}, \tilde{p}) \\
&\overset{(\dagger)}{\leq} \sup_{\tilde{r} \in \mathcal{R}_{\mathcal{Z}'}(t_{k(i)-1})} \sup_{\tilde{p} \in \mathcal{P}_{\mathcal{Z}'}(t_{k(i)-1})} g\left(\tilde{r} - \tfrac{\log(T/\delta)}{\log(t_{k(i)})} \cdot e_{\mathcal{Z}'_-} + \sqrt{\tfrac{c \log(t_{k(i)})}{t_{k(i)}}} \cdot e_{\mathcal{Z}'_+}, \tilde{p}\right) \\
&\overset{(\ddagger)}{\leq} g^\pi(\mathcal{M}(t_{k(i)-1})) + \sqrt{\tfrac{c \log(t)}{t}} - \eta(M, \pi) \tfrac{\log(T/\delta)}{\log(t_{k(i)})} \\
&\sim g^\pi(\mathcal{M}(t_{k(i)-1})) - \tfrac{\eta(M,\pi) \log(T/\delta)}{\log(t_{k(i)})} \\
&\overset{(46)}{\leq} g^*(M) + \mathrm{O}\left(\sqrt{\tfrac{\log(t_{k(i)-1})}{t_{k(i)-1}}}\right) - \tfrac{\eta(M,\pi) \log(T/\delta)}{\log(t_{k(i)})} < g^*(M)
\end{aligned}$$

where the last inequality hold for $t_{k(i)}$ large enough. In the above, $(\dagger)$ holds on the events specified by the shrinking-shaking behavior of confidence regions, see Lemmas 26 and 29; and $(\ddagger)$ is a technical result on exit probabilities, stating that even though we take a supremum on $\tilde{p} \in \mathcal{P}_{\mathcal{Z}'}(t_{k(i)} - 1)$, the choice of $\tilde{p}$ will put positive probability mass $\eta(M, \pi) > 0$ on $\mathcal{Z}'_-$ in its associated invariant probability measures.

This is justified as follows. On $\mathcal{Z}'_+ \equiv \mathcal{Z}' \cap \mathcal{Z}^{**}(M)$, the number of visits is $\omega(t_{k(i)-1})$ hence $\mathcal{P}_z(t_{k(i)-1})$ is nearly equal to $\{p_z\}$ for all $z \in \mathcal{Z}'_+$; In fact, for all fixed $\epsilon > 0$, we can assume that $\mathcal{P}_z(t_{k(i)-1}) \subseteq \{\tilde{p}_z : \|\tilde{p}_z - p_z\|_1 < \epsilon\}$ with overwhelming probability provided that $t_{k(i)-1}$ is large enough. Let $(\tilde{r}^\pi, \tilde{p}^\pi) \in \mathcal{M}^\pi(t_{k(i)-1})$ be an optimistic model of π (see Section B.1) and let $\tilde{\mu}^\pi$ be the empirical invariant measure of π starting from s under the optimistic model. Using that $\mathrm{sp}(g(\tilde{r}^\pi, \tilde{p}^\pi)) = 0$, we assume that $\tilde{p}^\pi$ has a single recurrent class $\mathcal{Z}''$ up to restricting to that class. By correctness of the confidence region, a policy output by `EVI` has optimistic gain higher than $g^*(M)$ and since the optimistic model is nearly equal to the true model on $\mathcal{Z}'_+$, we deduce that $\mathcal{Z}''$ must contain elements of $\mathcal{Z}'_-$ (otherwise π is optimal in M). We see that under $\tilde{p}^\pi$, for every element of $\mathcal{Z}'' \cap \mathcal{Z}'_+$ there must be a path to an element of $\mathcal{Z}'' \cap \mathcal{Z}'_-$ of length at most $|\mathcal{S}| - 1$ and probability at least $c_\epsilon(M) := (\min_{z\in\mathcal{Z}'_+} \min\{p(s|z) > 0 : s \in \mathcal{S}\} - \epsilon)^{|\mathcal{S}|-1}$, which is well-defined and positive for $\epsilon > 0$ small enough. So there must be $z \in \mathcal{Z}'' \cap \mathcal{Z}'_-$ such that $\tilde{\mu}(z) \geq |\mathcal{S}|^{-1} c_\epsilon(M)$. Set $\eta(M, \pi) := \frac{1}{2}c_0(M)$. For ϵ small enough and on mild concentration events, we have:

$$g\left(\tilde{r}^\pi - \tfrac{\log(T/\delta)}{\log(t_{k(i)})} \cdot e_{\mathcal{Z}'_-} + \sqrt{\tfrac{c\log(t_{k(i)})}{t_{k(i)}}} \cdot e_{\mathcal{Z}'_+}, , \tilde{p}^\pi\right) \leq g^\pi(\mathcal{M}(t_{k(i)-1})) + \sqrt{\tfrac{c\log(t)}{t}} - \eta(M, \pi)\tfrac{\log(T/\delta)}{\log(t_{k(i)})}.$$

This justifies (‡).

Overall, we have $g^\pi(\mathcal{M}(t)) < g^*(M) \leq g^*(\mathcal{M}(t))$ on the event $E_t := \bigcap_{z\in\mathcal{Z}} E^z_t$ with E^z_t given by, for $z \notin \mathcal{Z}^{**}(M)$:

$$\left(F^z_{t_{k(i)}}, \begin{bmatrix} \mathcal{P}_z(t) \subseteq \mathcal{P}_z(t_{k(i)-1}) \\ \text{or } N_z(t) \leq N_z(t_{k(i)}) + C\log\left(\frac{T}{\delta}\right)\end{bmatrix}, \begin{bmatrix} \sup \mathcal{R}_z(t) \leq \sup \mathcal{R}_z(t_{k(i)-1}) - \frac{N_z(t) - N_z(t_{k(i)})}{C\log(t_{k(i)})} \\ \text{or } N_z(t) \leq N_z(t_{k(i)}) + C\log\left(\frac{T}{\delta}\right)\end{bmatrix}\right)$$

and for $z \in \mathcal{Z}^{**}(M)$:

$$\left(F^z_{t_{k(i)}}, d_{\mathrm{H}}(\mathcal{P}_z(t), \mathcal{P}_z(t_{k(i)-1})) \leq \sqrt{\tfrac{c\log(t)}{t}}, d_{\mathrm{H}}(\mathcal{R}_z(t), \mathcal{R}_z(t_{k(i)-1}))\right)$$

where, for $z \notin \mathcal{Z}^{**}(M)$, $F^z_{t_{k(i)}}$ is the event appearing in the shrinking effect lemma (Lemma 26), and for $z \in \mathcal{Z}^{**}(M)$, $F^z_{t_{k(i)}}$ is the event appearing in the shaking effect lemma (Lemma 29); In both cases, we have $\mathbf{P}(\exists i, \forall j \geq i : F^z_{t_{k(j)}}) = 1$ provided that the rate $\lambda > 0$ in the definition of $F^z_{t_{k(i)}}$ is chosen accordingly to the asymptotic regime of the algorithm (Lemma 8). We deduce that on E_t, π will be rejected as soon as (VM) triggers, because its optimistic gain is no more optimistically optimal. By (9), as soon as a pair $z \notin \mathcal{Z}^{**}(M)$ is about to be visited for the second time in the episode, the episode will stop. We therefore have shown that while $g^\pi(S_t; M) < g^*(S_t; M)$ and on E_t, there exists $z \equiv (s, a)$ that is reachable from S_t under π such that $N_z(t) < N_z(t_k) + 1 + C\log(T/\delta)$ and $g^\pi(s; M) < g^*(s; M)$.

Accordingly, we have shown that the algorithm is $(E, t_{k(i)}, T, \varphi)$-coherent, with $\mathbf{P}(\exists t \in [t_{k(i)}, t_{k(i)} + T] : E^c_t) \leq \delta + \mathrm{o}(1)$ when $i \to \infty$ and $\varphi(t) = 1 + C\log(T/\delta)$. ∎

Appendix E. Model dependent regret guarantees via coherence

In the proof of the regret of exploration guarantees, Lemma 12 is used twice and two different coherence properties are invoked. Coherence is first used in a *global* form to derive the almost sure asymptotic regime. Indeed, the first step of the proof (see Section D.1) consists in showing that the algorithm is $((F_t), \lceil \log(T) \rceil, T, \varphi)$-coherent for $\varphi(\lceil \log(T) \rceil) = \mathrm{O}(\log(T))$ where the sequence of events (F_t) is asymptotically almost-sure, i.e., $\mathbf{P}(\exists T, \forall t \geq T : F_t) = 1$. Then, coherence is used in a *local* form to derive the regret of exploration guarantees. Indeed, the whole point of Section 4.4 is to show that the algorithm is $(E, t_{k(i)}, T, \varphi)$-coherent where $(t_{k(i)})$ is the sequence of exploration episodes, $\mathbf{P}(\exists T, \forall t \geq T : E_t) = 1$ and $\varphi(t_{k(i)}) = \mathrm{O}(\log(T))$.

In this appendix, we show a third application of coherence properties: model dependent regret guarantees.

E.1. A general model dependent regret bound via coherence

We provide first a general result.

Theorem 30 *Consider an episodic algorithm with* (1) *weakly regenerative episodes and* (2) *such that there exists an adapted sequence of events* (F_t) *with* $\mathbf{P}(\bigcup_{t=T}^{\infty} F_t^c) = \mathrm{O}(\frac{1}{T})$ *such that the algorithm is* $((F_t), T, T, \varphi)$*-coherent for all* $T \geq 1$*. Then, for all non-degenerate model* M*,*

$$\mathrm{Reg}(T; M) = \mathrm{O}\left(\sum_{m=0}^{\lceil \log_2(T) \rceil - 1} \varphi(2^m)\right) + \mathrm{O}(\log(T)) \tag{47}$$

when $T \to \infty$.

Proof Let $n := \lceil \log_2(T) \rceil$. For all $m \leq n$, the algorithm is $(F, 2^m, 2^m, \varphi)$-coherent, has weakly regenerative episodes, and M is non-degenerate, so we invoke Lemma 12 and obtain, for $x \geq 0$,

$$\begin{aligned}
&\mathbf{P}\big(\mathrm{Reg}(2^m, 2^{m+1}) \geq x + C_4 \varphi(2^n)\big) \\
&\leq \mathbf{P}\left(\mathrm{Reg}(2^m, 2^{m+1}) \geq x + C_4 \varphi(2^n), \bigcap_{t=2^m}^{2^{m+1}-1} F_t\right) + \mathbf{P}\left(\bigcup_{t=2^m}^{\infty} F_t^c\right) \\
&\leq \exp\left(-\frac{x}{C_2} + C_3 m \log(2) + \log(C_1)\right) + \mathrm{O}\big(2^{-m}\big)
\end{aligned}$$

where C_1, C_2, C_3, C_4 are model dependent constants. For $x \geq C_2(C_1 + (1 + C_3)\log(2) m)$, the RHS is $\mathrm{O}(2^{-m})$. In other words, $\mathrm{Reg}(2^m, 2^{m+1}) = \mathrm{O}(\varphi(2^m))$. Summing for $m \geq 1$, we get:

$$\begin{aligned}
\mathrm{Reg}(T) &:= \sum_{m=0}^{n-1} \mathrm{Reg}(2^m, 2^{m+1}) \\
&= \mathrm{O}\left(\sum_{m=0}^{n-1} \varphi(2^m) + 1\right)
\end{aligned}$$

$$= \mathrm{O}\left(\sum_{m=0}^{\lceil \log_2(T) \rceil - 1} \varphi(2^m) \right) + \mathrm{O}(\log(T)).$$

This is the intended result. ■

A few comments are in order. First, the requirement $\mathbf{P}(\bigcup_{t=T}^{\infty} F_t^c) = \mathrm{O}(\frac{1}{T})$ is slightly overshoot and can be weakened depending on the asymptotic properties of φ and the desired bound. Second, the proof technique can be directly adapted to obtain bounds in probability rather than in expectation. Last, but perhaps the most important, is that this bound only holds for non-degenerate models (Definition 4). While every model can be made non-degenerate up to smooth reward perturbations, non-degenerate models are a bit special, because the weakly optimal pair is unique from every state (unique Bellman optimal policy), and $\mathcal{Z}^{**}(M)$ has a unique communicating component (unique gain optimal component), see Section F.2. The proof of Lemma 12, which is key here, inevitably relies on non-degeneracy. Yet, degenerate models are easy to find. When Ortner (2010) discusses the necessity for episodes (see his Figure 2), he exhibits a degenerate model for that purpose. This simple example is a good starting point to understand why coherence and weakly regenerative episodes are insufficient to provide regret bounds on degenerate models.

E.2. A model dependent regret bound for (VM)

Theorem 30 is applied to `KLUCRL` managing episodes with a f-(VM) rule, by showing that such algorithms satisfy a $((F_t), T, T, \varphi)$-coherence property with a budget function $\varphi(T) = \mathrm{O}(\log(T))$, leading to $\mathrm{O}(\log(T) \log\log(T))$ regret bounds.

Theorem 31 *Let M be a non-degenerate model. Consider running `KLUCRL` with M satisfying Assumption 4 and assume that episodes are managed with a f-(VM) with $f > 0$. Then:*

$$\mathrm{Reg}(T; M) = \mathrm{O}(\log(T) \log\log(T)). \tag{48}$$

Proof Consider the good events $E_t := (M \in \mathcal{M}(t))$ and $F_t := \bigcap_{t'=(t-|\mathcal{Z}|)/2}^{t} E_t$.

By design $\mathbf{P}(\bigcup_{t=T}^{\infty} E_t^c) = \mathrm{O}(\frac{1}{T})$, see Lemma 18, so $\mathbf{P}(\bigcup_{t=T}^{\infty} F_t^c) = \mathrm{O}(\frac{1}{T})$ as well. We show that the algorithm is (F_t, T, T, φ)-coherent for $\varphi(T) = \mathrm{O}(\log(T))$. The result will then follow by Theorem 30 using that $\int \log(x) dx = x \log x - x$.

By Pinsker's inequality, for all $\epsilon > 0$, there exists $C \equiv C_\epsilon > 0$ such that, if $N_z(t) \geq C \log(t)$, then:

$$\mathcal{P}_z(t) \subseteq \left\{ \tilde{p}_z : \|\tilde{p}_z - \hat{p}_z(t)\|_1 < \tfrac{1}{2}\epsilon \right\} \quad \text{and} \quad \mathcal{R}_z(t) \subseteq \left\{ \tilde{r}_z : \|\tilde{r}_z - \hat{r}_z(t)\|_\infty < \tfrac{1}{2}\epsilon \right\} \tag{49}$$

Introduce the gain gap $\Delta_g := \min\{\|g^\pi(M) - g^*(M)\|_\infty : \pi \notin \Pi^*(M)\} > 0$. Whenever $M \in \mathcal{M}(t)$, we have $g^*(M) \leq g^*(\mathcal{M}(t))$. Let π be a policy output by `EVI` at time t and assume that $N_z(t) \geq C \log(t)$ for all $z \in \mathcal{Z}$. It has optimistic bias with span at most $D(M)$, hence by Lemma 32, we have:

$$\|g^\pi(\mathcal{M}_t) - g^\pi(M)\|_\infty \leq \epsilon\big(1 + \tfrac{1}{2} D(M)\big) \tag{50}$$

yet $g^\pi(\mathcal{M}_t) \geq g^*(\mathcal{M}_t) \geq g^*(M)$. So, provided that $\epsilon(1 + \frac{1}{2}D(M)) < \Delta_g$, π necessarily achieves optimal gain. We assume from now on that $\epsilon(1 + \frac{1}{2}D(M)) < \Delta_g$ is true.

Now, assume that π_t is such that $g^{\pi_t}(S_t, M) < g^*(S_t, M)$. By construction of `EVI`-based algorithms, π_t is the output of `EVI` for t_k with $t \in [t_k, t_{k+1})$, hence is the optimistically optimal policy at time t_k. By assumption $g^{\pi_{t_k}}(S_t; M) < g^*(S_t; M)$, so assuming that

$$E_{t_k} \equiv (M \in \mathcal{M}(t_k)) \tag{51}$$

holds, we deduce from the previous argument that there must be $z \in \mathcal{Z}$ such that $N_z(t_k) < C \log(t_k)$. Since $g^{\pi_{t_k}}(S_t, M) < g^*(S_t, M)$, $\mathrm{Reach}(\pi_{t_k}, S_t)$ must contain a recurrent component of π_{t_k} on which the achieved gain is sub-optimal. Pick one, denoted $\mathcal{Z}'$. Thanks to Assumption 4, the optimistic gain of $g^{\pi_{t_k}}(s, \mathcal{M}(t_k))$ for $s \in \mathcal{S}(\mathcal{Z}')$ only depends on pairs among $\mathrm{Reach}(\pi_{t_k}, s)$ and yet $g^{\pi_{t_k}}(s; \mathcal{M}(t_k)) \geq g^*(s, M)$. So there must be a sub-sampled pair in $\mathcal{Z}'$, i.e., there exists $(s, a) \in \mathcal{Z}'$ such that $N_{s,a}(t_k) < C \log(t_k)$; This pair is reachable from S_t under π_t and $g^{\pi_t}(s; M) < g^*(s; M)$ by construction of $\mathcal{Z}'$. Last, but not least, is that by construction of (`VM`), we have $t \leq 2t_k + |\mathcal{Z}|$ and $N_{s,a}(t) \leq 2N_{s,a}(t_k) + 1$. So, on the event $F_t := \bigcap_{t'=(t-|\mathcal{Z}|)/2}^{t} E_t$,

$$\exists z \equiv (s, a) \in \mathrm{Reach}(\pi_t, S_t) : \quad N_z(t) \leq 2C \log(t) + 1 \text{ and } g^{\pi_t}(s; M) < g^*(s; M). \tag{52}$$

Setting $\varphi(t) := 2C \log(2t) + 1$, we have shown that the algorithm is $((F_t), T, T, \varphi)$-coherent. We have $\varphi(T) = \mathrm{O}(\log(T))$ and $\mathbf{P}(\bigcup_{t=T}^{\infty} F_t^c) = \mathrm{O}(\frac{1}{T})$. Conclude by applying Theorem 30. ■

The result is remarkable in that f is basically arbitrary. It allows for $f(t)$ decreasing arbitrarily fast, hence for linearly many episodes, meaning that `KLUCRL` can nearly be episode-less on non-degenerate models, at the expense of minimax guarantees (see Theorem 14). This remark is to be combined with the observation that optimistic algorithms (Appendix B) cannot be episode-less on degenerate models in general, see Ortner (2010). In tandem, this indicates that coherence alone cannot provide regret guarantees beyond non-degenerate models. If the model dependent regret guarantees are obtained "for free" from coherence, extending such guarantees to degenerate models would require a different approach and most likely assumptions on the function f.

Whether the $\mathrm{O}(\log \log(T))$ factor can be removed remains an open question.

Appendix F. A few technical results on Markov decision processes

In this appendix, we provide a few useful technical results on Markov decision processes. In Section F.1, we provide a general result on the sensibility of the gain function to parameters, that is used at many places in this work. In Section F.2, we give a few insights regarding the non-degeneracy assumption of Definition 4. Lastly, we dedicate the last Section F.3 to the proof of Theorem 6, showing that the regret of exploration of existing algorithms is linear on explorative Markov decision processes.

F.1. Sensibility of the gain function to parameters

In Lemma 32, we explain how the gain function of a policy is subjected to vary under perturbation of the reward vector r and the transition kernel p. The gain function is shown to be 1-Lipschitz with respect to rewards, and $\frac{1}{2}\mathrm{sp}(h^\pi)$-Lipschitz with respect to kernels, where $h^\pi(s) := \lim \mathbf{E}_s^\pi[\sum_{t=1}^T (R_t - g^\pi(S_t))]$ is the bias function of the policy.

Lemma 32 *Let $M \equiv (\mathcal{Z}, p, r)$ and $\hat{M} \equiv (\mathcal{Z}, \hat{p}, \hat{r})$ be two Markov decision processes and fix $\pi \in \Pi$ a policy. If $\mathrm{sp}(g^\pi(M)) = 0$, then*

$$
\begin{aligned}
&\left\| g^\pi(\hat{M}) - g^\pi(M) \right\|_\infty \\
&\le \max_{s \in \mathcal{S}} \left\{ \left| \hat{r}(s, \pi(s)) - r(s, \pi(s)) \right| + \frac{1}{2}\mathrm{sp}(h^\pi(M)) \left\| \hat{p}(s, \pi(s)) - p(s, \pi(s)) \right\|_1 \right\}.
\end{aligned}
$$

Proof Let $T \ge 1$ and let $s \in \mathcal{S}$ be an initial state. Set $\epsilon_r^\pi := \|\hat{r}^\pi - r^\pi\|_\infty$ and $\epsilon_p^\pi := \|\hat{p}^\pi - p^\pi\|_1$.

$$
\begin{aligned}
&\mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} R_t\right] \\
&= \mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} \hat{r}^\pi(S_t)\right] \\
&\le \mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} r^\pi(S_t)\right] + T\epsilon_r^\pi \\
&\overset{(\dagger)}{=} \mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} (g^\pi(S_t) + (e_{S_t} - p(S_t, A_t))h^\pi)\right] + T\epsilon_r^\pi \\
&\overset{(\ddagger)}{\le} Tg^\pi(s) + \mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} \big((e_{S_{t+1}} - \hat{p}(S_t, A_t))h^\pi + (\hat{p}(S_t, A_t) - p(S_t, A_t))h^\pi \big)\right] \\
&\qquad\qquad + \mathrm{sp}(h^\pi) + T\epsilon_r^\pi \\
&\overset{(\S)}{=} T\big(g^\pi(s) + \epsilon_r^\pi + \tfrac{1}{2}\mathrm{sp}(h^\pi)\epsilon_p^\pi\big) + \mathrm{sp}(h^\pi)
\end{aligned}
$$

where $(\dagger)$ invokes the Poisson equation $g^\pi(S_t) + h^\pi(S_t) = r^\pi(S_t) + p^\pi(S_t)h^\pi$, $(\ddagger)$ uses that $g^\pi(S_t) = g^\pi(s)$ for all $t \ge 0$ and $(\S)$ that, if $p, p' \in \mathcal{P}(\mathcal{S})$ and $u \in \mathbf{R}^{\mathcal{S}}$ then $|(p' - p)u| \le$

$\frac{1}{2}\mathrm{sp}(u)\|p' - p\|_1$. Dividing by T and letting it go to infinity, we obtain the desired upper-bound. The lower bound is obtained similarly. ■

F.2. The space of non-degenerate Markov decision processes

In this section, we discuss of non-degeneracy assumption, found in Definition 4, and made in Theorem 7 for both model dependent regret guarantees and regret of exploration guarantees. We argue that while not all Markov decision processes are non-degenerate, most of them are.

Theorem 33 (Characterizations of non-degenerate MDPs) *Let $M \equiv (\mathcal{Z}, r, p)$ be a communicating Markov decision process. The following statements are equivalents.*

1. *M is non-degenerate in the sense of Definition 4: There is a unique policy satisfying the Bellman equations (i) $g^\pi(s) = \max_{a \in \mathcal{A}(s)}\{p(s,a)g^\pi\}$ and (ii) $g^\pi(s) + h^\pi(s) = \max_{a \in \mathcal{A}(s)}\{r(s,a) + p(s,a)h^\pi\}$ for all $s \in \mathcal{S}$, and this policy is unichain.*
2. *$\mathcal{Z}^*(M)$ is robust to reward noise, i.e., there exists $\epsilon > 0$ such that if $\|r' - r\|_\infty < \epsilon$, then $\mathcal{Z}^*(r', p) = \mathcal{Z}^*(r, p)$;*

The first characterization is the definition, stating that the Bellman optimal policy is unique and unichain. The second characterization states that the set of weakly optimal pairs is robust to reward perturbations, in other words, that the gap function $\Delta^*(-)$ has locally constant support.

We start with a lemma, showing that the uniqueness of weakly optimal actions is almost sure up to smooth perturbation of the reward function.

Lemma 34 *Let $M \equiv (\mathcal{Z}, r, p)$ be a communicating Markov decision process. Let $U(z)$ be i.i.d. random variables of distribution $\mathrm{N}(0, 1)$. Then $M_U := (\mathcal{Z}, r + U, p)$ has unique weakly optimal actions almost surely, i.e., $|\mathcal{Z}^*(M_U)| = |\mathcal{S}|$ almost surely.*

Proof If a model $M' \equiv (\mathcal{Z}, r', p)$ does not have unique optimal actions, then there exist $s \in \mathcal{S}$ as well as $a \neq a' \in \mathcal{A}(s)$ such that $(s, a), (s, a') \in \mathcal{Z}^*(M')$. In particular, we have:

$$r'(s,a) + p(s,a)h^*(r',p) = r'(s,a') + p(s,a')h^*(r',p). \tag{53}$$

Because h^* is obtained as the bias vector of some policy, we have in particular:

$$\exists \pi \in \Pi, \quad r'(s,a) + p(s,a)h^\pi(r',p) = r'(s,a') + p(s,a')h^\pi(r',p) \tag{54}$$

which is of the form "$\exists \pi \in \Pi, f^\pi(r') = 0$" where f^π are a linear forms. It happens that all are non-degenerate. Indeed, denoting (e_z) the canonical basis of $\mathbf{R}^{\mathcal{Z}}$, we see that for all $\pi \in \Pi$, either $f^\pi(e_{(s,a)}) \neq 0$ or $f^\pi(e_{(s,a')}) \neq 0$ depending on whether $\pi(s) = a$ or $\pi(s) \neq a$. It follows that the set of $r' \in \mathbf{R}^{\mathcal{Z}}$ satisfying (54) is a union of hyperplanes, hence is negligible with respect to the Lebesgue measure. It follows that $\mathbf{P}(r + U \text{ satisfies } (53)) = 0$. ■

Proof of Theorem 33 To begin with, note that if π^* is the unique Bellman optimal policy of a Markov decision process M, π^* is bias optimal by standard theory (Puterman, 1994, §9.2), i.e., $g^{\pi^*}(M) = g^*(M)$ and $h^{\pi^*}(M) = h^*(M)$. In particular, $\mathcal{Z}^*(M) = \{(s, \pi^*(s)) : s \in \mathcal{S}\}$.

Now, assume (*1.*) and let π^* be the unique Bellman optimal policy of M. We show (*2.*) by contradiction. Assume that $\mathcal{Z}^*(M)$ is not robust to reward noise. So, because $2^{\mathcal{Z}}$ is finite, there exists $\mathcal{Z}_0 \subseteq \mathcal{Z}$ with $\mathcal{Z}_0 \neq \mathcal{Z}^*(M)$, together with a sequence $r_n \to r$ such that $\mathcal{Z}^*(r_n, p) = \mathcal{Z}_0$ for all $n \geq 1$. Up to infinitesimal perturbation of r_n, we can further assume that $|\mathcal{Z}_0| = |\mathcal{S}|$ by Lemma 34. So, $\mathcal{Z}_0$ defines a policy π_0 where $\pi_0(s)$ picks the unique action a such that $(s, a) \in \mathcal{Z}_0$. By definition of $\mathcal{Z}^*(M)$, this policy is the unique bias optimal policy of every $M_n \equiv (\mathcal{Z}, r_n, p)$. In particular, π_0 is Bellman optimal in M_n. Now, the gain g^{π_0} and bias h^{π_0} are 1-Lipschitz in r. By taking $n \to \infty$, we conclude that π_0 is Bellman optimal in M as well. By (*1.*) uniqueness of the Bellman optimal policy of M, we have $\pi_0 = \pi^*$ hence $\mathcal{Z}_0 = \mathcal{Z}^*(M)$; A contradiction.

Conversely, assume (*2.*). By Lemma 34, it implies that $|\mathcal{Z}^*(M)| = |\mathcal{S}|$, hence that the bias optimal policy is unique. We prove (*1.*) by contradiction, so either M has multiple Bellman optimal policies, or its Bellman optimal policy is not unichain.

Assume that the Bellman optimal policy π^* is not unichain and let $\mathcal{Z}_1$ and $\mathcal{Z}_2$ be two disjoint recurrent components of π^*. Consider the model $M'_\epsilon := (\mathcal{Z}, r + \epsilon \mathbf{1}(\mathcal{Z}_1), p)$ for $\epsilon > 0$. We see that $g^*(M'_\epsilon) = g^*(M) + \epsilon$, while the gain of every unichain policy π with recurrent component $\mathcal{Z}_2$ is $g^\pi(M'_\epsilon) = g^*(M)$. Using the formula

$$g^\pi(s; M'_\epsilon) = g^*(s; M'_\epsilon) - \lim_{T\to\infty} \mathbf{E}_s^{\pi, M'_\epsilon}\left[\frac{1}{T}\sum_{t=1}^{T} \Delta^*(Z_t; M'_\epsilon)\right],$$

we conclude that there exists $z \in \mathcal{Z}_2$ such that $\Delta^*(z; M'_\epsilon) > 0$, hence $z \notin \mathcal{Z}^*(M'_\epsilon)$. So $\mathcal{Z}^*(M'_\epsilon) \neq \mathcal{Z}^*(M)$; A contradiction.

Now, assume that M has multiple Bellman optimal policies, say π_1^* and π_2^*. Without loss of generality, we assume that π_2^* is bias optimal in M, i.e., that $\pi_2^*(s)$ picks the unique action a such that $(s, a) \in \mathcal{Z}^*(M)$, which exists by (*2.*). With the same argument as before, we can show that both π_1^* and π_2^* are unichain. We can also show that π_1^* and π_2^* must have the same recurrent component; Otherwise, we introduce $\mathcal{Z}_1^*$ and $\mathcal{Z}_2^*$ their respective components, we consider $M'_\epsilon := (\mathcal{Z}, r + \epsilon\mathbf{1}(\mathcal{Z}_1^* \setminus \mathcal{Z}_2^*), p)$ and invoke the same rationale. So $h^{\pi_1^*} = h^{\pi_2^*} = h^*$ on the recurrent states of π_1^* and π_2^*. Let $\Delta_1^*(s, a) := g^{\pi_1^*}(s) + h^{\pi_1^*}(s) - r(s, a) - p(s, a) h^{\pi_1^*}$ be the gap function of π_1^* in M. Note that $\Delta_1^* \geq 0$ by definition of π_1^*. Now, let s_0 be a recurrent state of π_1^* and π_2^* and let $\tau_0 := \inf\{t \geq 1 : S_t = s_0\}$ be the reaching time to s_0. For all $s \in \mathcal{S}$, we have

$$\begin{aligned}
h^{\pi_2^*}(s) &= \mathbf{E}_s^{\pi_2, M}\left[\sum_{t=1}^{\tau_0 - 1}\left(h^{\pi_2^*}(S_t) - h^{\pi_2^*}(S_{t+1})\right)\right] + h^{\pi_2^*}(s_0) \\
&\overset{(\dagger)}{=} \mathbf{E}_s^{\pi_2, M}\left[\sum_{t=1}^{\tau_0 - 1}\left(r(Z_t) - g^*(S_t)\right)\right] + h^{\pi_2^*}(s_0) \\
&\overset{(\ddagger)}{=} \mathbf{E}_s^{\pi_2, M}\left[\sum_{t=1}^{\tau_0 - 1}\left(r(Z_t) - g^{\pi_1^*}(S_t)\right)\right] + h^{\pi_1^*}(s_0) \\
&\overset{(\S)}{=} \mathbf{E}_s^{\pi_2, M}\left[\sum_{t=1}^{\tau_0 - 1}\left(h^{\pi_1^*}(S_t) - h^{\pi_1^*}(S_{t+1}) - \Delta_1^*(Z_t)\right)\right] + h^{\pi_1^*}(s_0)
\end{aligned}$$

$$\overset{(\$)}{\leq} h^{\pi_1^*}(s)$$

where $(\dagger)$ invokes the Poisson equation of π_2^*; $(\ddagger)$ uses that π_1^* is gain optimal and that $h^{\pi_1^*}(s_0) = h^{\pi_2^*}(s_0)$; $(\S)$ follows by definition of the gap function Δ_1^*; and $(\$)$ uses that $\Delta_1^* \geq 0$. So $h^* = h^{\pi_2^*} \leq h^{\pi_1^*}$. So π_1^* is bias optimal. By (*2.*), the bias optimal policy of M is unique, so $\pi_1^* = \pi_2^*$; A contradiction. ■

Combining Theorem 33 and Lemma 34, we obtain the following result.

Corollary 35 (Non-degeneracy is almost-sure) *Let $M \equiv (\mathcal{Z}, r, p)$ be a communicating Markov decision process. Let $U(z)$ be i.i.d. random variables of distribution* $\mathrm{N}(0, 1)$. *Then $M_U := (\mathcal{Z}, r + U, p)$ is almost-surely non-degenerate.*

This result states that if a Markov decision process is degenerate, almost all its neighbors are non-degenerate. For instance, fixing the kernel then picking the reward function uniformly at random in $[0, 1]^{\mathcal{Z}}$, the resulting Markov decision process is non-degenerate with probability one. This supports the idea that, although many Markov decision processes are degenerate, most are non-degenerate.

F.3. Proof of Theorem 6: Algorithms based on (DT) have linear regret of exploration

In this paragraph, we prove Theorem 6.

Theorem 6. *Fix a pair space $\mathcal{Z}$ and let $\mathcal{M}$ be the space of all recurrent models with pairs $\mathcal{Z}$. Let $f : \mathbf{N} \to (0, \infty)$ be such that $\lim f(n) = +\infty$. Any no-regret episodic learner (π_t) satisfying:*

$$\begin{aligned} \forall k \geq 1, \exists z \in \mathcal{Z}, \quad & N_{t_{k+1}}(z) \geq N_{t_k}(z) + f(N_{t_k}(z)) \\ \exists c > 0, \forall t \geq 0, \forall (s,a) \in \mathcal{Z}, \quad & \pi_t(a|z) \geq c \text{ or } \pi_t(a|z) = 0 \end{aligned} \tag{55}$$

has linear regret of exploration on the explorative sub-space of $\mathcal{M}$, i.e., for all $M \in \mathcal{M}^+$, we have $\mathrm{RegExp}(T) = \Omega(T)$ *a.s. when $T \to \infty$.*

Proof Let $M \in \mathcal{M}^+$. By Theorem 36, $|\mathcal{K}^-| = \infty$ almost surely. Denote $(t_{k(i)})$ the enumeration of exploration times. Because M is recurrent, every policy is recurrent on M thus $\mathrm{Reach}(\pi, M, s) \cap \mathcal{Z}^-(M) \neq \emptyset$ if, and only if $g^\pi(M) < g^*(M)$, where s is an arbitrary state. From (8), we see that:

$$\mathbf{P}\Big(\lim_{t\to\infty} \min\{N_t(s,a) : \pi_t(a|s) > 0\} = \infty\Big) = 1. \tag{56}$$

It follows that $\liminf(t_{k(i)+1} - t_{k(i)}) = \infty$. Below, we write μ^π the asymptotic empirical measure of play of $\pi \in \Pi$, given by $\mu^\pi(z|s; M) := \lim \frac{1}{T}\mathbf{E}_s^{\pi,M}[\sum_{t=1}^T \mathbf{1}(Z_t = z)]$, i.e., $\mu^\pi(z|s; M)$ is the average amount of time that π spends playing z under M starting from $s \in \mathcal{S}$. In particular, for all $T \geq 0$, we have:

$$\mathrm{RegExp}(T)$$

$$
\begin{aligned}
&\overset{(*)}{\geq} \limsup_{i\to\infty} \left(\mathbf{E}^{(\pi_t),M} \left[\sum_{t=t_{k(i)}}^{t_{k(i)}+T-1} \Delta^*(Z_t; M) \right] \right) \\
&\overset{(\dagger)}{\geq} \limsup_{i\to\infty} \Big(\mathbf{E}^{(\pi_t),M} \Big[T \min(\mu^{\pi_{t_{k(i)}}}(M)) \Delta^*_{\min}(M) - D(\pi_{t_{k(i)}}; M) \Big] \Big) \\
&\overset{(\ddagger)}{\geq} T\alpha - \beta
\end{aligned}
$$

where $(*)$ follows from by definition; $(\dagger)$ is obtained by writing the Poisson equation of $\pi_{t_{k(i)}}$ for the reward function $f_i(z) = \mathbf{1}(z = z_i)$ where z_i is any sub-optimal pair played by $\pi_{t_{k(i)}}$, and $D(\pi_{t_{k(i)}}; M)$ is the span of the bias function of $\pi_{t_{k(i)}}$ under f; and $(\ddagger)$ introduces $\alpha := \min_\pi \min(\mu^\pi(M)) \Delta^*_{\min}(M) > 0$ and $\beta := \max_\pi D(\pi; M) < \infty$. ■

Appendix G. The class of explorative MDPs

In this appendix, we study the spaces of Markov decision processes for which the regret of exploration (Definition 3) is well-defined. By construction, the regret of exploration is well-defined if, and only if the number of exploration times (Definition 2) is infinite and we naturally investigate when this is exactly the case. As motivated in Section 3.2, we need a technical accommodation: We focus on learning algorithms with sub-linearly many episodes. These algorithms are these for which the performance in practice is actually comparable to the policies from which they pick actions. Under this technical assumption, explorative Markov decision processes correspond to those that, intuitively, cannot be learned by playing the optimal policy only. In Theorem 36, we provide four characterizations of explorative environments. Every one of them is of a different nature, that we explain below.

Theorem 36 (Characterizations of explorative MDPs) *Let $\mathcal{M} \equiv \prod_{z\in\mathcal{Z}}(\mathcal{R}_z \times \mathcal{P}_z)$ be a **convex** ambient space in product form. Let $M \in \mathcal{M}$ be a non-degenerate Markov decision process. The following assertions are equivalent:*

1. *$M \notin \mathcal{M}^+$, i.e., M is not explorative;*
2. *M has empty **confusing set**, i.e., $\mathrm{Cnf}(M) = \emptyset$, see (57);*
3. *There exists a consistent learner Λ, i.e., such that $\mathrm{Reg}(T; M', \Lambda) = \mathrm{o}(T^\epsilon)$ for all $M' \in \mathcal{M}$ and $\epsilon > 0$, such that $\mathrm{Reg}(T; M, \Lambda) = \mathrm{o}(\log(T))$;*
4. *There exists a robust learner Λ, i.e., such that $\sup_{M'\in\mathcal{M}} \mathrm{Reg}(T; M', \Lambda) = \mathrm{o}(T)$, such that $\mathrm{Reg}(T; M, \Lambda) = \mathrm{O}(1)$;*

Each characterization in Theorem 36 is to be understood as follows.

The first characterization (*1.*) is simply the definition from the main text (Definition 5): A Markov decision process is explorative if the regret of exploration of no-regret algorithms with sub-linearly many episodes is well-defined. The other characterizations relate the concept of explorative MDPs to more common settings. The second characterization (*2.*) is computational. A Markov decision process is explorative if its confusing set, given by[5]

$$\mathrm{Cnf}(M) := \left\{ M^\dagger \in \mathcal{M} : M \ll M^\dagger, M = M^\dagger \text{ on } \mathcal{Z}^{**}(M), \Pi^*(M^\dagger) \cap \Pi^*(M) = \emptyset \right\}, \quad (57)$$

is empty. The confusing set is a natural object that arises in instance dependent approaches to regret minimization, see Lai and Robbins (1985); Burnetas and Katehakis (1997); Tranos and Proutiere (2021); Boone and Maillard (2025)—although as shown by the fourth characterization in Theorem 36, it is also linked to instance independent frameworks. In practice, the second characterization provides a simple way to test if a Markov decision process is explorative. The third characterization (*3.*) is relevant to regret minimization in the instance dependent setting. It is known that most MDPs are such that consistent learners satisfy $\mathrm{Reg}(T; M, \Lambda) = \Omega(\log(T))$. However, non-explorative MDPs are those for which it is somehow possible to have regret $\mathrm{o}(\log(T))$. The fourth characterization (*4.*) is relevant to regret minimization in the problem

5. The notation "$M \ll M^\dagger$" is about the absolute continuity of M with respect to $M^\dagger$. It means that $r(z) \ll r^\dagger(z)$ and $p(z) \ll p^\dagger(z)$ for all $z \in \mathcal{Z}$.

independent (or minimax) setting, stating that we can find robust learners with bounded regret on M.

Outline The main goal of this appendix is to establish Theorem 36. Theorem 36 provides many characterizations of $\mathcal{M}^+$, but each of them come with long-winded and exhausting proofs, especially if written in full details. So, we begin by providing some intuition on why explorative MDPs are necessary in the first place. In Section G.1, we describe a simple Markov decision process that is *not* explorative, and for which we show that the famous `UCRL2` of Auer et al. (2009) has bounded regret. Because the regret of exploration is the object of focus in this work, we believe that the most important part of Theorem 36 is to show that if $\text{Cnf}(M) \neq \emptyset$, then M is explorative and the regret of exploration is well-defined. Therefore, Section G.2 is dedicated to a fully detailed proof of this result, hence providing a simple condition under which the analysis of the regret of exploration is meaningful in the first place. The remaining equivalences of Theorem 36 are bonus. In Section G.3, we show that Markov decision processes with non-empty confusing set cannot be learned trivially: the regret of consistent learning algorithms must grow logarithmically with T (Proposition 39) and the regret of robust learning algorithms must be unbounded (Proposition 40). We consider the converse results in Section G.4, showing that when the confusing set of M is empty, then M is non-explorative, that there are robust algorithms with bounded regret on M, and consistent learning algorithms with sub-logarithmic regret on M. This completes the proof of Theorem 36.

In Section G.5, we discuss how common the property "$M \in \mathcal{M}^+$" is. We explain that it depends on the amount of structure of $\mathcal{M}^+$: $\mathcal{M} \backslash \mathcal{M}^+$ can be large if $\mathcal{M}$ is heavily structured, and small otherwise. In particular, we show that all non-degenerate interior models (Assumption 4) are explorative in the ambient set of all Markov decision processes (see Proposition 44).

G.1. An example of non-explorative Markov decision process

Not every Markov decision process is explorative, and as a matter of fact, they are easily found. Such MDPs can be learned within a finite exploration phase because efficient learning algorithm can eliminated sub-optimal policies just by having information on optimal ones. In Figure 10, we provide an example of a non-explorative environment.

Notations and intuition Let p be the transition kernel of M as described by Figure 10. Let

$$\mathcal{M} := \left\{ M' : \forall z \in \mathcal{Z},\ p'(z) = p(z) \text{ and } r(z) \in [0,1] \right\}$$

be the set of Bernoulli-reward Markov decision processes with the same transition structure than M. On M, there are two policies π^* and π^-, respectively looping on the 5-cycle or the 3-cycle. By looping on the 5-cycle, the algorithm learns its rewards very well, hence can claim that the 3-cycle's average reward is upper bounded by $\frac{1+1+0.1+\varepsilon_t}{3}$ because unknown rewards are bounded by 1. This is smaller than a lower bound for the 5-cycle $\frac{0.9+0.9+0.9+0.9+0.1-\varepsilon_t}{5}$ (where ε_t is vanishing with t). Therefore, the algorithm has no need to visit the dashed arrows infinitely often. What we have just justified is that: (1) there is no $M' \in \mathcal{M}$ that coincide with M on the 5-cycle, which is such that $\Pi^*(M') \neq \{\pi^*\}$, meaning that $\text{Cnf}(M) = \emptyset$ and echoing the characterization (2.) of Theorem 36; (2) the property $\text{Cnf}(M) = \emptyset$ can be exploited by some optimistic algorithm to have uncommonly small regret on M specifically, echoing

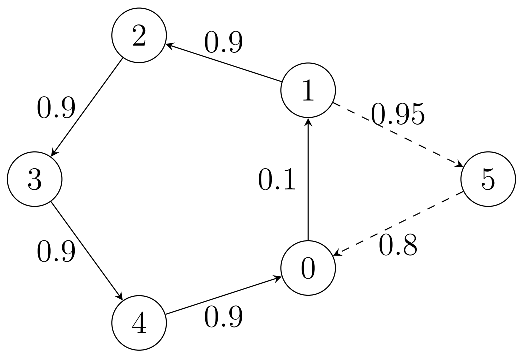


Figure 10: An example of a non-explorative Markov decision process. From all states, there is a single choice of action excepted at the marked state $(*)$ where there are two actions (dashed and solid lines). Choices of action deterministically lead to the state indicated by the arrow. Rewards are Bernoulli, with means indicated by the labels.

the characterizations (*3.*) and (*4.*) of Theorem 36. We show this second point more formally with `UCYCLE` Ortner (2010), a variant of `UCRL2` that is specialized to learning deterministic transition Markov decision processes such as in Figure 10.

Algorithm 2 `UCYCLE`: `UCRL2` for deterministic transition models

$$\mathcal{R}(t) := \prod_{z \in \mathcal{Z}} \left\{ \tilde{r}(z) \in [0,1] : \tilde{r}(z) \leq \hat{r}_t(z) + \sqrt{\tfrac{2\log(SAt)}{N_z(t)}} \right\} \quad \text{and} \quad \mathcal{P}(t) := \{p\}.$$

$k \leftarrow 0$, initialize π^0;
for $t = 0, 1, \ldots$ **do**
  if (`DT`) triggers **then**
    $k \leftarrow k+1$; $t_k \leftarrow t$;
    $\pi_{t_k} \leftarrow \mathtt{EVI}(\mathcal{M}(t_k), 0, 0^{\mathcal{S}})$;
  end if
  Set $\pi_t \leftarrow \pi_{t_k}$ and play $A_t \leftarrow \pi_t(S_t)$.
end for

To be absolutely accurate, Algorithm 2 is not exactly the same algorithm as Ortner (2010), that we have simplified to ease the exposition. It is essentially the same algorithm as `UCRL2` of Auer et al. (2009) with prior information on the transition kernel of M. The proof of its model independent regret guarantees on $\mathcal{M}$ can be directly adapted from Ortner (2010), or from our own Appendix B by removing the error terms relative to the learning of transition kernels.

Proposition 37 *`UCYCLE` (see Algorithm 2) is robust on $\mathcal{M}$, with*

$$\sup_{M' \in \mathcal{M}} \operatorname{Reg}(T; M', \mathtt{UCYCLE}) = \operatorname{O}\left(\sqrt{|\mathcal{Z}| T \log(T)}\right).$$

Moreover, for M' as given by Figure 10, we have $\operatorname{Reg}(T; M, \mathtt{UCYCLE}) = \operatorname{O}(1)$.

Remark The example of Figure 10 is robust to reward perturbation. It means that it is non-degenerate in the sense of Definition 4. It follows that by identifying $M' \in \mathcal{M}$ as a reward vector $r' \in [0,1]^{\mathcal{Z}}$, it means the set $\mathcal{M}'$ of $M' \in \mathcal{M}$ where the 5-cycle dominates the 3-cycle in the fashion described above has positive Lebesgue measure. As $\mathcal{M}^+ \supseteq \mathcal{M}'$, it follows that a large portion of $\mathcal{M}$ is made of non-explorative Markov decision processes: By picking r' uniformly at random, the obtained MDP is non-explorative with positive probability.

Proof of Proposition 37 The assertion on the model independent regret guarantees is well-known, see Ortner (2010) and Appendix B. We focus on proving that it has bounded regret on the model M given in Figure 10.

The model M is identified with its reward vector r. Remark that the only pair with positive Bellman gap is $(1, \dagger)$ with Bellman gap $\Delta^*(1, \dagger) \le 1$. So, the regret is upper-bounded by $|\{t \le T : \pi_t = \pi^-\}|$. We are left to bound how many times the sub-optimal policy π^- is played. A simple property induced by the doubling trick (DT) is that $t_{k+1} \le 3t_k$. So, if $\pi_t = \pi^-$, then there exists $t' \in [\frac{1}{3}t, t]$ such that π^- is the result of EVI, i.e., $g^{\pi^-}(\mathcal{M}(t')) > g^{\pi^*}(\mathcal{R}(t')) + \frac{1}{t'}$.

Let $c := 3 \cdot \frac{0.9+0.9+0.9+0.9+0.1}{5} - 2 = 0.22$, which is the threshold on the reward that one should have on $(0, *)$ in order to make π^- better than π^*. Since

$$g^{\pi^-}(\mathcal{M}(t)) \le \frac{1}{3}\left(2 + \hat{r}_t(0,*) + \sqrt{\frac{2\log(|\mathcal{Z}|t)}{N_{0,*}(t)}}\right) \text{ and } g^{\pi^*}(\mathcal{M}(t)) \ge \mathbf{1}(r \in \mathcal{R}(t)) g^{\pi^*}(M),$$

we have:

$$\begin{aligned}
(*) &:= \mathbf{E}\big|\{t \ge 1 : \pi_t \ne \pi^-\}\big| \\
&\le 300 + \sum_{t \ge 300} \sum_{t'=t/3}^{t} \mathbf{P}\Big(g^{\pi^-}(\mathcal{M}(t')) > g^{\pi^*}(\mathcal{M}(t')) + \tfrac{1}{100}\Big) \\
&\le 300 + \sum_{t \ge 300} \sum_{t'=t/3}^{t} \left(\mathbf{P}\left(\hat{r}_{t'}(0,*) + \sqrt{\frac{2\log(|\mathcal{Z}|t')}{N_{0,*}(t')}} > 0.21\right) + \mathbf{P}(M \notin \mathcal{M}(t'))\right).
\end{aligned} \tag{58}$$

For the first term, remark that $N_{0,*}(t') \ge \frac{1}{5}t'$ almost surely when $t' \ge 5$. For t' large enough so that $\sqrt{10\log(|\mathcal{Z}|t')/t'} < 0.01$, we have

$$\begin{aligned}
(**) &:= \mathbf{P}\left(\hat{r}_{t'}(0,*) + \sqrt{\frac{2\log(|\mathcal{Z}|t')}{N_{0,*}(t')}} > 0.21\right) \\
&\le \mathbf{P}\left(\exists n \in [\tfrac{1}{5}t', t'] : N_{0,*}(t') = n, \hat{r}_{t'}(0,*) + \sqrt{\tfrac{2\log(|\mathcal{Z}|t')}{n}} > 0.21\right) \\
&\le \sum_{n=\frac{1}{5}t'}^{\infty} \mathbf{P}(N_{0,*}(t') = n, \hat{r}_{t'}(0,*) - r(0,*) > 0.2) \\
&\overset{(\dagger)}{\le} \sum_{n=\frac{1}{5}t'}^{\infty} \exp\left(-\frac{8}{10000}n\right) = \frac{\exp\big(-\frac{1}{6250}t'\big)}{1 - \exp\big(-\frac{1}{1250}\big)} = \mathrm{O}\big(\exp(-\tfrac{1}{6250}t')\big)
\end{aligned}$$

where $(\dagger)$ follows from Azuma-Hoeffding's inequality. For the second term, we have

$$\begin{aligned}
\mathbf{P}(M \notin \mathcal{M}(t')) &= \mathbf{P}\left(\exists z \in \mathcal{Z}, |\hat{r}_{t'}(z) - r_z| > \sqrt{\frac{2\log(|\mathcal{Z}|t')}{N_z(t')}}\right) \\
&\le \sum_z \sum_{n=1}^{\infty} \mathbf{P}\left(N_z(t') = n, |\hat{r}_{t'}(z) - r(z)| > \sqrt{\frac{2\log(|\mathcal{Z}|t')}{n}}\right) \\
&\overset{(\dagger)}{\le} 2|\mathcal{Z}| \sum_{n=1}^{\infty} \exp(-4\log(|\mathcal{Z}|t') \cdot n) \\
&\le \frac{2|\mathcal{Z}|}{(t'|\mathcal{Z}|)^4} \cdot \frac{1}{1-(t'|\mathcal{Z}|)^{-4}} \le \frac{4}{|\mathcal{Z}|^3 t'^4} = \mathrm{O}(t'^{-4}).
\end{aligned}$$

where $(\dagger)$ follows from Azuma-Hoeffding's inequality. Overall, injecting it all in (58), we obtain $\mathbf{E}|\{t \ge 1 : \pi_t \ne \pi^-\}| < \infty$. We conclude accordingly that $\mathrm{Reg}(T; M, \texttt{UCYCLE}) = \mathrm{O}(1)$. ■

G.2. MDPs with non-empty confusing sets are explorative

In this section, we show that Markov decision processes with non-empty confusing set are explorative, see Proposition 38. This is *(1.)* $\Rightarrow$ *(2.)* in Theorem 36 for which we show the transposition $\neg$*(2.)* $\Rightarrow$ $\neg$*(1.)*. This result is absolutely necessary to justify that the analysis of the regret of exploration is formally based.

Proposition 38 *Let $M \in \mathcal{M}$. If* $\mathrm{Cnf}(M) \ne \emptyset$*, then every no-regret algorithm Λ with sub-linearly many episodes has infinitely many exploration episodes on M, almost surely.*

Proof sketch Recall that, by definition, $k \ge 1$ is an **exploration episode** if (1) $g^*(M) = g(\pi^k, S_{t_k}, M)$ and (2) $\mathrm{Reach}(\pi^k, S_{t_k}, M) \cap \mathcal{Z}^-(M) \ne \emptyset$, see (Definition 2). In order to show that there are infinitely many exploration episodes, we have to show that the learning process alternates infinitely often between periods of times when the played policy is gain optimal, and others when there is a reachable sub-optimal pair. (**STEP 1**) is a preliminary technical fact. In (**STEP 2**), we show with (62) that the process is infinitely many times on the recurrent part of a gain optimal policy. In (**STEP 2**), we show with (63) that the process must play sub-optimal pairs infinitely often. Combining both in (**STEP 4**), we show that the number of exploration times is infinite, and each are finite with probability one.

Notations For $\pi \in \Pi$, we write $\mathrm{Rec}(\pi)$ the set of states that are recurrent under π on M, i.e., $s \in \mathcal{S}$ such that $\mathbf{P}_s^{\pi,M}(\forall m, \exists n \ge m : S_n = s) = 1$.

(**STEP 1**) *For every model $M \in \mathcal{M}$, there exists a constant $C(M) > 0$ such that whatever the learning algorithm, we have:*

$$\mathbf{E}^M\left[\sum_{t=1}^{T}(g^*(S_t, M) - g^{\pi_t}(S_t, M) + \mathbf{1}(S_t \notin \mathrm{Rec}(\pi_t)))\right] \le \mathrm{Reg}(T; M) + C(M)\mathbf{E}^M|\mathcal{K}(T)|. \tag{59}$$

Proof In the proof below, we drop the dependency in M in the notations. If $\pi \in \Pi$, we denote $\mathrm{Rec}(\pi)$ the recurrent states of π in M. We have:

$$
\begin{aligned}
(*) &= \mathrm{Reg}(T; M) \\
&= \mathbf{E}\left[\sum_{t=1}^{T} \Delta^*(Z_t)\right] \\
&\overset{(\dagger)}{=} \mathbf{E}\left[\sum_{t=1}^{T} (g^*(S_t) - r(Z_t) + (e_{S_t} - p(Z_t))h^*)\right] \\
&\geq \mathbf{E}\left[\sum_{t=1}^{T} (g^* - r(Z_t))\right] - \mathrm{sp}(h^*) \\
&\overset{(\ddagger)}{\geq} \underbrace{\mathbf{E}\left[\sum_{k=1}^{|\mathcal{K}(T)|} \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}(S_t \in \mathrm{Rec}(\pi_t))(g^*(S_t) - r(Z_t))\right]}_{\mathrm{A}} - \underbrace{\mathbf{E}\left[\sum_{k=1}^{|\mathcal{K}(T)|} \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}(S_t \notin \mathrm{Rec}(\pi_t))\right]}_{\mathrm{B}} \\
&\quad - \mathrm{sp}(h^*)
\end{aligned}
$$

where $(\dagger)$ uses the Bellman equation $h^*(s) + g^*(s) = r(s,a) + p(s,a)h^* + \Delta^*(s,a)$, and $(\ddagger)$ uses that $g^*(S_t) - r(Z_t) \geq -1$. We bound A and B separately. Let $D_* := \max_\pi \max_s \mathbf{E}_s^\pi[\inf\{t \geq 1 : S_t \in \mathrm{Rec}(\pi)\}] < \infty$ be the worst hitting time to a recurrent class in M. We have:

$$
\mathrm{B} = \mathbf{E}\left[\sum_{k=1}^{|\mathcal{K}(T)|} \sum_{t=t_k}^{t_{k+1}-1} \inf\big\{t > t_k : S_t \in \mathrm{Rec}(\pi^k)\big\}\right] \leq D_* \mathbf{E}[|\mathcal{K}(T)|]. \tag{60}
$$

Meanwhile, introduce $t'_k := t_{k+1} \wedge \inf\{t > t_k : S_t \in \mathrm{Rec}(\pi^k)\}$ and $H := \max_\pi \mathrm{sp}(h^\pi) < \infty$ the worst bias span. We have:

$$
\begin{aligned}
\mathrm{A} &= \mathbf{E}\left[\sum_{k=1}^{|\mathcal{K}(T)|} \sum_{t=t'_k}^{t_{k+1}-1} (g^*(S_t) - r(Z_t))\right] \\
&\overset{(\dagger)}{=} \mathbf{E}\left[\sum_{k=1}^{|\mathcal{K}(T)|} \sum_{t=t'_k}^{t_{k+1}-1} \Big(g^*(S_t) - g^{\pi^k}(S_t) + (p(Z_t) - e_{S_t})h^{\pi^k}\Big)\right] \\
&\geq \mathbf{E}\left[\sum_{k=1}^{|\mathcal{K}(T)|} \sum_{t=t'_k}^{t_{k+1}-1} (g^*(S_t) - g^{\pi_t}(S_t))\right] - H\mathbf{E}[|\mathcal{K}(T)|] \\
&\overset{(\ddagger)}{\geq} \mathbf{E}\left[\sum_{t=1}^{T} (g^*(S_t) - g^{\pi_t}(S_t))\right] - H\mathbf{E}[|\mathcal{K}(T)|]
\end{aligned} \tag{61}
$$

where (†) uses the Poisson equation $h^{\pi^k}(s) + g^{\pi^k}(s) = r(s, \pi^k(s)) + p(s, \pi^k(s))h^{\pi^k}$ and (‡) that $g^*(S_t) \geq g^{\pi_t}(S_t)$ for all $t \geq 1$. Combining (60) and (61), we get:

$$\mathbf{E}\left[\sum_{t=1}^{T}(g^*(S_t) - g^{\pi_t}(S_t))\right] + \mathbf{E}\left[\sum_{t=1}^{T}\mathbf{1}(S_t \notin \mathrm{Rec}(\pi_t))\right] \leq \mathrm{Reg}(T) + (2D_* + H)\mathbf{E}[|\mathcal{K}(T)|].$$

Conclude the proof by setting $C := 2D_* + H < \infty$. ■

(STEP 2) *Assume that the algorithm is no-regret and has sub-linearly many episodes in expectation. Then:*

$$\mathbf{P}(\forall T, \exists t \geq T : g^*(S_t, M) = g^{\pi_t}(S_t, M) \textit{ and } S_t \in \mathrm{Rec}(\pi_t)) = 1. \tag{62}$$

Proof Assume on the contrary that $\mathbf{P}(\forall T, \exists t \geq T : g^*(S_t, M) = g^{\pi_t}(S_t, M) \wedge S_t \in \mathrm{Rec}(\pi_t)) = 1 - \delta$ with $\delta > 0$. Accordingly, there exists $T_0 \geq 1$ such that:

$$\mathbf{P}\big(\forall t \geq T_0, g^*_{S_t}(M) > g^{\pi}_t(S_t, M) \text{ or } S_t \notin \mathrm{Rec}(\pi_t)\big) \geq \tfrac{1}{2}\delta.$$

Let $\Delta_g := \min\{g^*(s, M) - g^\pi(s, M) : \pi \in \Pi, s \in \mathcal{S}, g^*(s, M) > g^\pi(s, M)\}$ be the gain-gap of M. We have $\Delta_g \in (0, 1]$ and thus:

$$\begin{aligned}
(*) &:= \mathbf{E}\left[\sum_{t=1}^{T}(g^*(S_t, M) - g^{\pi_t}(S_t, M))\right] + \mathbf{E}\left[\sum_{t=1}^{T}\mathbf{1}(S_t \notin \mathrm{Rec}(\pi_t))\right] \\
&\geq \Delta_g \mathbf{E}\left[\sum_{t=1}^{T}\mathbf{1}(g^*(S_t, M) > g^{\pi_t}(S_t, M) \text{ or } S_t \notin \mathrm{Rec}(\pi_t))\right] \\
&\geq \Delta_g (T - T_0)\mathbf{P}(\forall t \geq T_0, g^*(S_t, M) > g^{\pi_t}(S_t, M) \text{ or } S_t \notin \mathrm{Rec}(\pi_t)) \\
&\geq \tfrac{1}{2}\Delta_g \delta (T - T_0) = \Omega(T).
\end{aligned}$$

Meanwhile, we know that $\mathrm{Reg}(T; M) = \mathrm{o}(T)$ and $\mathbf{E}[|\mathcal{K}(T)|] = \mathrm{o}(T)$, so that by **(STEP 1)** (59), we also have $(*) = \mathrm{o}(T)$, a contradiction. ■

(STEP 3) *If* $\mathrm{Cnf}(M) \neq \emptyset$*, then every no-regret algorithm satisfies*

$$\mathbf{P}^M(\forall T, \exists t > T : \Delta^*(Z_t) > 0) = 1. \tag{63}$$

Proof On the contrary, assume that $\mathbf{P}^M(\forall T, \exists t > T : \Delta^*(Z_t) > 0) = 1 - \delta$ with $\delta > 0$. Accordingly, there exists $m \geq 1$ such that:

$$\tfrac{1}{2}\delta \leq \mathbf{P}^M(\forall t > m : \Delta^*(Z_t) = 0) \leq \mathbf{P}^M\left(\forall t \geq 1 : \sum_{z \in \mathcal{Z}^-(M)} N_t(z) \leq m\right). \tag{64}$$

We show that $z \in \mathcal{Z}^-(M)$ can be changed to $z \notin \mathcal{Z}^{**}(M)$ in (64), see (65). To see this, introduce the reward function $f(z) := \mathbf{1}(z \in \mathcal{Z}^{**}(M))$ and let g^f, h^f and Δ^f the respective gain, bias and gap functions of the optimal policy π^* of M (defined by $\pi^*(s) = a$ the unique

$a \in \mathcal{A}(s)$ such that $(s,a) \in \mathcal{Z}^*(M)$) under reward function f and kernel $p(M)$. Remark that $g^f(s) = 1$ for all $s \in \mathcal{S}$ and that, by construction of π^*, $\Delta^f(z) = 0$ for all $z \in \mathcal{Z}^*(M)$. Denote $H^f := \mathrm{sp}(h^f) \vee \max_z |\Delta^f(z)|$. We have:

$$\begin{aligned}
&\sum_{z \in \mathcal{Z}^{**}(M)} N_z(T) \\
&= \sum_{t=1}^{T} f(Z_t) \\
&\overset{(\dagger)}{=} \sum_{t=1}^{T} \big(1 + (e_{S_t} - p(Z_t))h^f - \Delta^f(Z_t)\big) \\
&\geq T - H^f - \sum_{t=1}^{T} \Delta^f(Z_t) + \sum_{t=1}^{T} \big(e_{S_{t+1}} - p(Z_t)\big)h^f \\
&\overset{(\ddagger)}{\geq} T - H^f - H^f \sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathcal{Z}^*(M)) + \sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathcal{Z}^{**}(M))\big(e_{S_{t+1}} - p(Z_t)\big)h^f
\end{aligned}$$

where $(\dagger)$ uses the Bellman equation $1 + h^f(s) = f(s,a) + p^f(s,a)h^f + \Delta^f(s,a)$, and $(\ddagger)$ that $h^f(s) = 0$ for all $(s, \pi^*(s)) \in \mathcal{Z}^{**}(M)$. For $\mathcal{Z}' \subseteq \mathcal{Z}$, denote $N_T(\mathcal{Z}') := \sum_{z \in \mathcal{Z}'} N_T(z)$. The first sum is equal to $\sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathcal{Z}^*(M)) = N_T(\mathcal{Z}^-(M))$. The RHS of the above equation is bounded using a time-uniform Azuma-Hoeffding inequality (see (Bourel et al., 2020, Lemma 5)), showing that:

$$\mathbf{P}\left(\exists T \geq 1 : \begin{array}{c} \sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathcal{Z}^{**}(M))\big(e_{S_{t+1}} - p(Z_t)\big)h^f \\ < -H^f \sqrt{N_T(\mathcal{Z}^{**}(M)^c) \log\left(\frac{4N_{\mathcal{Z}^{**}(M)^c}(T)}{\delta}\right)} \end{array} \right) \leq \tfrac{1}{4}\delta$$

Using that $N_T(\mathcal{Z}^{**}(M)^c) = T - N_T(\mathcal{Z}^{**}(M))$, we obtain that, with probability at least $\frac{1}{4}\delta$, for all $T \geq 1$, we have:

$$\begin{aligned}
&T - N_T(\mathcal{Z}^{**}(M)^c) \\
&\geq T - H^f\big(1 + N_T(\mathcal{Z}^-(M))\big) - H^f \sqrt{N_T(\mathcal{Z}^{**}(M)^c) \log\left(\tfrac{4N_T(\mathcal{Z}^{**}(M)^c)}{\delta}\right)} \\
&\geq T - H^f(1+m) - H^f \sqrt{N_T(\mathcal{Z}^{**}(M)^c) \log\left(\tfrac{4N_T(\mathcal{Z}^{**}(M)^c)}{\delta}\right)}.
\end{aligned}$$

Rearranging terms, we get that with probability at least $\frac{1}{4}\delta$, for all $T \geq 1$, we have:

$$\begin{aligned}
&N_T(\mathcal{Z}^{**}(M)^c) \\
&\leq H^f\left(1 + m + \sqrt{N_T(\mathcal{Z}^{**}(M)^c) \log(N_T(\mathcal{Z}^{**}(M)^c))} + \sqrt{N_T(\mathcal{Z}^{**}(M)^c) \log\left(\tfrac{4}{\delta}\right)}\right).
\end{aligned}$$

Denoting $n := N_T(\mathcal{Z}^{**}(M))$, we have an equation of the form $n \leq \alpha + \beta\sqrt{n \log(n)} + \gamma\sqrt{n}$. For $n \geq 3$, $n \log(n) \geq n$ hence we can simplify the upper-bound to $n \leq \alpha + (\beta+\gamma)\sqrt{n \log(n)}$.

Dividing by $\log(n) \geq 1$ and setting $m := n/\log(n)$, we get $m \leq \alpha + (\beta+\gamma)\sqrt{m}$, and simple algebra leads to:

$$\frac{n}{\log(n)} = m \leq 2\big(\alpha + (\beta+\gamma)^2\big).$$

Further using $\log(n) \leq \sqrt{n}$, we get $n \leq 4(\alpha + (\beta+\gamma)^2)^2$. We conclude that there exists a constant m' such that:

$$\mathbf{P}^M\left(\forall t \geq 1, \sum_{z \notin \mathcal{Z}^{**}(M)} N_t(z) \leq m'\right) \geq \tfrac{1}{4}\delta. \tag{65}$$

Now that (65) is established, we finally derive a contradiction by relying on a change of measure argument. Let $M^\dagger \in \mathrm{Cnf}(M)$, which is non-empty by assumption. For short, the transition kernels and reward distributions of M (respectively $M^\dagger$) are denoted p and r (respectively $p^\dagger$ and $r^\dagger$). We introduce the log-likelihood-ratio of observations $H_t := (S_t, A_t, R_1, \ldots, A_{t-1}, R_{t-1}, S_t)$ as:

$$L(t) \equiv L(H_t) := \sum_{s,a} \sum_{i<t-1} \mathbf{1}(S_i = s, A_t = a) \log\left(\frac{r_{s,a}(R_i)}{r^\dagger_{s,a}(R_i)} \frac{p_{s,a}(S_{i+1})}{p^\dagger_{s,a}(S_{i+1})}\right).$$

It is known since Marjani et al. (2021) that if $\mathcal{E}$ is a $\sigma(H_t)$-measurable event, then $\mathbf{P}^{M^\dagger}(\mathcal{E}) = \mathbf{E}^M[\mathbf{1}(\mathcal{E})\exp(-L(t))]$. Since $M \ll M^\dagger$, there exists a constant $c > 0$ such that, for all $z \in \mathcal{Z}$, we have $\log[(r_z(\alpha)/r^\dagger_z(\alpha)) \cdot (p_z(s')/p^\dagger_z(s'))] \leq \log(c)$ with the convention $0/0 = 0$. For $z \in \mathcal{Z}^{**}(M)$, the LHS logarithm is null. Therefore, we have:

$$\begin{aligned}
&\mathbf{P}^{M^\dagger}\left(\sum_{z \notin \mathcal{Z}^{**}(M)} N_t(z) \leq m'\right) \\
&= \mathbf{E}^M\left[\mathbf{1}\left(\sum_{z \notin \mathcal{Z}^{**}(M)} N_t(z) \leq m'\right)\exp(-L(t))\right] \\
&\geq \mathbf{E}^M\left[\mathbf{1}\left(\sum_{z \notin \mathcal{Z}^{**}(M)} N_t(z) \leq m'\right)\exp\left(-\sum_{z \notin \mathcal{Z}^{**}(M)} N_t(z)\log(c)\right)\right] \\
&\geq c^{-m'}\mathbf{P}^M\left(\sum_{z \notin \mathcal{Z}^{**}(M)} N_t(z) \leq m'\right) \geq c^{-m'}\delta := \delta' > 0.
\end{aligned}$$

Accordingly, the algorithm has probability at least δ' to spend at most m' visits outside $\mathcal{Z}^{**}(M)$ when running on $M^\dagger$. This will be in contradiction $M^\dagger \in \mathrm{Cnf}(M)$ and the consistency of the algorithm. Indeed, since $M^\dagger \gg M$ coincides with M on $\mathcal{Z}^{**}(M)$, we see that the optimal policy π^* of M has unique recurrent class $\mathcal{Z}^{**}(M)$ in $M^\dagger$. Yet, $\pi^* \notin \Pi^*(M^\dagger)$, hence $\mathcal{Z}^{**}(M) \cap \mathcal{Z}^-(M^\dagger) \neq \emptyset$, i.e., there exists $z \in \mathcal{Z}^{**}(M)$ such that $\Delta^*(z; M^\dagger) > 0$. We further link the number of visits of this z to the total number of visits of $\mathcal{Z}^{**}(M)$ with the same technique that the one used to convert (64) to (65).

Introduce the reward function $f(z') := \mathbf{1}(z' = z)$, and let g^f, h^f, Δ^f be the gain, bias and gaps functions of the policy π^* in $M^\dagger$. There exists $\epsilon > 0$ such that $g^f(s) = \epsilon$ for all $s \in \mathcal{S}$. Letting $C := \mathrm{sp}(h^f) \vee \max_{z'} |\Delta^f(z')| < \infty$. For all $T \geq 1$, we have

$$\begin{aligned} N_T(z) = \sum_{t=1}^{T} f(Z_t) &= \sum_{t=1}^{T} \big(\epsilon + (e_{S_t} - p(Z_t))h^f - \Delta^f(Z_t)\big) \\ &\geq T\epsilon - C - CN_{\mathcal{Z}^{**}(M)^c}(T) + \sum_{t=1}^{T} \big(e_{S_{t+1}} - p(Z_t)\big)h^f \\ &\overset{(\dagger)}{\geq} T\epsilon - C(1+m') - C\sqrt{T\log\left(\tfrac{2T}{\delta'}\right)} \sim T\epsilon. \end{aligned}$$

where $(\dagger)$ holds with probability $\frac{1}{2}\delta' > 0$ uniformly for $T \geq 1$, by invoking a time-uniform Azuma-Hoeffding (see (Bourel et al., 2020, Lemma 5)) to lower-bound the right-hand martingale. We accordingly obtain, when $T \to \infty$,

$$\mathrm{Reg}(T; M^\dagger) \gtrsim \tfrac{1}{2}\epsilon\delta'\Delta^*(z; M^\dagger)T = \Omega(T). \tag{66}$$

So (66) is in contradiction with the consistency of the algorithm. ■

(**STEP 4**) *If the algorithm is no-regret, has sub-linearly many episodes, then for all $M \in \mathcal{M}$ such that* $\mathrm{Cnf}(M) \neq \emptyset$*, we have:*

$$\mathbf{P}^M\big(\forall T, \exists t \geq T : g^*(M) = g^{\pi_{t-1}}(S_{t-1}, M) \textit{ and } \mathrm{Reach}(\pi_t, S_t, M) \cap \mathcal{Z}^-(M) \neq \emptyset\big) = 1. \tag{67}$$

Moreover, the stopping times t enumerating times such that $g^*(S_{t-1}, M) = g^{\pi_{t-1}}(S_{t-1}, M)$ and $\mathrm{Reach}(\pi_t, S_t, M) \cap \mathcal{Z}^-(M) \neq \emptyset$ are exploration times; Hence there are infinitely many of them with probability one.

Proof This is obtained by combining (62) of (**STEP 2**) and (63) of (**STEP 3**). We have:

$$\mathbf{P}^M(\forall T, \exists t \geq T : g^*(S_{t-1}, M) = g^{\pi_t}(S_t, M) \text{ and } S_t \in \mathrm{Rec}(\pi_t)) = 1, \text{ and} \tag{68}$$

$$\mathbf{P}^M\big(\forall T, \exists t \geq T : \mathrm{Reach}(\pi_t, S_t, M) \cap \mathcal{Z}^-(M) \neq \emptyset\big) = 1. \tag{69}$$

By non-degeneracy of M, if $g^*(S_t, M) = g^{\pi_t}(S_t, M)$ and $S_t \in \mathrm{Rec}(\pi_t)$, then $\mathrm{Reach}(\pi_t, S_t, M) = \mathcal{Z}^{**}(M)$ which is disjoint from $\mathcal{Z}^-(M)$. Define:

$$\begin{aligned} \tau_1 &:= \inf\{t \geq 1 : g^*(S_t, M) = g^{\pi_t}(S_t, M) \text{ and } S_t \in \mathrm{Rec}(\pi_t)\}, \\ \tau_{2i} &:= \inf\big\{t > \tau_{2i-1} : \mathrm{Reach}(\pi_t, S_t, M) \cap \mathcal{Z}^-(M) \neq \emptyset\big\}, \\ \tau_{2i+1} &:= \inf\{t > \tau_{2i} : g^*(S_t, M) = g^{\pi_t}(S_t, M) \text{ and } S_t \in \mathrm{Rec}(\pi_t)\} \end{aligned}$$

Then (τ_i) is an increasing sequence of stopping times, and by (68) (69) applied in tandem, we show by induction that $\mathbf{P}^M(\tau_i < \infty) = 1$ for all $i \geq 1$. By non-degeneracy of M, at $t = \tau_{2i+1}$, the current policy is gain optimal and the process is currently on the optimal class $\mathcal{Z}^{**}(M)$. Because $\mathcal{Z}^{**}(M)$ is the disjoint union of sink components of $\mathcal{Z}^*(M)$, hence the only way to exit $\mathcal{Z}^{**}(M)$ is by playing a $z \in \mathcal{Z}^-(M)$. Therefore, we see that for $t = \tau_{2i}$, we must have $g^*(S_{t-1}, M) = g^{\pi_{t-1}}(S_{t-1}, M)$ with $\pi_{t-1} \neq \pi_t$. Accordingly, every τ_{2i} are change of episodes that are exploration episodes. ■

This proves Proposition 38. ■

G.3. Instance (in)dependent regrets for MDPs with non-empty confusing sets

In this section, we show (*3.*) $\Rightarrow$ (*2.*) and (*4.*) $\Rightarrow$ (*2.*) in Theorem 36 by showing the transpositions $\neg$(*2.*) $\Rightarrow$ $\neg$(*3.*) in Proposition 39 and $\neg$(*2.*) $\Rightarrow$ $\neg$(*4.*) in Proposition 40. In the statements below, we borrow the terminology introduced by Theorem 36. A learning algorithm Λ is said **consistent** (on $\mathcal{M}$) if $\mathrm{Reg}(T; M, \Lambda) = \mathrm{o}(T^\epsilon)$ for all $\epsilon > 0$ and $M \in \mathcal{M}$. A learning algorithm Λ is said **robust** (relatively to $\mathcal{M}$) if $\sup_{M' \in \mathcal{M}} \mathrm{Reg}(T; M', \Lambda) = \mathrm{o}(T)$.

Proposition 39 (Consistent algorithms) *Let $M \in \mathcal{M}$ be such that* $\mathrm{Cnf}(M) \neq \emptyset$*. Then every consistent learning algorithm Λ satisfies* $\mathrm{Reg}(T; M, \Lambda) = \Omega(\log(T))$.

Proof This is a consequence of (Boone and Maillard, 2025, Corollary 7), that shows that every consistent learning algorithm Λ is such that, for all $M^\dagger \in \mathrm{Cnf}(M)$, we have:

$$\mathbf{E}^{M,\Lambda}\left[\sum_{z \in \mathcal{Z}} N_z(T) \mathrm{KL}(q(z)||q^\dagger(z))\right] \geq \log(T) + \mathrm{o}(\log(T))$$

where $q(z) = (r(z), p(z))$ is the reward-kernel tuple.

Fix $M^\dagger \in \mathrm{Cnf}(M)$. Let $c := \max\{\mathrm{KL}(q(z)||q^\dagger(z)) : q(z) \neq q^\dagger(z)\}$, that satisfies $c < \infty$ since $M \ll M^\dagger$. By definition of $\mathrm{Cnf}(M)$, $M = M^\dagger$ coincide on $\mathcal{Z}^{**}(M)$, so $\mathrm{KL}(q(z)||q^\dagger(z)) = 0$ for all $z \in \mathcal{Z}^{**}(M)$. Together with $\mathbf{E}[\sum_{z \in \mathcal{Z}} N_z(T) \mathrm{KL}(q(z)||q^\dagger(z))] \leq |\mathcal{Z}| \max_{z \in \mathcal{Z}} \mathbf{E}[N_z(T)] \mathrm{KL}(q(z)||q^\dagger(z))$, we deduce that:

$$\forall T \geq 1, \qquad \max_{z \notin \mathcal{Z}^{**}(M)} \mathbf{E}^{M,\Lambda}[N_z(T)] \geq \frac{\log(T) + \mathrm{o}(\log(T))}{|\mathcal{Z}|c}.$$

Now, by (Boone and Maillard, 2025, Lemma 8), there exist constants $\alpha, \beta > 0$ such that $\mathbf{E}^{M,\Lambda}[\sum_{t=1}^T \mathbf{1}(Z_t \notin \mathcal{Z}^{**}(M))] \leq \alpha \mathrm{Reg}(T; M, \Lambda) + \beta$ for all $T \geq 1$, so

$$\mathrm{Reg}(T; M, \Lambda) \geq \frac{\max_{z \notin \mathcal{Z}^{**}(M)} \mathbf{E}^{M,\Lambda}[N_z(T)] - \beta}{\alpha} \geq \frac{\log(T) + \mathrm{o}(\log(T))}{|\mathcal{Z}|c\alpha},$$

hence $\mathrm{Reg}(T; M, \Lambda) = \Omega(\log(T))$. ■

Proposition 40 (Robust algorithms) *Let $M \in \mathcal{M}$ be such that* $\mathrm{Cnf}(M) \neq \emptyset$*. Then every robust learning algorithm Λ satisfies* $\mathrm{Reg}(T; M, \Lambda) = \omega(1)$*, i.e.,* $\mathrm{Reg}(T; M, \Lambda) \to \infty$.

Proof This is a direct consequence of (**STEP 3**) of the proof of Proposition 38, see (63). Indeed, robust algorithms are by no-regret. So, by (63), the function given by

$$f(T) := \inf_{T' \geq T} \left\{ \mathbf{P}^{M,\Lambda}\left(\exists t \in \{T, \ldots, T'-1\} : \Delta^*(Z_t) > 0\right) \geq \frac{1}{2} \right\}$$

satisfies $T + 1 \leq f(T) < \infty$ for all $T \geq 0$. Introduce the (deterministic) sequence $T_1 := 1$ and $T_{k+1} := f(T_k)$, and introduce its pseudo-inverse $g(T) := \sup\{k \geq 1 : T_{k+1} \leq T\}$. Since $f(T) < \infty$, we have $T_k \to \infty$ and $g(T) \to \infty$. For $T \geq T_2$, we have:

$$\mathrm{Reg}(T; M, \Lambda) = \mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^T \Delta^*(Z_t)\right]$$

$$
\begin{aligned}
&\geq \sum_{k=1}^{g(T)} \mathbf{E}^{M,\Lambda}\left[\sum_{t=T_k}^{T_{k+1}-1} \Delta^*(Z_t)\right] \\
&\overset{(\dagger)}{\geq} \sum_{k=1}^{g(T)} c\mathbf{P}^{M,\Lambda}\big(\exists t \in \{T_k, \ldots, T_{k+1}-1\} : \Delta^*(Z_t) > 0\big) \overset{(\ddagger)}{\geq} \frac{c \cdot g(T)}{2}
\end{aligned}
$$

where $c := \min\{\Delta^*(z) : \Delta^*(z) > 0\} > 0$ is the minimum positive Bellman gap. We have $\frac{cg(T)}{2} \to \infty$ when $T \to \infty$, hence the conclusion. ■

G.4. MDPs with empty confusing sets are non-explorative

In this section, we show that if $\text{Cnf}(M) = \emptyset$, then M is non-explorative, there exists a consistent learning algorithm Λ such that $\text{Reg}(T; M, \Lambda) = \text{o}(\log(T))$ and there exists a robust learning algorithm Λ' such that $\text{Reg}(T; M, \Lambda') = \text{O}(1)$. This corresponds to *(2.)* $\Rightarrow$ *(1.)*, *(3.)*, and *(4.)* in Theorem 36 hence completing the proof of all the equivalences. The implication *(2.)* $\Rightarrow$ *(4.)*, stating the existence of a robust learning algorithm Λ with $\text{Reg}(T; M, \Lambda) = \text{O}(1)$ is done first, with Proposition 41. The proof is constructive, as we introduce a biased variant of KLUCRL Filippi et al. (2010) that is specialized to have bounded regret on M, see Algorithm 3. We prove *(2.)* $\Rightarrow$ *(1.)*, i.e., that M is non-explorative, in Proposition 42 and with the same algorithm. For *(2.)* $\Rightarrow$ *(1.)* and the proof of the existence of a consistent learning algorithm Λ such that $\text{Reg}(T; M, \Lambda) = \text{o}(\log(T))$, we provide the construction of the algorithm and simply sketch the proof.

G.4.1. A ROBUST ALGORITHM SPECIALIZED TO A NON-EXPLORATIVE MODEL

We begin by providing a robust algorithm Λ such that $\text{Reg}(T; M, \Lambda) = \text{O}(1)$ for M specifically.

Proposition 41 *Consider a* ***convex*** *ambient space $\mathcal{M}^* \equiv \prod_{z\in\mathcal{Z}}(\mathcal{R}^*_z \times \mathcal{P}^*_z)$ in product form and let $M \in \mathcal{M}^*$ be non-degenerate. If* $\text{Cnf}(M) = \emptyset$*, there exists a learning algorithm Λ that (1) is robust, (2) makes sub-linearly many episodes and (3) satisfies* $\text{Reg}(T; M, \Lambda) = \text{O}(1)$.

The algorithm that we consider is a variant of KLUCRL managing episodes with (DT), that is specialized for M. Also, we have to take into account that $\mathcal{M}^*$ may not be the whole set of Markov decision processes with pair space $\mathcal{Z}$, i.e., we may have $\mathcal{M}^* \neq \prod_{z\in\mathcal{Z}}([0,1] \times \mathcal{P}(\mathcal{S}))$. It must be taken into account by the learning algorithm, as the property "$\text{Cnf}(M) = \emptyset$" depends on $\mathcal{M}^*$—by definition (57), $\text{Cnf}(M) \subseteq \mathcal{M}^*$ so if one increases $\mathcal{M}^*$ to $\mathcal{M}'$, the confusing set of M relatively to $\mathcal{M}'$ may become non-empty. So, the confidence region of KLUCRL, $\mathcal{M}(t)$, is constrained to $\mathcal{M}^*$ to eventually exploit that $\text{Cnf}(M) = \emptyset$ relatively to $\mathcal{M}^*$.

Notations We introduce the natural optimal policy of M, given by $\pi^*(s) = a$ where $a \in \mathcal{A}(s)$ is the unique element such that $(s,a) \in \mathcal{Z}^*(M)$. Further introduce:

$$
\mathcal{R}_z(t; \mathcal{M}^*) := \left\{\tilde{r}_z \in \mathcal{R}^*_z : N_z(t)\text{KL}(\hat{r}_z(t)||\tilde{r}_z) \leq \log(2t) + \log(e(1 + N_z(t)))\right\}
$$

$$
\mathcal{P}_z(t; \mathcal{M}^*) := \left\{\tilde{p}_z \in \mathcal{P}^*_z : N_z(t)\text{KL}(\hat{p}_z(t)||\tilde{p}_z) \leq \log(2t) + |\mathcal{S}|\log\left(e\left(1 + \frac{N_z(t)}{|\mathcal{S}| - 1}\right)\right)\right\}
$$

Note that unlike (5) of the vanilla KLUCRL, $\log(t)$ is changed to $\log(2t)$. This is done so that $\mathbf{P}(\exists t \geq T : M \notin \mathcal{M}(t; \mathcal{M}^*)) = \mathrm{O}(T^{-2})$ instead of $\mathrm{O}(T^{-1})$ as in the vanilla version. The confidence region for π^* is $\mathcal{M}_{\pi^*}(t; \mathcal{M}^*) := \prod_{s \in \mathcal{S}}(\mathcal{R}_{s,\pi^*(s)}(t; \mathcal{M}^*) \times \mathcal{P}_{s,\pi^*(s)}(t; \mathcal{M}^*))$. Similarly to the whole confidence region $\mathcal{M}(t; \mathcal{M}^*)$, it can be seen as a Markov decision process with compact action space by extending its action space (see Section B.1) and EVI can be run on $\mathcal{M}_{\pi^*}(t; \mathcal{M}^*)$ to compute the optimistic gain of π^*, written $g^{\pi^*}(\mathcal{M}(t; \mathcal{M}^*))$.

Idea of the algorithm The designed algorithm is working by epochs of doubling sizes. Given an epoch $\{2^m, \ldots, 2^{m+1} - 1\}$, it starts by iterating π^* $(2^m)^{2/3}$ times in a row. After that initial phase, the algorithms runs an altered version of KLUCRL that uses EVI specifically biased for π^*, that, when several policies are nearly optimistically optimal, prioritizes π^*.

Algorithm 3 KLUCRL$(\pi^*, \mathcal{M}^*)$

for epochs $m = 0, 1, 2, \ldots$ **do**
  Iterate π^* for $t = 2^m, \ldots, 2^m + 2^{2m/3}$;
  for $t = 2^m + 2^{2m/3}, \ldots, 2^{m+1}$ **do**
    if (DT) triggers **or** $t = 2^m + 2^{2m/3}$ **then**
      $k \leftarrow k + 1$, $t_k \leftarrow t$;
      $\pi_{t_k} \leftarrow$ EVI-b$_{\pi^*}(\mathcal{M}(t; \mathcal{M}^*), t)$;
    end if
    Set $\pi_t \leftarrow \pi_{t_k}$ and play $A_t \leftarrow \pi_t(S_t)$.
  end for
end for

Algorithm 4 EVI-b$_\pi(\widetilde{\mathcal{M}}, t)$

Compute $\tilde{\pi} \leftarrow$ EVI$(\widetilde{\mathcal{M}})$;
Compute $\tilde{g}^* \leftarrow g^*(\widetilde{\mathcal{M}})$;
Compute $\tilde{g}^\pi \leftarrow g^\pi(\widetilde{\mathcal{M}})$;
if $\tilde{g}^* > \tilde{g}^\pi + \frac{\log(t)}{t^{1/3}}$ **then**
  return $\tilde{\pi}$;
else
  return π.
end if

Proof of Proposition 41 Proving that the algorithm is robust on $\mathcal{M}^*$ follows a similar line that Appendix B, that we won't detail here. The idea is that the forced exploration with π^* last for at most $T^{2/3}$ time steps of the learning process, accounting for a regret of order $T^{2/3}$ if π^* is not optimal. Later, the algorithm deploys policy that are $\frac{\log(T)}{T^{1/3}}$-optimistically optimal, inducing an extra cost of $T^{2/3} \log^2(T)$ compared to the vanilla analysis of KLUCRL. Therefore, the model independent regret is KLUCRL$(\pi^*, \mathcal{M}^*)$ is of order $T^{2/3} \log^2(T) = \mathrm{o}(T)$.

Meanwhile, KLUCRL$(\pi^*, \mathcal{M}^*)$ makes $\mathrm{O}(\log(T))$ episodes: one for each epoch when playing π^*, and the others are triggered by (DT) that is known to produce logarithmically many episodes, see Auer et al. (2009) or Section B.5.

Last but not least, we argue that $\mathrm{Reg}(T; M) = \mathrm{O}(1)$. Because this is an instance dependent result, the argument is different from Appendix B. The idea is to show that

$$\mathbf{P}^M\big(\exists t \in \{2^m, \ldots, 2^{m+1} - 1\} : \pi_t \neq \pi^*\big) = \mathrm{O}\big(4^{-m}\big). \tag{70}$$

Following (70), we have:

$$\mathrm{Reg}(T; M) \leq \sum_{m=0}^{\infty} \mathbf{E}^M \left[\sum_{t=2^m}^{2^{m+1}-1} \Delta^*(Z_t) \right]$$

$$\leq \sum_{m=0}^{\infty} 2^m \max(\Delta^*) \mathbf{P}^M\left(\exists t \in \{2^m, \ldots, 2^{m+1}-1\} : \pi_t \neq \pi^*\right)$$
$$\overset{(\dagger)}{\leq} \max(\Delta^*) \sum_{m=0}^{\infty} \mathrm{O}\big(2^{-m}\big) = \mathrm{O}(1)$$

where $(\dagger)$ follows from (70). We now explain how (70) is established.

(STEP 1) *There exists $c > 0$ such that, for m large enough and $z \in \mathcal{Z}^{**}(M)$, we have:*

$$\mathbf{P}^M\Big(\exists t \in \{2^m + 2^{2m/3}, \ldots, 2^{m+1}-1\} : N_z(t) < c2^{2m/3}\Big) = \mathrm{O}(4^{-m}).$$

Proof The recurrent pairs of π^* are precisely $\mathcal{Z}^{**}(M)$, i.e., $\mathbf{P}^{M,\pi^*}\{\forall m, \exists n \geq m : Z_n = z\} = 1$. Because the state space is finite, it follows that $\min_{z \in \mathcal{Z}^{**}(M)} \mathbf{E}^{M,\pi^*}[N_z(t)] \geq ct$ for some $c > 0$ when $t \to \infty$. It means that under π^*, every optimal pair is visited linearly many times in expectation. Fixing $z_0 \in \mathcal{Z}^{**}(M)$ and setting $f(z) = \mathbf{1}(z = z_0)$, we show that $N_{z_0}(t) \geq ct$ holds in probability as well. This is done as follows. Seeing f as a reward function, π^* has an associated gain and bias functions that we denote g^f and h^f. These satisfy a Poisson equation $f(s,\pi(s)) + p(s,\pi(s))h^f = g^f(s) + h^f(s)$. By non-degeneracy, π^* is unichain so $\mathrm{sp}(g^f) = 0$, and we see that $g^f(s) \geq c$ for all $s \in \mathcal{S}$. We continue as follows: If we only iterate π^*, then

$$\begin{aligned} N_{z_0}(t) &= \sum_{i=1}^{t} f(Z_t) \\ &= \sum_{i=1}^{t} \big(g^f(S_t) + h^f(S_t) - p(Z_t)h^f\big) \geq ct - \mathrm{sp}(h^f) + \sum_{i=1}^{t} \big(e_{S_{t+1}} - p(Z_t)\big)h^f. \end{aligned}$$

The RHS is a martingale and each term is almost surely bounded by $\mathrm{sp}(h^f)$. By Azuma-Hoeffding's inequality, it is therefore bounded by $\mathrm{sp}(h^f)\sqrt{t\log(\alpha t)/2} = \mathrm{o}(t)$ with probability $1 - \frac{1}{t^\alpha}$. So, provided that t is large enough, we have $N_z(t) \geq \frac{1}{2}ct$ with probability $1 - \frac{1}{t^\alpha}$.

Now, we know that on the time range $\{2^m, \ldots, 2^m + 2^{2m/3}\}$, the algorithm exclusively iterates π^*, so π^* is iterate $t = 2^{2m/3}$ times. Pick $\alpha = 3$. ■

(STEP 2) *There exists $C > 0$ such that for all $z \in \mathcal{Z}^{**}(M)$, we have*

$$\mathbf{P}^M\left(\begin{matrix}\exists t \in \{2^m + 2^{2m/3}, \ldots, 2^{m+1}-1\} \\ \exists \widetilde{\mathcal{M}}_t \equiv (\mathcal{Z}, \tilde{r}_t, \tilde{p}_t) \in \mathcal{M}(t; \mathcal{M}^*)\end{matrix} : \|\tilde{p}_t(z) - p(z)\|_1 + |\tilde{r}_t(z) - r(z)| > \frac{C\sqrt{\log(t)}}{t^{1/3}}\right)$$
$$= \mathrm{O}\big(4^{-m}\big)$$

Proof Fix $z \in \mathcal{Z}^{**}(M)$. By **(STEP 1)**, we know that $N_z(t) > c2^{2m/3}$ with probability $1 - \mathrm{O}(4^{-m})$ uniformly for $t \in \{2^m + 2^{2m/3}, \ldots, 2^{m+1}-1\}$. Using (Jonsson et al., 2020, Proposition 1) we find that for m large enough,

$$\mathbf{P}^M\left(\exists t \geq 1 : N_z(t) \geq c2^{2m/3} \text{ and } \mathrm{KL}(\hat{r}_t(z)||r(z)) > \frac{\log(2e \cdot 4^m)}{c2^{2m/3}}\right) \leq 4^{-m}.$$

So, by Pinsker's inequality, it follows that for m large enough, we have $|\hat{r}_t(z) - r(z)| \le C_r\sqrt{\log(t)}t^{-1/3}$ with probability $1 - \mathrm{O}(4^{-m})$ uniformly for $t = 2^m + 2^{2m/3}, \ldots, 2^{m+1} - 1$, where $C_r > 0$ is some constant. Now, by design of the confidence region $\mathcal{R}_z(t; \mathcal{M}^*)$, every $\tilde{r}_t(z) \in \mathcal{R}_z(t; \mathcal{M}^*)$ satisfies $|\tilde{r}_t(z) - \hat{r}_t(z)| \le \frac{1}{N_z(t)} \log(4et)$. By triangular inequality, we conclude that, for $z \in \mathcal{Z}^{**}(M)$, the inequality

$$|\tilde{r}_t(z) - r(z)| \le \frac{C\sqrt{\log(t)}}{t^{1/3}}$$

holds uniformly for $t = 2^m + 2^{2m/3}, \ldots, 2^{m+1} - 1$ with probability $1 - \mathrm{O}(4^{-m})$. Transition kernels are treated similarly. ■

(**STEP 3**) *There exists a constant $C > 0$ such that*

$$\mathbf{P}\left(\exists t = 2^m + 2^{2m/3}, \ldots, 2^{m+1} - 1 : g^{\pi^*}(\mathcal{M}(t; \mathcal{M}^*)) + \frac{C\sqrt{\log(t)}}{t^{1/3}} \le g^*(\mathcal{M}(t; \mathcal{M}^*))\right)$$
$$= \mathrm{O}(4^{-m}).$$

Proof This is where we use that $\mathrm{Cnf}(M) = \emptyset$.

Let $\mathcal{E}_m$ be the event under which $M \in \mathcal{M}(t; \mathcal{M}^*)$ and $\|\tilde{p}_t(z) - p_z\|_1 + |\tilde{r}_t(z) - r(z)| \le C\sqrt{\log(t)}t^{-1/3}$ for all optimal pair $z \in \mathcal{Z}^{**}(M)$, uniformly for $t = 2^m + 2^{2m/3}, \ldots, 2^{m+1} - 1$ and $\tilde{M}_t \equiv (\mathcal{Z}, \tilde{r}_t, \tilde{p}_t) \in \mathcal{M}(t; \mathcal{M}^*)$, where $C > 0$ is given by (**STEP 2**). Then, following (**STEP 2**) and the design of $\mathcal{M}(t; \mathcal{M}^*)$, we have $\mathbf{P}(\mathcal{E}_m) = 1 - \mathrm{O}(4^{-m})$.

Fix $t \in \{2^m + 2^{2m/3}, \ldots, 2^{m+1} - 1\}$ and let $\tilde{M} \in \mathcal{M}(t; \mathcal{M}^*)$ be a plausible model. Since $\mathcal{M}^*$ is convex, we can assume that $\tilde{M} \gg M$ up to changing $\tilde{M}$ to $(1 - \lambda)\tilde{M} + \lambda M$ for some arbitrarily small $\lambda > 0$. We show that, on $\mathcal{E}_m$,

$$g^{\pi^*}(M) + \frac{C_g\sqrt{\log(t)}}{t^{1/3}} \ge g^*(\tilde{M}) \tag{71}$$

for $C_g > 0$ some constant, independent from $\tilde{M}$ and m. Since, on $\mathcal{E}_m$ again, we further have $g^{\pi^*}(\mathcal{M}(t; \mathcal{M}^*)) \ge g^*(M)$, the result will follow from $\mathbf{P}(\mathcal{E}_m) = 1 - \mathrm{O}(4^{-m})$.

We now show (71). Let $M \cup \tilde{M}$ be the Markov decision process with states $\mathcal{S}$ where the choice of an action from s consists in (1) choosing $a \in \mathcal{A}(s)$ in the vanilla sense and (2) choosing whether the transition is made using $(r(s, a), p(s, a))$ or $(\tilde{r}(s, a), \tilde{p}(s, a))$. Note that $M \cup \tilde{M}$ is still a MDP with finite action space, that $g^*(M \cup \tilde{M}) \ge \max\{g^*(M), g^*(\tilde{M})\}$ and $D(M \cup \tilde{M}) \le D(M)$. In particular, $M \cup \tilde{M}$ is communicating and $\mathrm{sp}(h^*(M \cup \tilde{M})) \le D(M)$. Using `EVI` on $M \cup \tilde{M}$ (see Section B.1) to compute its optimal gain, we extract a MDP $M^\dagger$ such that $g^*(M^\dagger) = g^*(M \cup \tilde{M})$ and $\mathrm{sp}(h^*(M^\dagger)) \le D(M)$. By construction, $M^\dagger$ is a blend of M and $\tilde{M}$, in the sense that $r^\dagger(z) \in \{r(z), \tilde{r}(z)\}$ and $p^\dagger(z) \in \{p(z), \tilde{p}(z)\}$. Let $M_0^\dagger$ be obtained from $M^\dagger$ by setting

$$M_0^\dagger(z) := \begin{cases} M(z) & \text{if } z \in \mathcal{Z}^{**}(M); \\ M^\dagger(z) & \text{if } z \notin \mathcal{Z}^{**}(M). \end{cases}$$

Note that since $\mathcal{M}^*$ is in product form and $\mathcal{M}(t;\mathcal{M}^*) \subseteq \mathcal{M}^*$, we have $M_0^\dagger \in \mathcal{M}^*$. It follows that $M_0^\dagger \in \mathcal{M}^*$ and $M_0^\dagger = M$ on $\mathcal{Z}^{**}(M)$. Moreover, since $\tilde{M} \gg M$, we have $M^\dagger \gg M$ and $M_0^\dagger \gg M$. So, because $\mathrm{Cnf}(M) = \emptyset$, we have $\pi^* \in \Pi^*(M_0^\dagger)$. Accordingly,

$$g^{\pi^*}(M_0^\dagger) = g^*(M_0^\dagger) = g^*(M) \tag{72}$$

where the second equality follows from the observation that $M_0^\dagger$ is a copy of M on $\mathcal{Z}^{**}(M)$. Now, by construction $\mathcal{E}_m$, we know that the width of the confidence region is $\mathrm{O}(\sqrt{\log(t)}t^{-1/3})$ on pairs of $\mathcal{Z}^{**}(M)$. Using the gain deviation inequality of Lemma 32, we conclude that on $\mathcal{E}_m$,

$$\begin{aligned}
\left\|g^{\pi^*}(M) - g^{\pi^*}(M^\dagger)\right\|_\infty &\leq \left(1 + \mathrm{sp}(h^*(M^\dagger))\right) \mathrm{O}\left(\frac{\sqrt{\log(t)}}{t^{1/3}}\right) = \mathrm{O}\left(\frac{\sqrt{\log(t)}}{t^{1/3}}\right) \\
\left\|g^*(M_0^\dagger) - g^*(M^\dagger)\right\|_\infty &\leq \left(1 + \mathrm{sp}(h^*(M^\dagger))\right) \mathrm{O}\left(\frac{\sqrt{\log(t)}}{t^{1/3}}\right) = \mathrm{O}\left(\frac{\sqrt{\log(t)}}{t^{1/3}}\right)
\end{aligned} \tag{73}$$

where every $\mathrm{O}(-)$ hides constants that are independent of $\tilde{M}$ and m. Since $g^*(M^\dagger) \geq g^*(\tilde{M})$, we conclude accordingly. ■

Finally, (70) is an immediate consequence of ((**STEP 3**)). ((**STEP 3**)) states that, uniformly for $t = 2^m + 2^{2m/3}, \ldots, 2^{m+1} - 1$, we have:

$$g^{\pi^*}(\mathcal{M}(t;\mathcal{M}^*)) + \frac{C\sqrt{\log(t)}}{t^{1/3}} > g^*(\mathcal{M}(t;\mathcal{M}^*)) \tag{74}$$

with probability $1 - \mathrm{O}(4^{-m})$, where $C > 0$ is some constant independent of m. Now, by design of $\texttt{EVI-b}_{\pi^*}$ (Algorithm 4), if at $t = t_k \in \{2^m + 2^{2m/3}, \ldots, 2^{m+1}\}$, we have $g^{\pi^*}(\mathcal{M}(t;\mathcal{M}^*)) + \log(t)t^{-1/3} \geq g^*(\mathcal{M}(t;\mathcal{M}^*))$, then $\texttt{EVI-b}_{\pi^*}$ outputs π^*. So, on the event prescribed by (74) and for $t \geq \exp(C^2)$, $\texttt{EVI-b}_{\pi^*}$ outputs π^*. Hence (70). ■

G.4.2. MODELS WITH NON-EMPTY CONFUSING SET ARE NON-EXPLORATIVE

With the same algorithm, we show that non-degenerate Markov decision processes with empty confusing set are non-explorative.

Proposition 42 *Consider a **convex** ambient space $\mathcal{M}^* \equiv \prod_{z\in\mathcal{Z}}(\mathcal{R}_z^* \times \mathcal{P}_z^*)$ in product form and let $M \in \mathcal{M}^*$ be non-degenerate. If* $\mathrm{Cnf}(M) = \emptyset$*, then M is non-explorative, i.e., there exists a learning algorithm Λ (1) with sub-linearly many episodes, (2) which is no-regret on $\mathcal{M}^*$ and (3) that has finitely many exploration episodes.*

Proof We consider the algorithm $\texttt{KLUCRL}(\pi^*, \mathcal{M}^*)$ (Algorithm 3), introduced for the proof of Proposition 41. Following (70), we have:

$$\mathbf{P}^M(\exists t \geq 2^m : \pi_t \neq \pi^*) \leq \sum_{n=m}^{\infty} \mathbf{P}^M\left(\exists t \in \{2^n, \ldots, 2^{n+1} - 1\} : \pi_t \neq \pi^*\right)$$

$$= \mathrm{O}\left(\sum_{n=m}^{\infty} 4^{-n}\right) = \mathrm{O}(4^{-m}) = \underset{m\to\infty}{\mathrm{o}}(1).$$

So $\mathbf{P}^M(\exists T, \forall t \geq T : \pi_t = \pi^*) = 1$. Because every pair z that π^* can reach satisfies $\Delta^*(z; M) = 0$ by construction of π^*, it follows that $\mathbf{P}^M(\exists T, \forall t \geq T : \Delta^*(Z_t) = 0) = 1$. So necessarily, the number of exploration times (Definition 2) is almost-surely finite. ■

G.4.3. A CONSISTENT ALGORITHM SPECIALIZED TO A NON-EXPLORATIVE MODEL

We conclude by arguing that the robust algorithm of for Proposition 41, $\texttt{KLUCRL}(\pi^*, \mathcal{M}^*)$, can be adapted into a consistent learning algorithm Λ' such that $\mathrm{Reg}(T; M, \Lambda') = \mathrm{o}(\log(T))$.

Proposition 43 *Consider a **convex** ambient space $\mathcal{M}^* \equiv \prod_{z\in\mathcal{Z}}(\mathcal{R}^*_z \times \mathcal{P}^*_z)$ in product form and let $M \in \mathcal{M}^*$ be non-degenerate. If* $\mathrm{Cnf}(M) = \emptyset$*, there exists a learning algorithm Λ that (1) is consistent; and (2) satisfies* $\mathrm{Reg}(T; M, \Lambda) = \mathrm{o}(\log(T))$.

The considered algorithm is a reworked version of $\texttt{KLUCRL}(\pi^*, \mathcal{M}^*)$ (Algorithm 3). When tuning Algorithm 3 and Algorithm 4 to provide a robust algorithm, Algorithm 3 forces $\Omega(T^{2/3})$ iterations of the policy π^*. This is incompatible with consistency, as the latter implies that the model dependent regret is $\Omega(T^{2/3})$ when π^* is not gain optimal. Instead, the algorithm will force $\mathrm{O}(\log^3(T))$ iterations of π^*, losing robustness but achieving consistency. Then, the biased $\texttt{EVI}$ adds a bonus to favor the selection of π^*, compensating the noise on the estimation of g^{π^*}.

Algorithm 5 $\texttt{KLUCRL}'(\pi^*, \mathcal{M}^*)$

1: **for** epochs $m = 0, 1, 2, \ldots$ **do**
2: Set $\psi(2^m) \leftarrow \log(2^m)$;
3: Play π^* for $t = 2^m, \ldots, 2^m + \psi(2^m)$
4: **for** $t = 2^m + \psi(2^m), \ldots, 2^{m+1}$ **do**
5: **if** (DT) triggers **or** $t = 2^m + \psi(2^m)$ **then**
6: $k \leftarrow k + 1$, $t_k \leftarrow t$;
7: $\pi_{t_k} \leftarrow \texttt{EVI-b}_{\pi^*}(\mathcal{M}(t; \mathcal{M}^*), t)$;
8: **end if**
9: Set $\pi_t \leftarrow \pi_{t_k}$ and play $A_t \leftarrow \pi_t(S_t)$.
10: **end for**
11: **end for**

Algorithm 6 $\texttt{EVI-b}'_\pi(\widetilde{\mathcal{M}}, t)$

1: Compute $\tilde{\pi} \leftarrow \texttt{EVI}(\widetilde{\mathcal{M}})$;
2: Compute $\tilde{g}^* \leftarrow g^*(\widetilde{\mathcal{M}})$;
3: Compute $\tilde{g}^\pi \leftarrow g^\pi(\widetilde{\mathcal{M}})$;
4: **if** $\tilde{g}^* > \tilde{g}^\pi + \sqrt{\frac{1}{\log(t)}}$ **then**
5: **return** $\tilde{\pi}$;
6: **else**
7: **return** π.
8: **end if**

Sketch of proof Proving Proposition 43 in full detail is, again, tedious. As the proof follows from techniques that are quite similar to Proposition 41, we only leave the main ideas. We fix $M' \in \mathcal{M}^*$ and look at whether $M' = M$ or $M' \neq M$.

If $M' \neq M$, we prove that $\mathrm{Reg}(T; M') = \mathrm{O}(\log^3(T))$ when $T \gg \exp\{(g^*(M') - g^{\pi^*}(M'))^{-1}\}^2$. T needs to be large in order to compensate for the bonus that Algorithm 6 puts on the optimistic gain of π^*, that intentionally overshoots the empirical noise on g^{π^*}, i.e., the value of $|g^{\pi^*}(\hat{M}_t) - g^{\pi^*}(M)|$. Beyond that subtlety, we enumerate $z \notin \mathcal{Z}^*(M)$ and we distinguish the cases where $Z_t = z$ depending on whether (1) z is transient under the currently deployed policy or (2) z is recurrent under the currently deployed policy and (3) z is a recurrent pair of π^* and $t \in \bigcup_m \{2^m, \ldots, 2^m + \psi(m)\}$ is within a forced exploration phase. The first is proportionally to the number of episodes, which is $\mathrm{O}(\log(T))$, while the second implies that confidence regions are wrong. As discussed above, when taking account of the bonus in Algorithm 6, confidence regions start to be correct when $T \gg \exp\{(g^*(M') - g^{\pi^*}(M'))^{-1}\}^2$, leading to an error that sums as $\mathrm{O}(1)$ overall. The third accounts for $\mathrm{O}(\log^3(T))$ which is eventually the dominant term in the regret bound.

If $M' = M$, we prove that $\mathrm{Reg}(T; M) = \mathrm{O}(1)$ with the exact same proof technique as Proposition 41, by establishing an equation in the style of (70) with the same proof strategy (**STEP 1, 2, 3**), adapted to a forced exploration of $\Theta(\log^3(T))$ rather than $\Theta(T^{1/3})$. ■

G.5. Interior Markov decision processes are universally explorative

We conclude the discussion on explorative spaces by discussing examples of explorative spaces, and how common the property "$M \in \mathcal{M}^+$" may be. As shown in Section G.1, when the ambient space $\mathcal{M}$ is a fixed kernel space (i.e., is of the form $\mathcal{M} \equiv \prod_{z \in \mathcal{Z}} (\mathcal{R}_z \times \{p(z)\}$ for some fixed $p \in \mathcal{P}(\mathcal{S})^{\mathcal{Z}}$) and $\mathcal{R}_z \subseteq [0, 1]$, large portions of $\mathcal{M}$ may be non-explorative. For instance, taking

$$\mathcal{M} := \left\{ M \equiv (\mathcal{Z}, r, p) : r \in [0,1]^{\mathcal{Z}} \text{ and } \forall z \in \mathcal{Z}, \forall s \in \mathcal{S}, p(s|z) \in \{0, 1\} \right\}$$

the space of deterministic transition Markov decision processes with pair space $\mathcal{Z}$, one can generalize the example of Section G.1 to show that as soon as $|\mathcal{Z}| > |\mathcal{S}|$, a model picked in $\mathcal{M}$ uniformly at random is non-explorative with positive probability.

The take-away of this observation is that if $\mathcal{M}$ is structured, then $\mathcal{M}^+$ can be large. In Proposition 44 below, we prove a complementary result: If $\mathcal{M}$ has no structure, then every (non-degenerate) model in the interior (see Assumption 4) of $\mathcal{M}$ is explorative. It follows (see Corollary 45) that if the ambient space $\mathcal{M}$ has no structure and if $M \in \mathcal{M}$ is picked uniformly at random, then M is explorative almost surely. This property is very convenient for experiments: Any Markov decision processes that is generated randomly is a good environment to investigate regret of exploration guarantees.

Proposition 44 (Interior implies explorative) *Let $\mathcal{M} \equiv \prod_{z \in \mathcal{Z}} ([0, 1] \times \mathcal{P}(\mathcal{S}))$ be the space of all Markov decision processes with pair space $\mathcal{Z}$, with $|\mathcal{Z}| > |\mathcal{S}|$. Then every non-degenerate (n.d.) interior model is explorative, i.e.,*

$$\{M \equiv (\mathcal{Z}, r, p) \textit{ n.d.} : \forall z \in \mathcal{Z}, \mathrm{supp}(r(z)) = \{0, 1\} \textit{ and } \mathrm{supp}(p(z)) = \mathcal{S}\} \subseteq \mathcal{M}^+.$$

Proof Let $M \in \mathcal{M}$ be a non-degenerate interior model. We show that $\mathrm{Cnf}(M) \neq \emptyset$. Because $|\mathcal{Z}| > |\mathcal{S}|$ and M is non-degenerate, there exists $z_0 \in \mathcal{Z} \setminus \mathcal{Z}^*(M)$. By definition, $\Delta^*(z_0; M) > 0$ so that this pair cannot be recurrent under any gain optimal policy of M. For $\epsilon > 0$, define $M_\epsilon \equiv (\mathcal{Z}, r_\epsilon, p_\epsilon)$ the Markov decision process given by:

$$p_\epsilon(z) := \begin{cases} p(z) & \text{if } z \neq z_0 \\ (1-\epsilon) e_{s_0} + \epsilon \cdot p(z) & \text{if } z = z_0 \end{cases} \quad \text{and} \quad r_\epsilon(z) := \begin{cases} r(z) & \text{if } z \neq z_0 \\ 1 - \epsilon + \epsilon \cdot r(z) & \text{if } z = z_0 \end{cases}$$

where $z_0 \equiv (s_0, a_0)$ and e_{s_0} is the Dirac at s_0. By construction, $M_\epsilon \gg M$ for all $\epsilon > 0$. Let π be the (randomized) policy that picks actions uniformly at random from $s \neq s_0$ and with $\pi(s_0) = a_0$. It is clear that as $\epsilon \to 0$, we have $g^\pi(M_\epsilon) \to 1$ because π converges to a unichain policy that converges to a loop on s_0 where it scores $1 - \epsilon$. Now, if $\pi^* \in \Pi^*(M)$, we have $g^{\pi^*}(M_\epsilon) = g^{\pi^*}(M)$ since π^* does not pick a_0 from s_0. So, for $\epsilon > 0$ small enough, we have $g^\pi(M_\epsilon) > g^{\pi^*}(M_\epsilon)$ for all $\pi^* \in \Pi^*(M)$. For such $\epsilon > 0$, we have $\Pi^*(M_\epsilon) \cap \Pi^*(M) = \emptyset$ and it follows that $M_\epsilon \in \mathrm{Cnf}(M)$. So $\mathrm{Cnf}(M) \neq \emptyset$ and M is explorative from Theorem 36. ■

Corollary 45 *Let $\mathcal{M} \equiv \prod_{z \in \mathcal{Z}}([0,1] \times \mathcal{P}(\mathcal{S}))$ be the space of all Markov decision processes with pair space $\mathcal{Z}$, with $|\mathcal{Z}| > |\mathcal{S}|$. Let $M \equiv (\mathcal{Z}, r, \mathcal{Z})$ where $r(z) \sim \mathrm{U}[0,1]$ and $p(z) \sim \mathrm{U}(\mathcal{P}(\mathcal{S}))$ are sampled independently. Then $M \in \mathcal{M}^+$ almost surely.*

Proof If M is picked at random as described above, then M is interior with probability one. Meanwhile, the set of degenerate models of a fixed arbitrary kernel $p \in \mathcal{P}(\mathcal{S})$ is of measure zero, see Corollary 35. Integrating, the set of degenerate models is negligible in $\mathcal{M}$ for the Lebesgue measure. Hence, if M is picked at random described, then M is non-degenerate with probability one. ■