The Regret of Exploration and the Control of Bad Episodes in Reinforcement Learning

Victor Boone [*] [1] **Bruno Gaujal** [*] [1]

Abstract

The first contribution of this paper is the introduction of a new performance measure of a RL algorithm that is more discriminating than the regret, that we call the *regret of exploration* that measures the asymptotic cost of exploration. The second contribution is a new *performance test* (PT) to end episodes in RL optimistic algorithms. This test is based on the performance of the current policy with respect to the best policy over the current confidence set. This is in contrast with all existing RL algorithms whose episode lengths are only based on the number of visits to the states. This modification does not harm the regret and brings an additional property. We show that while all current episodic RL algorithms have a linear regret of exploration, our method has a $O(\log T)$ regret of exploration for non-degenerate deterministic MDPs.

1. Introduction and Motivation

In infinite horizon undiscounted reinforcement learning (RL), an algorithm dynamically learns a Markov Decision Process (MDP) by picking actions. Its performance is measured by the *regret* that aggregates the rewards and compare this accumulated score to the best achievable score of an algorithm that knows the MDP in hindsight. An algorithm is no-regret if its regret is sublinear, *i.e.*, its average performance is asymptotically optimal.

Many RL algorithms are model-based, meaning that they maintain a confidence region for the underlying unknown MDP modeling the reward mechanism. These model-based algorithms use episodes over which the policy is fixed. At the end of each episode, the observations made during the episode are used to update the confidence region. A new policy is computed from the updated estimate and is used for the next episode.

This paper focuses on model-based algorithms following the principle of *optimism in the face of uncertainty*, that pick the policy achieving maximal gain in the confidence region. Model based optimistic algorithms originate with UCRL (Auer & Ortner, 2006), which is itself an adaptation to RL of the celebrated UCB algorithm (Auer et al., 2002) for multi-armed bandits. Episodes, understand time-windows where the algorithm sticks to a single policy,[1] were added to the UCB template for the following reason: If the lengths of episodes are uniformly bounded (e.g., by 1 just like in UCB), then there exist MDPs over which the regret will grow linearly (see Example 1 in (Ortner, 2010a)). Therefore, every model-based RL algorithm uses episodes with lengths increasing with time. The popular UCRL2 (Auer et al., 2009) and its variants (Fruit et al., 2020; Tossou et al., 2019b; Bourel et al., 2020; Filippi et al., 2010) use the *doubling trick*: An episode terminates when the visit count of a transition doubles. Some other papers do not use the doubling trick. For example (Tossou et al., 2019a) uses what is called the extended doubling trick (the sum of the visits to all the states doubles at each episode). In the Bayesian context (Ouyang et al., 2017) uses episodes whose length grows by one at each step. However, up to our knowledge, in all cases, the lengths of all the episodes grow to infinity – with a few exceptions, e.g. the recent IMED-RL (Pesquerel & Maillard, 2022), but this one is not episodic nor optimistic, and limited to the ergodic setting. This grow also concerns bad episodes (i.e., where a sub-optimal policy is used), hence all current episodic algorithms behave arbitrarily bad, for arbitrarily long episodes. To illustrate this, let us consider the following simple example, given in Figure 1.

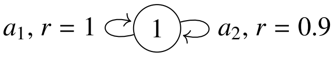

Figure 1: (Two-arm bandit) A MDP with one state and two actions, rewards of the two actions are both Gaussian distributed with unit variance and respective means $r(a_1) = 1$ and $r(a_2) = 0.9$.

[*] Equal contribution [1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France. Correspondence to: Victor Boone <victor.boone@univ-grenoble-alpes.fr>, Bruno Gaujal <bruno.gaujal@inria.fr>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Our use of the term “episode” is the same as in (Auer et al., 2009) and is not to be confused with the broader use of (Sutton & Barto, 2018), that call “episode” the inner part of generic loops.

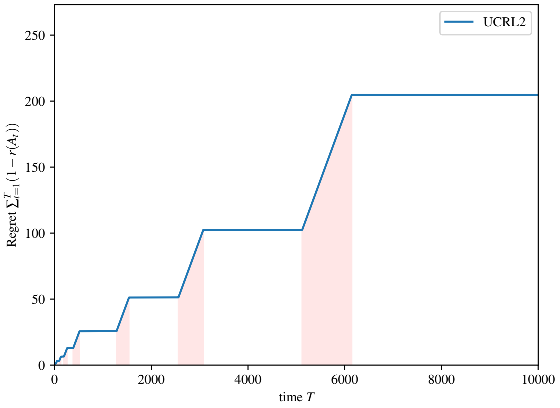

Figure 2: Regret of a run of UCRL2 over the MDP given in Figure 1. The bad episodes (in light red) correspond to time intervals where the current action is $A_t = a_2$ and good episodes to ones where $A_t = a_1$.

A run of UCRL2 on this MDP is shown in Figure 2. The regret only grows during bad episodes, where the current action A_t is the bad a_2, yielding the expected reward 0.9. On the other hand, during good episodes, the played action is a_1 with expected reward 1 and the regret does not grow. By definition, the duration of the k-th bad episode is 2^k here, so UCRL2 will stick to the bad action for arbitrarily long periods of time.

However letting all episodes grow to infinity may not be necessary especially because the compromise between exploration and exploitation that is advocated as being the key to efficient learning is not present here: The length of the episodes is solely based on an exploration criterion (the number of visits) and does not take the performance of the current policy into account. Following this observation, we raise the following points:

1. How to measure the cost of these bad episodes more discriminatingly than with the overall regret?

2. Under the metric designed by the previous point, how to design an algorithm with efficiently managed bad episodes without harming the regret guarantees?

We provide an answer to both questions. Our first contribution is to design a new metric, that we call the *regret of exploration*, that measures the phenomenon reported above and hence is of higher order than the traditional regret (Section 3). Then, we provide a new model-based optimistic algorithm, called UCRL-PT, whose regret of exploration grows sublinearly in deterministic MDPs (Section 4). The way UCRL-PT manages episodes can be adapted to most episodic algorithms cited above, leading to an improvement of their regret of exploration without experimental degradation of their regret performance.

2. Episodic Reinforcement Learning

This section introduces standard material from MDP theory and reinforcement learning.

2.1. Markov Decision Processes

In this paper, we consider tabular MDPs $M := \langle \mathcal{S}, \mathcal{A}, P, q \rangle$ consisting in a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, transition kernels $P(\cdot|s, a)$ where $P(s'|s, a)$ is the probability to switch to state s from s' by picking action a, and reward distributions $q(x, a)$ whose means are denoted $r(x, a)$. The state-action space is the product $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$. The MDP is *deterministic* if for all x, a, y, $P(y|x, a) \in \{0, 1\}$. A *policy* is a stationary deterministic map $\pi : \mathcal{S} \to \mathcal{A}$. Under the execution of a policy (or an algorithm), we write $Z_t := (X_t, A_t)$ the random state-action pair at time t, and R_t the obtained reward. In particular, $X_{t+1} \sim P(\cdot|X_t, A_t)$ and $R_t \sim q(X_t, A_t)$. In undiscounted reinforcement learning, one is interested to maximize the average aggregate rewards in expectation, i.e., $\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} R_t]$ when T grows large. Under a fixed policy, this quantity is called the *gain*,

$$g_x(\pi, M) := \lim_{T\to\infty} \mathbb{E}_x^{M,\pi}\left[\frac{1}{T}\sum\nolimits_{t=1}^{T} R_t\right] \tag{1}$$

where $\mathbb{E}_x^{M,\pi}[\cdot]$ is the expectation on M under π starting from the initial state $x \in \mathcal{S}$. We will also use the notation $\mathbb{P}_x^{M,\pi}(\cdot)$ for the associated probability distribution. We write P^π the transition matrix under π and r^π the reward vector $r^\pi(x) := r(x, \pi(x))$.

Communicating MDPs. In the rest of the paper, we will make the standard *communicating MDP* assumption, that guarantees that, for all $x, y \in \mathcal{S}$, there is a policy π and $t > 0$ such that $(P^\pi)^t(x, y) > 0$. Under this assumption, the optimal gain

$$g_x^*(M) = \max_\pi g_x(\pi, M) \tag{2}$$

is independent of the initial state x (Puterman, 1994), and is simply denoted $g^*(M)$ or g^*.

Bias and diameter. To every policy is associated a bias vector $h^\pi(x, M)$ given by

$$h^\pi(x, M) := \lim_{T\to\infty} \mathbb{E}_x^{M,\pi}\left[\sum\nolimits_{t=1}^{T}(R_t - g^*)\right] \tag{3}$$

or the Cesàro-limit when the above doesn't converge. The bias and the gain of a policy are well-known to be linked via the matrix identity $(P^\pi - I)h^\pi = r^\pi - g(\pi)$. Also, among policies achieving optimal gain, there is one that achieves maximal bias from all state x, that is said *bias optimal*. Its bias vector is denoted $h^*(x, M)$.

The optimal bias is known to be connected to the diameter of the MDP, which is the maximal expected amount of time to transit from a state to another. Specifically,

$$D(M) := \max_{x \neq y} \min_{\pi} \mathbb{E}_x^{M,\pi}[\tau_y] \tag{4}$$

where $\tau_y := \inf\{t \geq 1 : X_t = y\}$ is the reaching time to y. The diameter is finite if the MDP is communicating. Then it is well-known that $h^*(x, M) \leq D(M)\mathrm{sp}(r)$ where $\mathrm{sp}(r) := \max_z r(z) - \min_z r(z)$ is the *span* of the mean reward vector $r \in \mathbb{R}^{\mathcal{Z}}$ (Bartlett & Tewari, 2009, Theorem 4).

2.2. Reinforcement Learning

We consider an online simulation model for the true MDP: At time t, a sample of $q(X_t, A_t)$ and the next state X_{t+1} are observed when the current state is X_t and action A_t is chosen by the learner. A *reinforcement learning algorithm* is a (random) sequence of policies $\{\pi_t\}$ where each π_t only depends on the previous observations X_t, $(X_i, A_i, R_i)_{i<t}$ and controls the stochastic process by picking the next action. That is, $A_t := \pi_t(X_t)$ at time t.

Regret of a RL algorithm. The standard way to measure the online performances of a learner $\mathcal{L}$ is the *regret*, (sometimes called the pseudo-regret) here defined as in (Auer et al., 2009) with

$$\mathrm{Reg}(T) := Tg^* - \sum\nolimits_{t=1}^{T} r(Z_t). \tag{5}$$

Episodic algorithms. In the following, we focus on online learning algorithms with episodes. Although the regret (5) is not episodic, all model-based RL algorithms are episodic – there is usually a cost of switching policies and an algorithm may only pick optimal policies $\pi_t \in \Pi^*$ yet endure linear regret, see (Ortner, 2010b). So algorithms pick a policy, stick to it until a stopping condition is met then only they may change it. The collection of time instants $[1, T]$ is thus split into *episodes* $\{[t_k, t_{k+1} - 1] : k\}$ over which the policy is constant. Specifically, the policy π_t used at time t doesn't change over $[t_k, t_{k+1} - 1]$ and is denoted π^k.

2.3. Model-Based Optimistic Algorithms

As their name indicates, *model-based* algorithms maintain a model of their environment M. Usually, from their observations X_t, $(X_i, A_i, R_i)_{i<t}$, the algorithm builds a confidence set $\widetilde{\mathcal{M}}_t$ of the plausible models from which it takes its decisions. This confidence set is built using concentration inequalities (such as UCRL2 or UCRL3) or using KL-divergence (such as KL-UCRL).

Optimism. The *optimism-in-the-face-of-uncertainty* (OFU) principle states that you shall pick the policy achieving maximal gain in $\widetilde{\mathcal{M}}_t$. Denote $\tilde{g}_t(\pi) := \sup_{\tilde{M} \in \widetilde{\mathcal{M}}_t} g(\pi, \tilde{M})$ and $\tilde{g}_t^* := \max_\pi \tilde{g}_t(\pi)$. By optimism, the policy achieving $\tilde{g}_t^*$ is a good pick at time t.

To streamline our discussion, we present below the pseudo-code of a generic optimist episodic RL algorithm. All algorithms mentioned above are based on this scheme. This generic algorithm uses the doubling trick to manage its episodes and Extended Value Iteration (EVI) to compute $\tilde{g}_t^*$ together with the optimistically optimal policy.

Algorithm 1 Generic Episodic Learning Algorithm.

```
$t \leftarrow 1$;
for $k = 1, 2, \ldots$ do
    $t_k \leftarrow t$;
    $\pi^k \leftarrow \mathrm{argmax}_\pi \tilde{g}_t(\pi)$ ;
    repeat
        Execute action $A_t = \pi^k(X_t)$ ;
        Observe $R_t$ and next state $X_{t+1}$;
        $t \leftarrow t + 1$;
    until (Doubling Trick) $N_t(Z_t) \geq (\max\{1, 2N_{t_k}(Z_t)\})$
end for
```

In practice, the operation "$\pi^k \leftarrow \mathrm{argmax}_\pi \tilde{g}_t(\pi)$" is done using Extended Value Iteration (EVI), see (Auer et al., 2009). When the confidence set $\widetilde{\mathcal{M}}_t$ is well chosen any algorithm based on the generic algorithm 1 has a sub-linear regret. In the original paper (Auer et al., 2009), the first algorithm using this approach, namely UCRL2, uses a confidence set built on Hoeffding concentration inequalities and it is shown that with probability of at least $1 - \delta$, whatever the starting state X_1 and the time $T \geq 1$, the regret of UCRL2 is bounded by

$$\mathrm{Reg}(T) = \tilde{O}\left(DS\sqrt{AT}\right)$$

where $\tilde{O}(\cdot)$ is a $O(\cdot)$ hiding polylog(T) factors. Its subsequent variants improve on this bound and their respective regret upper bounds have been pushed closer and closer to the known lower bound $\Omega(\sqrt{DSAT})$, see (Wei et al., 2020) for a recent overview.

3. Regret of Exploration

It is quite obvious, at least experimentally (Figure 2), that algorithms relying on the doubling trick will use suboptimal policies for arbitrarily long time-periods infinitely often. From the viewpoint of the regret alone, this doesn't mater. Indeed, the regret aggregates all the rewards and measures how well, the algorithm behaves using a Cezaro average.

But this time average is sometimes not good enough. In real life, rewards don't simply aggregate from time 0. Consider the very first problem for which multi-armed bandits have been invented: providing medicine to ill patients (Thompson, 1933). The regret measures how well, the medicine provided to patients is useful to the average patient. One that sees Healthcare as an online service cannot be satisfied of aggregate rewards alone: You can't justify giving inefficient treatments to people because you've provided the right drug for however long past period. Of course, because the inefficiency of a treatment cannot be objected from finitely many observations, one has to check infinitely often that the seemingly bad option is truly bad. The point is that it should be done without harming *today*'s healthcare quality.

Another motivation for introducing a new metric is to assess the behavior of a learning algorithm once the MDP is already well estimated. Indeed, the transient phase, where some actions have not yet been properly sampled, can be seen as a necessary untidy burn-in phase where the performance of a learning algorithm is not pertinent. However, once all states and actions have been well estimated a "good" learning algorithm should be able to cut exploration periods to a minimum.

This is precisely what is measured by a new performance function that we introduce, namely the *regret of exploration*. It comes in addition to the regret and completes it. It is especially sound when one cannot tolerate an algorithm that behaves poorly for long periods of times.

Other works going beyond regret. The present work isn't the first to claim that "regret is not enough" in RL. The work of (Dann et al., 2018) introduces the *uniform PAC learning* setting that overcomes some blind spots of no-regret guarantees; Yet the suggested learning setting is done in finite horizon and is incompatible with sublinear expected regret guarantees (see (Dann et al., 2018, Theorem 1)), which is the setting our paper is built upon.

3.1. Definition of the Regret of Exploration

Consider an episodic learning algorithm $\mathcal{L}$ with sublinear regret in expectation. One consequence of the celebrated Theorem of (Lai & Robbins, 1985) is $\mathcal{L}$ will sample transitions from suboptimal policies infinitely often. Therefore, infinitely many times, $\mathcal{L}$ will finish an episode where the optimal policy π^* is used, and switch to a sub-optimal policy $\pi \notin \Pi^*$ for the next episode. The episode where π^* is used can be seen as a period where exploitation is maximal because the algorithm gains as much reward as possible. However, the very purpose of the next episode is to explore the environment. During this episode, $\mathcal{L}$ improves its confidence on suboptimal actions' rewards and transitions at the expense of getting suboptimal rewards.

The performance measure that we introduce, the regret of exploration, measures the cost of such exploration episodes. We first introduce the *sliding regret* that measures the regret starting from an arbitrary time t (and not at time 1, as the classical regret).

$$\mathrm{SReg}(t, t+T) = Tg^* - \sum\nolimits_{u=t}^{t+T-1} r(Z_u). \tag{6}$$

Definition 1 (Regret of exploration)**.** Denote $\{t_k : k \geq 1\}$ the (random) sequence of episode starting times. Let $\mathcal{K}_{\mathrm{exp}} := \left\{k : \pi^{k-1} \in \Pi^* \text{ and } \pi^k \notin \Pi^*\right\}$ be the sub-set of exploration episodes and let $\{t_{k(n)} : n \geq 1\}$ be the sub-sequence of the starting times of exploration episodes, i.e., $t_{k(n)}$ is the starting time of the n-th exploration episode. The *regret of exploration* at horizon T is given by the asymptotic quantity

$$\mathrm{RegExp}(T) := \limsup_{n\to\infty} \mathbb{E}\left[\mathrm{SReg}(t_{k(n)}; t_{k(n)} + T)\right] \tag{7}$$

To emphasize on the dependence on the learning algorithm $\mathcal{L}$ and/or the MDP M, we will sometimes write $\mathrm{RegExp}(\mathcal{L}; T)$ or $\mathrm{RegExp}(\mathcal{L}, M; T)$.

In the rest of this paper, we show that the regret of exploration of current RL algorithms is linear in T, showcasing their asymptotic instability, and we design a new algorithm (call UCRL-PT) with a *logarithmic* regret of exploration.

Remark 1. A sub-linear regret of exploration is stronger than short bad episodes, because an episodic algorithm may have short bad episodes yet may also have several of them in short succession, resulting in many bad choices of actions for a long time period. An algorithm with no regret of exploration is also immune to that kind of bad behavior.

3.2. Regret of Exploration of Classical Algorithms

The behavior of UCRL2 displayed in Figure 2 suggests that the doubling trick used to set the length of the episodes will lead to a linear regret of exploration.

Actually, a much more general statement can be shown for all *consistent* learning algorithms (an algorithm $\mathcal{L}$ is consistent according to (Burnetas & Katehakis, 1997) if its expected regret is sub-linear in a strong sense: $\mathbb{E}[\mathrm{Reg}(\mathcal{L}; T)] = o(T^\eta)$ for all $\eta < 1$).

Theorem 1. *Consider a consistent episodic reinforcement learning algorithm $\mathcal{L}$ with episode starting times $\{t_k : k \geq 1\}$ and assume that the event $\{t_{k+1} - t_k \to \infty\}$ is almost-sure. For all ergodic MDP M with at least one suboptimal state-action,*

$$\mathrm{RegExp}(\mathcal{L}, M; T) = \Omega(T)$$

when $T \to \infty$.

Although a few subtleties are on the way, this result isn't that surprising: A consistent algorithm will use suboptimal policies infinitely often, so if its episodes morally grow in size, the regret of exploration will be linear. A complete proof is provided in the Appendix.

This result is applicable to most algorithms of the literature. In particular, UCRL2, UCRL2B, KL-UCRL and UCRL3 have linear regret of exploration. [2]

4. UCRL-PT: UCRL with Performance Test

Theorem 1 assesses that all known model-based episodic algorithms are *unstable* in the asymptotic regime. The lengths of their suboptimal episodes are out-of-control. The culprit is the management of the episodes whose lengths are only based on visit counts; These algorithms are too lazy in the revision of their exploration policy. In practice (see Figure 2), UCRL-like algorithms have bad episodes that are way longer that they need to be. The time needed to update the confidence on a policy should be proportional to how bad the policy actually is: If the exploration policy is bad, it should be possible to figure it out fast.

4.1. A New Update Rule Based on the Policy's Performance

An important concept of the analysis of UCRL-PT – which is new regarding the analysis of RL algorithms, is the notion of the optimistic model *relative* to a policy π.

Definition 2. In the spirit of the (OFU) principle, the optimistic value of a policy π (with respect to a confidence region $\widetilde{\mathcal{M}_t}$) is

$$\tilde{g}_t(\pi) := \sup \left\{ \max_x g_x(\pi, \tilde{M}) : \tilde{M} \in \widetilde{\mathcal{M}_t} \right\} \tag{8}$$

The model's optimistic value is $\tilde{g}_t^* := \max_\pi \tilde{g}_t(\pi)$. Given a policy π, the optimistic model under π at time t is

$$\widetilde{M}_t^\pi = (\tilde{r}_t^\pi, \tilde{P}_t^\pi) := \operatorname{argmax} \left\{ \max_x g_x(\pi, \tilde{M}) : \tilde{M} \in \widetilde{\mathcal{M}_t} \right\} \tag{9}$$

The optimistic model differs from a policy to another, and in particular, $\tilde{g}_t(\pi^k)$ and $\tilde{g}_t^*$ refer to two different MDPs that have absolutely no reason to resemble one another. While $\tilde{P}_t^\pi$ depends on π, notice that the optimistic reward $\tilde{r}_t^\pi$ doesn't depend on π, because rewards are always maximized under the (OFU) principle, so one can just write $\tilde{r}_t$.

The performance test. This suggests a new episode management rule that goes like this: *If the current policy's optimistic value is too small with respect to the current model's optimistic value, change the policy. Specifically, at time t, end episode k if*

$$\tilde{g}_t(\pi^k) + \sqrt{\frac{\alpha \log(t)}{t}} \le \tilde{g}_t^*, \tag{\textbf{PT}}$$

This is completed with the doubling trick (**DT**) that guarantees that optimism (widths of confidence intervals) is never paid twice. The algorithm is given below, see Algorithm 2. The confidence bounds are generic.

Algorithm 2 UCRL-PT: Upper Confidence Reinforcement Learning with Performance Tests

```
Environment: Unknown MDP $M = (\mathcal{S}, \mathcal{A}, P, q)$.
Input:  Parameter $\alpha > 0$.
$t \leftarrow 1$;
for all episode $k = 1, 2, \dots$ do
    $t_k \leftarrow t$;
    Compute $\pi^k \leftarrow \operatorname{argmax}_\pi \tilde{g}_t(\pi)$ using EVI;
    repeat
        Execute action $A_t := \pi(X_t)$;
        Observe $R_t$ and next state $X_{t+1}$;
        $t \leftarrow t + 1$;
    until $\tilde{g}_t(\pi) \le \tilde{g}_t^* - \sqrt{\alpha \log(t)/t}$ or (Doubling Trick (DT));
end for
```

4.2. Number of Episodes

Actually, the main reason why bounded episodes in a learning algorithm are bad for learning is that if the number of policy changes is linear in T, then the regret will also be linear in T on some MDPs. Therefore, one of the main benefit of the doubling trick in classical episodic learning is to guarantee that the number of episodes is not too large. More precisely, it implies that the number of episodes grows as $\log T$. However, there is a big gap between a linear number of episodes and a logarithmic one. If we can design a learning algorithm whose number of episodes grows as $\sqrt{T}$, then the hope to keep a regret that grows as $\sqrt{T}$ remains intact. This is precisely what we show for UCRL-PT.

Since the main difference between classical episodic algorithms and UCRL-PT is the length of the episodes, the key new ingredient in the analysis of the regret of UCRL-PT is to bound the total number of episodes, i.e., the size K of the (random) set $\mathcal{K} := \{k : t_k \le T\}$ for some fixed $T \ge 1$.

Hoeffding confidence set In the rest of this section and to get precise statements, we consider the classical case where the confidence set is built using Hoeffding inequalities. However, keep in mind that the PT trick can be used for any kind of confidence set. Here, confidence bounds on rewards and transition probabilities ξ^r, ξ^p, are defined using Hoeffding-type concentration inequalities: for $i \in \{r, p\}$,

$$\xi^i(s, t) := \sqrt{\frac{\kappa_1^i \log(\kappa_2^i t)}{\max(1, s)}} \tag{10}$$

In the above, κ_1^i and κ_2^i are well chosen parameters to be specified later. $N_t(z) := \sum_{i=1}^{t-1} \mathbf{1}\{Z_i = z\}$ is the standard num-

[2] To be fair, the mentionned UCRL2B, KL-UCRL and UCRL3 are designed to have sublinear regret in *strong probability* hence they work with a fixed confidence threshold δ. For Theorem 1 to be applicable, these algorithms must be adapted to have guarantees on their *expected* regret. This is usually done by setting δ to the time-dependent $\delta(t) := \frac{1}{t}$.

ber of visits to the state-action pair $z = (x, a)$ at time t.

As usual the number of visits to the state-action pair $z = (x, a)$ at time t is $N_t(z) := \sum_{i=1}^{t-1} \mathbf{1}\{Z_i = z\}$ and the confidence bounds on the rewards and the transition probabilities ξ^r, ξ^p, are adapted to Hoeffding-type concentration inequalities: for $i \in \{r, p\}$, The associated time-dependent confidence set at time t is

$$\widetilde{\mathcal{M}}_t := \left\{ (\tilde{r}, \tilde{P}) : \begin{array}{c} \forall z \in \mathcal{Z}, |\tilde{r}(z) - \hat{r}_t(z)| \le \xi^r(N_t(z), t) \\ \text{and} \\ \forall z \in \mathcal{Z}, \left\|\tilde{P}(z) - \hat{P}_t(z)\right\|_1 \le \xi^p(N_t(z), t) \end{array} \right\}$$

This instance of UCRL-PT will be refered to as UCRL2-PT.

The good event. The *good event* is a refinement of the assertion "$M \in \widetilde{\mathcal{M}}_t$" which basically states that the confidence region is valid. The use of a high probability good event is a standard tool in the analysis of reinforcement learning algorithms. Here, our good event is stating that for all $t \in [1, T]$, $t' \in [t, T]$ and all $z \in \mathcal{Z}$,

$$\left\|\hat{P}_{t:t'}(z) - P(z)\right\|_1 \le \sqrt{4S \log(SAt'^4/\delta)} \qquad (11)$$

where $\hat{P}_{t:t'}(z)$ is the empirical distribution of the transition z over $[t, t']$, with a similar condition on rewards. Accordingly, if the event in (11) is $\mathcal{E}_{t,t',z}(T)$, then $\mathcal{E}(T) := \bigcap_{t=1}^{T} \bigcap_{t'=t}^{T} \bigcap_z \mathcal{E}_{t,t',z}(T)$. It can be shown (see Lemma 1) that $\mathcal{E}(T)$ holds with probability at least $1 - 2\delta$.

Confidence constants. The good event also tells how to pick the confidence constants κ_1, κ_2 in (10). Here, for a confidence level $\delta > 0$, we pick

$$\kappa_1 = 16S \quad \text{and} \quad \kappa_2 = SA/\delta \qquad (12)$$

Theorem 2. *For all $T \ge 1$, on the good event $\mathcal{E}(T)$, the number K of episodes of UCRL-PT is upper-bounded by*

$$K \le \frac{2^9 \cdot DS^{3/2}A \operatorname{sp}(r)}{\sqrt{\alpha}} \sqrt{T} \log(SAT) + \tilde{O}\left(T^{1/3}\right). \qquad (13)$$

Sketch of the proof of Theorem 2. The doubling trick accounts only for logarithmically many episodes which is negligible in front of the number of other episodes. We thus ignore episodes interrupted by (**DT**). The fact that episode k ends at time t_{k+1} implies that

$$\tilde{g}_{t_{k+1}}(\pi^k) + \sqrt{\frac{\alpha \log(t_{k+1})}{t_{k+1}}} \le \tilde{g}_{t_{k+1}}(\pi^{k+1}) \qquad (14)$$

Because π^k is optimistically optimal at time t_k, it means that over $[t_k, t_{k+1}]$, either $\tilde{g}_t(\pi^k)$ or $\tilde{g}_t(\pi^{k+1})$ has varied by about $\sqrt{\alpha \log(t_{k+1})/t_{k+1}}$. But here is the thing: the gain cannot vary too fast. Specifically, we show that if $\pi = \pi^k$ or π^{k+1}, then

$$\left\|\tilde{g}_{t_{k+1}}(\pi) - \tilde{g}_{t_k}(\pi)\right\| \le D\left(\left\|\hat{P}_{t_{k+1}} - \hat{P}_{t_k}\right\|_1 + \left\|\xi_{t_{k+1}} - \xi_{t_k}\right\|_\infty\right)$$

Therefore, from (14), we deduce that, over $[t_k, t_{k+1}]$, there must be a variation of (a) empirical kernels $\hat{P}_t$ or (b) optimistic bonuses ξ_t of order at least $D^{-1}\sqrt{\alpha \log(t_{k+1})/t_{k+1}}$. On the good event, these variations can be related to variations of time (i.e., $t_{k+1} - t_k$) and visit counts (i.e., $N_{t_{k+1}}(z) - N_{t_k}(z)$). We then derive a collection of inequalities that guarantees that, when there is a change of episode, visit counts or time increase enough relatively to $\alpha \log(t_{k+1})/t_{k+1}$, hence relatively to the a priori fixed $\alpha \log(T)/T$. The inequality that later accounts for the dominant part in the number of episodes is the following: For some $z \in \mathcal{Z}$,

$$N_{t_{k+1}}(z) \ge N_{t_k}(z)\left(1 + \frac{\alpha \log(T) \cdot N_{t_k}(z)}{2^{10} D^2 R_{\max}^2 ST \log(4SAT^3/\delta)}\right) \qquad (15)$$

By quantifying the growth of *integer-valued* sequences $\{u_k\}$ (here $u_k := N_{t_k}(z)$) satisfying an inequality such as (15), we deduce an upper bound of K in the form of (13). □

4.3. Regret Guarantees of UCRL2-PT

Once the number of episodes of UCRL2-PT is bounded, finding a bound on the regret follows well-known tracks. Here is an overall view of our approach used to bound the regret of UCRL2-PT.

Main line of the regret analysis. To simplify the analysis, we assume in the rest that rewards are bounded (i.e., that $q(z)$ has compact support) by some $R_{\max} \ge 0$.[3]

Here are the main lines of the regret decomposition (full details are in appendix A). First decompose Reg(T) as:

$$\sum_{k=1}^{K}\left[\underbrace{\left(\sum_{t=t_k}^{t_{k+1}-1}\left(g_M^* - \tilde{g}_{t_k}^*\right)\right)}_{\textcircled{1_k}} + \underbrace{\left(\sum_{t=t_k}^{t_{k+1}-1}\left(\tilde{g}_{t_k}^* - \tilde{g}_{t_k}^{\pi^k}\right)\right)}_{\textcircled{2_k}} + \underbrace{\left(\sum_{t=t_k}^{t_{k+1}-1}\left(\tilde{g}_{t_k}^{\pi^k} - r(Z_t)\right)\right)}_{\textcircled{3_k}}\right].$$

Under the good event, $\mathcal{E}(T) := \{\forall t \le T, M \in \widetilde{\mathcal{M}}_t\}$, the first term $\textcircled{1_k}$ is negative and the second term $\textcircled{2_k}$ is just 0 by choice of π^k. The last term $\textcircled{3_k}$ accounts for more work. Given π^k, we rely on the quantities introduced in Subsection 4.1 as well as the well-known gain-bias identity:

$$\tilde{g}_t(\pi^k) - \tilde{r}_t^{\pi^k} = (\tilde{P}_t^{\pi^k} - I)\tilde{h}_t^{\pi^k},$$

where $\tilde{r}_t^{\pi^k}$ is a short-hand for the $\mathcal{S}$-dimensional vector given by $\tilde{r}_t^{\pi^k}(x) := \tilde{r}_t(x, \pi^k(x))$. Therefore, introducing the interme-

[3] Many authors further assume that $R_{\max} = 1$, but we keep $R_{\max}$ in this work.

diate term $\tilde{h}^{\pi^k}_{t_k}(X_{t+1})$, we get

$$\begin{aligned}
\sum_{k=1}^{K} \textcircled{3_k} = &\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \left(\left(\tilde{P}^{\pi^k}_{t_k} - P^{\pi^k}_{t} \right) \tilde{h}^{\pi^k}_{t_k} \right)(X_t) && \left(\textstyle\sum_k \textcircled{4_k} \right) \\
&+ \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \left(\left(P^{\pi^k}_{t_k} \tilde{h}^{\pi^k}_{t_k} \right)(X_t) - \tilde{h}^{\pi^k}_{t_k}(X_{t+1}) \right) && \left(\textstyle\sum_k \textcircled{5_k} \right) \\
&+ \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \left(\tilde{h}^{\pi^k}_{t_k}(X_{t+1}) - \tilde{h}^{\pi^k}_{t_k}(X_t) \right) && \left(\textstyle\sum_k \textcircled{6_k} \right)
\end{aligned}$$

Terms $\textcircled{$4_k$}$ and $\textcircled{$5_k$}$ are estimated using the method of (Auer et al., 2009), refer to Appendix A for a complete proof. To give a few directions, term $\textcircled{$4_k$}$ is bounded using Weissman's concentration inequality (settling the coefficients κ_1 and κ_2 used in the bonus) and by bounding the span of $\tilde{h}^{\pi^k}_{t_k}$ w.r.t. to the diameter; and term $\textcircled{$5_k$}$ is bounded using Azuma-Hoeffding's inequality for martingales difference sequences (see Lemma 15). This leads to the upper bound:

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \left(\textcircled{4_k} + \textcircled{5_k} \right) \leq\ & 2D \operatorname{sp}(r) \sqrt{2SA\kappa_1 T \log(\kappa_2 T)} \\
&+ D \operatorname{sp}(r) \sqrt{2T \log(1/\delta)}
\end{aligned}$$

with, remember, κ_1 of order S. Finally, the last term $\textcircled{$6_k$}$ is telescopic, and bounded using with the span of the bias:

$$\sum_{k=1}^{K} \textcircled{6_k} \leq \sum_{k=1}^{K} \operatorname{sp}(\tilde{h}^{\pi^k}_{t_k}) \leq KD \operatorname{sp}(r).$$

In total this gives the following upper-bound on the regret.

Theorem 3. *With probability at least* $1 - \delta$ *it holds that for any initial state and any* $T > 1$, *the regret of UCRL2-PT is bounded by*

$$\operatorname{Reg}(T) = \tilde{O}\left(R_{\max} DS \sqrt{AT \log\left(\tfrac{1}{\delta}\right)} + R_{\max} KD \right).$$

The detailed proof of the theorem is given in Appendix A.

How to tune α. The dependence of the regret in α is hidden in the number of episodes, K. For any $\alpha > SAD^2$ the second term in the bound is dominated by the first one, and the regret becomes

$$\operatorname{Reg}(T) = \tilde{O}\left(R_{\max} DS \sqrt{AT \log\left(\tfrac{1}{\delta}\right)} \right). \tag{16}$$

Notice that the bound (16) on the regret of UCRL2-PT is the same as for UCRL2. Actually this should not come as a surprise because when α goes to infinity, UCRL-PT with Hoeffding confidence sets and UCRL2 are the same algorithm.

If D is not known, a possible choice is $\alpha = A^2 S^3$. In this case,

$$\operatorname{Reg}(T) = \tilde{O}\left(R_{\max} D^2 \sqrt{T \log\left(\tfrac{1}{\delta}\right)} \right). \tag{17}$$

Expected Regret. Theorem 3 gives an upper bound on the regret with high probability $1 - \delta$. This can be used to get a similar bound in $O(\sqrt{T})$ for the expected regret by using the usual trick: choose $\delta = 1/T$ and discard the first $\sqrt{T}$ terms in the regret similarly to (Auer et al., 2009).

4.4. Regret of Exploration of UCRL2-PT on DMDP

Here we prove that UCRL2-PT has a small regret of exploration when the MDP is *deterministic* (DMDP), meaning that all transitions are deterministic. In this setting, we can assume without loss of generality that kernels $P(\cdot|z)$ are known in advance and that only rewards have to be learned. The optimal policies of a DMDP are hence characterized by the mean reward vector $r \in \mathbb{R}^{\mathcal{Z}}$. We further assume that rewards are Bernoulli, although compact support is enough.

Non-degeneracy assumption. We further assume that the MDP is *non-degenerate*. This is a condition on the uniqueness of the optimal policy. More precisely a DMDP is non-degenerate if there is a unique bias optimal policy π^* that has a unique terminal cycle (that is, a unique recurrent class). In practice, if the mean rewards are chosen at random (according any absolutely continuous measure w.r.t. the Lebesgue measure), the induced MDP is non-degenerate with probability one, see (Boone & Gaujal, 2023).

The lazy version of UCRL2-PT. The algorithm needs to be slightly adapted into a lazy version (Algorithm 3) in order to guarantee some stability properties that help with the analysis. In the lazy version, the policy doesn't change if the current transition Z_t has not been visited at least twice in the current episode. This patch doesn't harm the regret guarantees of Theorem 3: First, it decreases the number of episodes K and in fact, the current proof of the bound on K (see Theorem 2) fits without modifications. Then, the regret analysis works as is.

Algorithm 3 UCRL2-PT, lazy version

```
Input: Confidence $\delta > 0$ and $\alpha > 0$;
for all episodes $k = 1, 2, \ldots$ do
    $t_k \leftarrow t$;
    $\pi^k \leftarrow \operatorname{argmax}_\pi \tilde{g}_t(\pi)$;
    repeat
        Execute action $A_t = \pi^k(X_t)$ ;
        Observe $R_t$ and next state $X_{t+1}$. ;
        $t \leftarrow t + 1$;
    until $N_t(Z_t) - N_{t_k}(Z_t) > 1$ and $[\tilde{g}_t(\pi^k) + \sqrt{\frac{\alpha \log(t)}{t}} \leq \tilde{g}^*_t$ or (DT)]
end for
```

Before stating the theorem on the regret of exploration, let us compare the behavior of UCRL2-PT and UCRL2 over the bandit example given in Figure 1 with two Gaussian arms, the good a_1 and the bad a_2 with respective means 1 and 0.9. Figure 3 shows the regret of both algorithms and one can notice that the bad episodes are shorter for UCRL2-PT, suggesting a better regret of exploration. A more detailed view of the behavior of UCRL2-PT is displayed in Figure 4. The current policy π^k corresponds to which arm is drawn over $[t_k, t_{k+1}]$. We know that the algorithm changes of episode when $\tilde{g}_t(\pi^k) + \sqrt{\alpha \log(t)/t} \leq \tilde{g}_t^*$, so the behavior of the algorithm is driven by the optimistic gap $\tilde{\Delta}_t(a_1; a_2) := \tilde{g}_t(a_1) - \tilde{g}_t(a_2)$. This quantity is plotted on a run of UCRL2-PT in Figure 4.

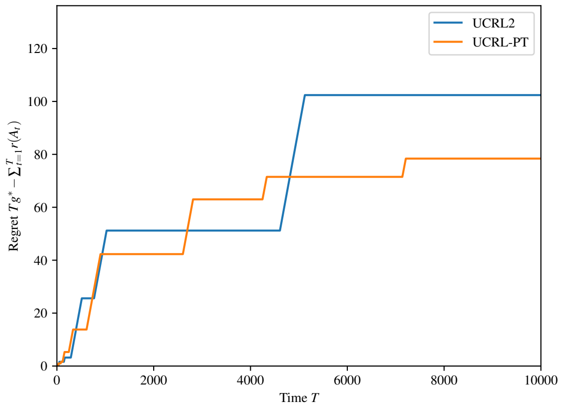

Figure 3: UCRL2 against UCRL2-PT on the example of Figure 1. The seed responsible for the generation of rewards is the same in both algorithms. On can observe that the bad episodes of UCRL2-PT are shorter.

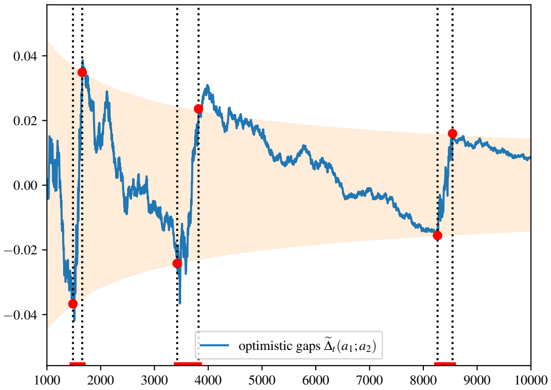

Figure 4: Optimistic Bellman gaps $\tilde{\Delta}_t(a_1; a_2) := \tilde{g}_t(a_1) - \tilde{g}_t(a_2)$ over time. We plot the convex hull of $\pm\sqrt{\alpha \log(t)/t}$ (orange region). Change of episodes are indicated by red points. Bad episodes are highlighted in red on the time-axis.

On a time-window $[t, t + T]$, the regret $\mathrm{RegExp}(t, t + T)$ is proportional to the amount of time spent drawing arm a_2, highlighted in red on the time-axis. A bad episode happen when $\tilde{\Delta}_t(a_1; a_2)$ goes below $-\sqrt{\alpha \log(t)/t}$, then is interrupted as soon as it goes above $\sqrt{\alpha \log(t)/t}$ again. When a bad episode occurs, arm a_2 is triggered, the empirical estimate of its value is updated over time, and its bonus decreases. The update of its empirical estimate is the noise that we observe on Figure 4 and is clearly non-negligible with respect to the upward drift induced by a decrease of the optimistic bonus of a_2. Still, that drift forces the bad episode to end rather quickly.

We then switch to a good episode where the noise amplitude is much smaller. This is because the optimal a_1 is visited more often than a_2 (the regret is small), hence its empirical value changes slower. There is also a drift (there downward) due to the evolution of bonuses. Yet the drift is weaker, because the visit counts of a_1 is much higher, so the associated bonus will decrease slower than the one of a_2 did, and the evolution of the bonus of a_2 is only due to a $\sqrt{\log t}$ over the good episode, which is almost constant over time.

Therefore, good episodes last, bad episodes don't; The regret of exploration is small. The precise – and much more general – statement is given in the following theorem.

Theorem 4. *Consider the lazy version of UCRL2-PT. Let $\alpha, \delta > 0$ the parameters of UCRL2-PT with $\kappa_1 := 16S$ and $\kappa_2 := SA/\delta$. For any non-degenerate MDP M with Bernoulli rewards, there exists a constant $C(M) > 0$ such that*

$$\mathrm{RegExp}(\text{UCRL2-PT}, M; T) \leq C(M)\left(1 + \log(T)\right). \quad (18)$$

The analysis is model dependent and relies on how the confidence region (hence the optimistic values of policies) are subject to change within small time ranges initialized at the beginning of an exploration episode, see Appendix B.

5. Experimental Validations

To conclude the paper, we study the behavior of (**PT**)-based algorithms in several MDPs. We consider two types of environments: the well-known River-Swim (see Figure 5) and randomly generated ergodic MDPs.

5.1. Running the Performance Test (PT) on a Laptop

Let us start with a discussion on the implementation of (**PT**). In practice, the (**PT**) variants of classical RL algorithms are much heavier to run. This is due to the fact that at each timestep, one must run EVI twice to compute both $\tilde{g}_t^*$ and $\tilde{g}_t(\pi^k)$. This additional computation can make (**PT**) impossible to rely on when the numbers of states and actions grow large. Thankfully, (**PT**)-based algorithms can be substantially accelerated with the combination of two methods.

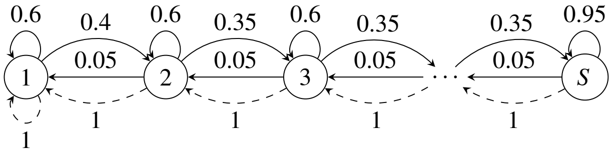

Figure 5: The River-Swim environment with S states. There are two actions, $\mathcal{A} = \{\text{left}, \text{right}\}$ whose associated transition are respectively drawned with dashed and solid lines. The only non-zero reward is the the state-action pair (S, right) with $r(S, \text{right}) = 1$, hence the optimal policy always picks the action "right".

	Original	Pure (**PT**)	(**PT**) + M	(**PT**) + S	(**PT**) + M + S
UCRL2	200k	1.0k	5.2k	74k	82k
KL-UCRL	157k	0.2k	2.7k	36k	62k
UCRL2B	167k	0.5k	2.9k	59k	77k

Table 1: Iterations per second of UCRL2, KL-UCRL and UCRL2B; the originals and the (**PT**) corrections with various acceleration options. The environment is a 5-state RiverSwim. These values have been obtained by running each algorithm for 100k iterations and take the average per-step time.

(1) **Memorisation** (M): From t to $t+1$, the confidence region is barely changed. Therefore, by initializing the EVI at time $t+1$ with the result of EVI at time t, one should expect EVI to converge much faster. This doesn't modify the algorithm's behavior.

(2) **Sparse** (**PT**) (S): Even if EVI converges almost instantly, running EVI at each time-step significantly slows down the algorithm. To address this, instead of always checking (**PT**), only test it when $t - t_k$ is a power of 2. Formally, (**PT**) is replaced by (**PT***):

$$\log_2(t - t_k) \in \mathbb{N} \quad \text{and} \quad \tilde{g}_t(\pi^k) + \sqrt{\frac{\alpha \log(t)}{t}} \leq \tilde{g}_t^*, \quad \textbf{(PT*)}$$

Although this modification slightly alter the behavior of the algorithm, its analysis is similar to (**PT**)'s.

As shown in Table 1, the combination of these two modifications makes the runtimes of (**PT**) variants acceptable in comparison to the originals.

5.2. The Performance Trick on the Regret Benchmark

The first set of experiments displayed in Figure 6 shows that introducing the performance test in the three learning algorithms (UCRL2, UCRL2B and KL-UCRL) does not harm the regret in both the RiverSwim and the random MDP examples. Actually, the PT-versions seem to perform slightly better than the originals but there is a lack of satistical evidence to fully support the claim.

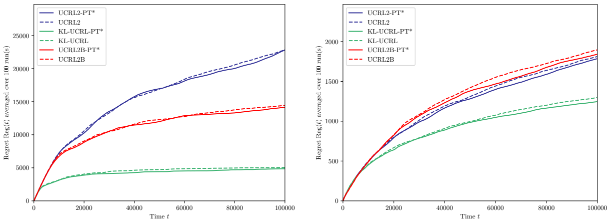

Figure 6: Regret comparison of original and PT-version of the algorithms. On the left, a 5-state RiverSwim; and on the right, a 5-state random MDP. α is set to 1.

5.3. Sensitivity to α

The hyperparameter α tunes the sensitivity of the performance trick. Experiments show that when α is too large, the term $\sqrt{\alpha \log(t)/t}$ takes too much time to decay and (**PT**) rarely triggers within reasonable time-horizons. When α is small, most episodes are ended with (**PT**) and exploration is efficient. Taking α too small is dangerous however. In the one hand, the regret scales with $\alpha^{-1/2}$ according to Theorem 2; On the other hand, when $\alpha = 0$, the algorithm always picks the optimistically optimal policy and behaves like UCB, whose regret will grow linearly on some instances (Ortner, 2010b). These experiments suggests to pick α as small as possible within an a priori fixed acceptable range.

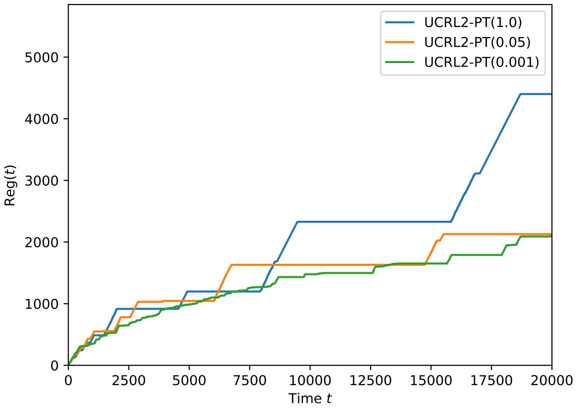

Figure 7: (Gap) regret of UCRL2-PT with various values of α on a RiverSwim(3), one run. The gap-regret is given by $\sum_{u=1}^{t} V^*(X_u) - Q^*(Z_u)$, where $Q^*(x,a) := r(x,a) + p(x,a)h^*$ is the Q-value of (x,a) and $V^*(x) := \max_a Q^*(x,a)$ is the V-value. In expectation, the gap-regret and the regret differ by at most $\text{sp}(h^*)$, but the gap-regret is much less noisy.

References

Auer, P. and Ortner, R. Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning. *Proceedings of the 19th International Conference on Neural*

Information Processing Systems, December 2006. doi: 10.7551/mitpress/7503.003.0011.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-Time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.*, 47(2–3), May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352.

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

Bartlett, P. L. and Tewari, A. 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, pp. 35–42, Arlington, Virginia, USA, June 2009. AUAI Press. ISBN 978-0-9749039-5-8.

Boone, V. and Gaujal, B. Identification of Blackwell Optimal Policies for Deterministic MDPs. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7392–7424. PMLR, April 2023.

Bourel, H., Maillard, O., and Talebi, M. S. Tightening Exploration in Upper Confidence Reinforcement Learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1056–1066. PMLR, July 2020.

Burnetas, A. and Katehakis, M. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 1997.

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. *arXiv:1703.07710 [cs, stat]*, January 2018.

Filippi, S., Cappé, O., and Garivier, A. Optimism in Reinforcement Learning and Kullback-Leibler Divergence. *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 115–122, September 2010. doi: 10.1109/ALLERTON.2010.5706896.

Fruit, R., Pirotta, M., and Lazaric, A. Improved Analysis of UCRL2 with Empirical Bernstein Inequality. *ArXiv*, abs/2007.05456, 2020.

Kaufmann, E. and Koolen, W. M. Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. *Journal of Machine Learning Research*, December 2021.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6 (1):4–22, March 1985. ISSN 01968858. doi: 10.1016/0196-8858(85)90002-8.

Ortner, R. Online regret bounds for markov decision processes with deterministic transitions. *Theoretical Computer Science*, 411(29-30):2684–2695, 2010a.

Ortner, R. Online regret bounds for Markov decision processes with deterministic transitions. *Theoretical Computer Science*, 411(29):2684–2695, 2010b. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2010.04.005.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown markov decision processes: A thompson sampling approach. In *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

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

Puterman, M. L. *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. doi: 10.1002/9780470316887.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Thompson, W. R. On the Likelihood that One Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3-4):285–294, December 1933. ISSN 0006-3444. doi: 10.1093/biomet/25.3-4.285.

Tossou, A., Basu, D., and Dimitrakakis, C. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities, 2019a.

Tossou, A., Basu, D., and Dimitrakakis, C. Near-optimal Optimistic Reinforcement Learning using Empirical Bernstein Inequalities. *arXiv:1905.12425 [cs, stat]*, December 2019b.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. Model-free Reinforcement Learning in Infinite-horizon Average-reward Markov Decision Processes. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10170–10180. PMLR, July 2020.

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

A. Bound on the Regret of UCRL-PT

We study the regret of UCRL2-PT.

To simplify the presentation, we consider that the rewards are *known* and only care abound learning transitions. This assumption is mainly for convenience. The analysis of the regret when rewards are also unknown is similar, and the additional regret induced in the analysis is at most of the same order as the one induced by transitions alone. In fact, it is smaller – this meets a traditional belief that "transitions are harder to learn".

Assumption 1. Along the analysis, we assume that rewards are known.

Long story short: This assumption has no impact on the generality of our regret bound nor of our method.

A.1. General Concepts for the Analysis of UCRL-PT

- **Confidence intervals.** Following Assumption 1, the confidence intervals on transitions are simplified to

$$\xi_t(z) = \sqrt{\frac{\kappa_1 \log(\kappa_2 t)}{\max\{1, N_t(z)\}}} \tag{A.1}$$

  instead of $\xi_t^p(z) = \sqrt{\kappa_1^p \log(\kappa_2^p t)/N_t(z)}$. The confidence region is just

$$\widetilde{\mathcal{M}}_t := \left\{\tilde{P} : \forall z \in \mathcal{Z}, \left\|\tilde{P}(z) - \hat{P}_t(z)\right\|_1 \leq \xi_t(z)\right\}.$$

- **Optimistic models.** A important concept of the analysis of UCRL-PT – which is new regarding the analysis of RL algorithms, is the notion of the optimistic model *relative* to a policy π. Given a policy π, the optimistic model under π at time t is

$$\widetilde{\mathcal{M}}_t^\pi = \tilde{P}_t^\pi := \operatorname{argmax}\left\{\max_x g_x(\pi; \tilde{M}) : \tilde{M} \in \widetilde{\mathcal{M}}_t\right\} \tag{A.2}$$

  This model differs from a policy to another, and in particular, $\tilde{g}_t(\pi^k)$ and $\tilde{g}_t^*$ refer to two different MDPs that have absolutely no reason to resemble one another. Also, $\tilde{P}_t^\pi$ may not evolve "continuously" with respect to time.

- **The good event.** The *good event* is a refinement of the assertion "$M \in \widetilde{\mathcal{M}}_t$" which basically states that the confidence region is valid. The use of a high probability good event is a standard tool in the analysis of reinforcement learning algorithms. Here, our good event is stating that for all $t \in [1, T]$, $t' \in [t, T]$ and all $z \in \mathcal{Z}$,

$$\mathcal{E}(T) := \bigcap_{t=1}^{T} \bigcap_{t'=t+1}^{T} \bigcap_{z \in \mathcal{Z}} \left\{\left\|\hat{P}_{t:t'}(z) - P(z)\right\|_1 \leq \sqrt{4S \log(SAt'^4/\delta)}\right\} \tag{A.3}$$

  where $\hat{P}_{t:t'}(z)$ is the empirical distribution of transition z over $[t, t']$, that is, define first $N_{t:t'}(z) := \sum_{i=t}^{t'-1} \mathbf{1}\{Z_i = z\}$ then set $\hat{P}_{t:t'}(z) = N_{t:t'}(z)^{-1} \sum_{i=t}^{t'-1} \mathrm{Dirac}(X_{i+1})\mathbf{1}\{Z_i = z\}$. It can be shown (see Lemma 1) that $\mathcal{E}(T)$ holds with probability at least $1 - 2\delta$.

Lemma 1 (Good event)**.** *Consider the good event $\mathcal{E}(T)$ given as in (A.3). For all $T \geq 1$, $\mathbb{P}(\mathcal{E}(T)) \geq 1 - 2\delta$.*

Proof. This is a straight forward computation. For all $z \in \mathcal{Z}$, introduce $\{W_k(z)\}$ i.i.d. random variables of distribution $P(z)$.

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}(T)^{\complement}\right) &\le \mathbb{P}\left(\bigcup_{t=1}^{T}\bigcup_{t'=t+1}^{T}\bigcup_{z\in\mathcal{Z}}\left\{\left\|\hat{P}_{t:t'}(z)-P(z)\right\|_1 \le \sqrt{4S\log(SAt'^4/\delta)}\right\}\right) && \text{(A.4)}\\
&\le \sum_{t=1}^{T}\sum_{t'=t+1}^{T}\sum_{z\in\mathcal{Z}}\mathbb{P}\left(\bigcup_{n=0}^{t'-t}\left\{N_{t:t'}(z)=n \text{ and } \left\|\hat{P}_{t:t'}(z)-P(z)\right\|_1 \le \sqrt{4S\log(SAt'^4/\delta)}\right\}\right) && \text{(A.5)}\\
&\le \sum_{t=1}^{T}\sum_{t'=t+1}^{T}\sum_{z\in\mathcal{Z}}\mathbb{P}\left(\bigcup_{n=0}^{t'-t-1}\left\{\|W_n(z)-nP(z)\|_1 \le \sqrt{4S\max\{1,n\}\log(SAt'^4/\delta)}\right\}\right) && \text{(A.6)}\\
&\le \sum_{t=1}^{T}\sum_{t'=t+1}^{T}\sum_{z\in\mathcal{Z}}\sum_{n=0}^{t'-t-1}\mathbb{P}\left\{\|W_n(z)-nP(z)\|_1 \le \sqrt{4S\max\{1,n\}\log(SAt'^4/\delta)}\right\} && \text{(A.7)}\\
&\le \sum_{t=1}^{T}\sum_{t'=t+1}^{T}\sum_{z\in\mathcal{Z}}\sum_{n=0}^{t'-t-1}\frac{\delta}{SAt'^4} && \text{(A.8)}\\
&\le \sum_{t=1}^{T}\sum_{t'=1}^{T}\frac{\delta}{t'^3}\mathbf{1}\{t'>t\} = \sum_{t'=1}^{T}\frac{\delta}{t'^3}\sum_{t=1}^{T}\mathbf{1}\{t'>t\} = \sum_{t'=1}^{T}\frac{\delta}{t'^2} \le 2\delta. && \square
\end{aligned}$$

A.2. Bound on the Number of Episodes

We consider the vanilla version of the algorithm enriched with explicit episode management, see Algorithm 2. This section is dedicated to a proof of Theorem 2 that we restate below.

Theorem 2. *For all $T \ge 1$, on the good event $\mathcal{E}(T)$, the number of episodes of UCRL2-PT is bounded above as*

$$K \le \frac{2^9 \cdot DS^{3/2}AR_{\max}}{\sqrt{\alpha}}\sqrt{T}\log(SAT) + \tilde{O}\left(T^{1/3}\right). \tag{A.9}$$

General proof strategy. The doubling trick accounts only for logarithmically many episodes which is negligible in front of the number of other episodes. Concerning other episodes, the fact that episode k ends at time t_{k+1} implies that

$$\tilde{g}_{t_{k+1}}(\pi^k) + \sqrt{\frac{\alpha\log(t_{k+1})}{t_{k+1}}} \le \tilde{g}_{t_{k+1}}(\pi^{k+1}) \tag{A.10}$$

Because π^k is optimistically optimal at time t_k, it means that over $[t_k, t_{k+1}]$, either $\tilde{g}_t(\pi^k)$ or $\tilde{g}_t(\pi^{k+1})$ has varied by about $\sqrt{\alpha\log(t_{k+1})/t_{k+1}}$. But here is the thing: the gain cannot vary too fast. Specifically, we show that (STEP 1) if $\pi = \pi^k$ or π^{k+1}, then

$$\left\|\tilde{g}_{t_{k+1}}(\pi) - \tilde{g}_{t_k}(\pi)\right\| \le D\left(\left\|\hat{P}_{t_{k+1}} - \hat{P}_{t_k}\right\|_1 + \left\|\xi_{t_{k+1}} - \xi_{t_k}\right\|_\infty\right)$$

Therefore, from (A.10), we deduce that, over $[t_k, t_{k+1}]$, there must be a variation of (STEP 2) empirical kernels $\hat{P}_t$ or (STEP 3) optimistic bonuses ξ_t of order at least $D^{-1}\sqrt{\alpha\log(t_{k+1})/t_{k+1}}$. On the good event, these variations can be related to variations of time (i.e., $t_{k+1} - t_k$) and visit counts (i.e., $N_{t_{k+1}}(z) - N_{t_k}(z)$). We then derive (STEP 4) a collection of inequalities that guarantees that, when there is a change of episode, visit counts or time increase enough relatively to $\alpha\log(t_{k+1})/t_{k+1}$, hence relatively to the a priori fixed $\alpha\log(T)/T$. The inequality that later accounts for the dominant part in the number of episodes is the following: For some $z \in \mathcal{Z}$,

$$N_{t_{k+1}}(z) \ge N_{t_k}(z)\left(1 + \frac{\alpha\log(T)\cdot N_{t_k}(z)}{2^{10}D^2R_{\max}^2 ST\log(4SAT^3/\delta)}\right) \tag{A.11}$$

Finally (STEP 5), by quantifying the growth of *integer-valued* sequences $\{u_k\}$ (here $u_k := N_{t_k}(z)$) satisfying an inequality such as (A.11), we deduce an upper bound of K in the form of (13).

▸ **STEP 1: Fundamental equation of episode renewal** (A.27) The goal is to show that if there is a change of episode, then necessarily, transition kernels or confidence bound must have moved by some tractable quantity. Introduce the set of *standard* episodes $\mathcal{K}_0$

$$\mathcal{K}_0 := \{k \in \mathcal{K} : \forall z \in \mathcal{Z}, N_{t_{k+1}}(z) < 2N_{t_k}(z)\}. \tag{A.12}$$

Therefore, an episode k is non-standard if either (a) there is $z \in \mathcal{Z}$ visited on $[t_k, t_{k+1} - 1]$ such that $N_{t_k}(z) = 0$ or (b) there is $z \in \mathcal{Z}$ such that $N_{t_k}(z) \geq 1$ and $N_{t_{k+1}}(z) \geq 2N_{t_k}(z)$. Accordingly, non-standard episodes are exactly those interrupted by the doubling trick (**DT**) rule. Intuitively, because on non-standard episodes, one transition at least doubles its visit counts, $\mathcal{K} \setminus \mathcal{K}_0$ is of cardinality at most logarithmic. It will be negligible in front of bounds on the cardinality of $\mathcal{K}_0$.

In practice, we will show the following:

Lemma 2. *If $k \in \mathcal{K}_0$ then on the good event $\mathcal{E}(T)$,*

$$\frac{1}{2}\sqrt{\frac{\alpha \log(T)}{T}} \leq 2D \operatorname{sp}(r)\left(\sup_z \left\|\hat{P}_{t_k}(z) - \hat{P}_{t_{k+1}}(z)\right\|_1 + \sup_z \left\|\xi_{t_k}(z) - \xi_{t_{k+1}}(z)\right\|_\infty\right) \tag{A.13}$$

Hence, we relate a change of episode to a variation of either (a) an empirical transition kernel $\hat{P}_t(z)$; or (b) a transition bonus $\xi_t(z)$. In both cases, this will translate into an increase of visit count (i.e., $N_{t_{k+1}}(z) - N_{t_k}(z)$ large) or in time (i.e., $t_{k+1} - t_k$ large), where the so-called increase is relative to the time barrier T, see STEP 4.

Proof of Lemma 2. By definition of $k \in \mathcal{K}_0$, there is a change of episode when

$$\tilde{g}_{t_{k+1}}(\pi^k) + \sqrt{\frac{\alpha \log(t_{k+1})}{t_{k+1}}} \leq \tilde{g}^*_{t_{k+1}} = \tilde{g}_{t_{k+1}}(\pi^{k+1}) \tag{A.14}$$

As we know that $\tilde{g}_{t_k}(\pi^k) = \tilde{g}^*_{t_k}$, we can write the above as follows:

$$\tilde{g}^*_{t_k}(\pi^k) + \left[\tilde{g}_{t_{k+1}}(\pi^k) - \tilde{g}_{t_k}(\pi^k)\right] + \sqrt{\frac{\alpha \log(t_{k+1})}{t_{k+1}}} \leq \tilde{g}_{t_k}(\pi^{k+1}) + \left[\tilde{g}_{t_{k+1}}(\pi^{k+1}) - \tilde{g}_{t_k}(\pi^{k+1})\right] \tag{A.15}$$

or, equivalently,

$$\left[\mathcal{D}_h \tilde{g}_t(\pi^k)\right](t_k) - \left[\mathcal{D}_h \tilde{g}_t(\pi^{k+1})\right](t_k) \leq \tilde{g}_{t_k}(\pi^{k+1}) - \tilde{g}_{t_k}(\pi^k) - \sqrt{\frac{\alpha \log(t_{k+1})}{t_{k+1}}} \tag{A.16}$$

$$\leq -\sqrt{\frac{\alpha \log(t_{k+1})}{t_{k+1}}} \leq -\sqrt{\frac{\alpha \log(T)}{T}} \tag{A.17}$$

where (A.17) is because by definition of π^k, $\tilde{g}_{t_k}(\pi^k) = \tilde{g}^*_{t_k}$ so $\tilde{g}_{t_k}(\pi^{k+1}) - \tilde{g}_{t_k}(\pi^k) \leq 0$. Therefore, a change of episode necessarily implies one of the two cases below:

$$\left[\mathcal{D}_h \tilde{g}_t(\pi^k)\right](t_k) \leq -\frac{1}{2}\sqrt{\frac{\alpha \log(T)}{T}} \tag{Type I episodes}$$

$$\left[\mathcal{D}_h \tilde{g}_t(\pi^{k+1})\right](t_k) \geq \frac{1}{2}\sqrt{\frac{\alpha \log(T)}{T}} \tag{Type II episodes}$$

that we respectively refer to type I and type II episodes. In the following, recall that $\tilde{P}^\pi_t$ the optimistic transition model of the policy π at time t, i.e., the transition matrix of the MDP $\tilde{M}^\pi_t \in \widetilde{\mathcal{M}}_t$ achieving $\tilde{g}_t(\pi)$.

In general $\tilde{P}^{\pi^k}_{t_k} - \tilde{P}^{\pi^k}_{t_{k+1}}$ is large and the optimistic model of a given policy may be hard to track over time. But, this not an issue. Introduce the Hausdorff distance on subsets of $\mathcal{M}$, $d_\mathcal{H}(\mathcal{U}, \mathcal{V}) := \max\{\sup_{P_1} \inf_{P_2} \|P_1 - P_2\|_1, \sup_v \inf_u \|u - v\|_1\}$. Recall that $\widetilde{\mathcal{M}}^\pi_t := \left\{\tilde{P}^\pi : \forall x \in \mathcal{S}, \left\|\tilde{P}^\pi(x) - \hat{P}^\pi_t(x)\right\|_1 \leq \xi^\pi_t(x)\right\}$. One checks that

$$\begin{aligned} d_\mathcal{H}\left(\tilde{\mathcal{M}}^{\pi^k}_{t_k}, \tilde{\mathcal{M}}^{\pi^k}_{t_{k+1}}\right) &\leq \left\|\hat{P}^{\pi^k}_{t_k} - \hat{P}^{\pi^k}_{t_{k+1}}\right\|_1 + \left\|\xi^{\pi^k}_{t_k} - \xi^{\pi^k}_{t_{k+1}}\right\|_\infty \\ &= \sup_{z \in \pi} \left\|\hat{P}_{t_k}(z) - \hat{P}_{t_{k+1}}(z)\right\|_1 + \sup_{z \in \pi} \left\|\xi_{t_k}(z) - \xi_{t_{k+1}}(z)\right\|_\infty. \end{aligned} \tag{A.18}$$

Write $\mathrm{Proj}_{\mathcal{U}}(\cdot)$ a projection on $\mathcal{U} \subseteq \mathcal{M}$ for one norm, i.e., $\mathrm{Proj}_{\mathcal{U}}(P_1)$ is any $P_2 \in \mathcal{U}$ minimizing the distance to P_1. In particular, for all $P_1 \in \mathcal{V}$, we have $\|P_1 - \mathrm{Proj}_{\mathcal{U}}(P_1)\|_1 \leq d_{\mathcal{H}}(\mathcal{U}, \mathcal{V})$. Now applying (A.18) to (Type I episodes) and (Type II episodes). For (Type I episodes), we have

$$
\begin{aligned}
\frac{1}{2}\sqrt{\frac{\alpha \log(T)}{T}} &\leq \tilde{g}_{t_k}(\pi^k) - \tilde{g}_{t_{k+1}}(\pi^k) = g\left(\pi^k; \tilde{P}^{\pi^k}_{t_k}\right) - g\left(\pi^k; \tilde{P}^{\pi^k}_{t_{k+1}}\right) && \text{(A.19)}\\
&\leq g\left(\pi^k; \tilde{P}^{\pi^k}_{t_k}\right) - g\left(\pi^k; \mathrm{Proj}_{\widetilde{\mathcal{M}}^{\pi^k}_{t_{k+1}}}\left(\tilde{P}^{\pi^k}_{t_k}\right)\right) && \text{(A.20)}\\
&\leq 2\,\mathrm{sp}\left(h\left(\pi^k; \tilde{P}^{\pi^k}_{t_k}\right)\right)\left(\sup_{z \in \pi^k}\left\|\hat{P}_{t_k}(z) - \hat{P}_{t_{k+1}}(z)\right\|_1 + \sup_{z \in \pi^k}\left\|\xi_{t_k}(z) - \xi_{t_{k+1}}(z)\right\|_\infty\right) && \text{(A.21)}\\
&\leq 2D\,\mathrm{sp}(r)\left(\sup_{z}\left\|\hat{P}_{t_k}(z) - \hat{P}_{t_{k+1}}(z)\right\|_1 + \sup_{z}\left\|\xi_{t_k}(z) - \xi_{t_{k+1}}(z)\right\|_\infty\right) && \text{(A.22)}
\end{aligned}
$$

Above, (A.20) follows from $g(\pi^k; \tilde{P}^{\pi^k}_{t_{k+1}}) \geq g(\pi^k; \mathrm{Proj}_{\widetilde{\mathcal{M}}^{\pi^k}_{t_{k+1}}}(\tilde{P}^{\pi^k}_{t_k}))$ which is from the definition of $\tilde{P}^{\pi^k}_{t_{k+1}}$; (A.21) is an application of Lemma 12 that bounds the variations of the gain with respect to the kernel variations; and (A.22) is by property of EVI on the good event. With a similar computation, for (Type II episodes), we have

$$
\begin{aligned}
\frac{1}{2}\sqrt{\frac{\alpha \log(T)}{T}} &\leq \tilde{g}_{t_{k+1}}(\pi^{k+1}) - \tilde{g}_{t_k}(\pi^{k+1}) = g\left(\pi^{k+1}; \tilde{P}^{\pi^{k+1}}_{t_{k+1}}\right) - g\left(\pi^{k+1}; \tilde{P}^{\pi^{k+1}}_{t_k}\right) && \text{(A.23)}\\
&\leq g\left(\pi^{k+1}; \tilde{P}^{\pi^{k+1}}_{t_{k+1}}\right) - g\left(\pi^{k+1}; \mathrm{Proj}_{\widetilde{\mathcal{M}}^{\pi^{k+1}}_{t_{k+1}}}\left(\tilde{P}^{\pi^{k+1}}_{t_k}\right)\right) && \text{(A.24)}\\
&\leq 2\,\mathrm{sp}\left(h\left(\pi^{k+1}; \tilde{P}^{\pi^{k+1}}_{t_{k+1}}\right)\right)\Bigg(\sup_{z \in \pi^{k+1}}\left\|\hat{P}_{t_k}(z) - \hat{P}_{t_{k+1}}(z)\right\|_1 \\
&\qquad\qquad + \sup_{z \in \pi^{k+1}}\left\|\xi_{t_k}(z) - \xi_{t_{k+1}}(z)\right\|_\infty\Bigg) && \text{(A.25)}\\
&\leq 2D\,\mathrm{sp}(r)\left(\sup_{z}\left\|\hat{P}_{t_k}(z) - \hat{P}_{t_{k+1}}(z)\right\|_1 + \sup_{z}\left\|\xi_{t_k}(z) - \xi_{t_{k+1}}(z)\right\|_\infty\right) && \text{(A.26)}
\end{aligned}
$$

Accordingly, in both (Type I episodes) and (Type II episodes), on the good event, we have

$$
\frac{1}{2}\sqrt{\frac{\alpha \log(T)}{T}} \leq 2D\,\mathrm{sp}(r)\left(\sup_{z}\left\|\hat{P}_{t_k}(z) - \hat{P}_{t_{k+1}}(z)\right\|_1 + \sup_{z}\left\|\xi_{t_k}(z) - \xi_{t_{k+1}}(z)\right\|_\infty\right) \tag{A.27}
$$

□

▸ **STEP 2: Upper bounding the variations of $\hat{P}_t(z)$.** We rely on Weissman's inequality (Weissman et al., 2003), see Lemma 14.

We establish the following result.

Lemma 3. *On the good event $\mathcal{E}(T)$, for all $z \in T$, we have*

$$
\left\|\hat{P}_{t_{k+1}}(z) - \hat{P}_{t_k}(z)\right\|_1 \leq 4\sqrt{S \log(4SAT^3/\delta)}\frac{\sqrt{N_{t_{k+1}}(z) - N_{t_k}(z)}}{N_{t_k}(z)} \tag{A.28}
$$

Proof. Pick $z \in \mathcal{Z}$. For short, denote $n = N_{t_k}(z)$ and $s = N_{t_{k+1}}(z) - N_{t_k}(z)$. Because $k \notin \mathcal{K}_{\mathrm{DT}}$, we know that $s \leq n$. Denote $W_t(z) := N_t(z)\hat{p}_t(z)$ the aggregate empirical distribution of the transition z. Then a straight forward computations shows that

$$
\begin{aligned}
\hat{P}_{t_{k+1}}(z) - \hat{P}_{t_k}(z) &= \frac{1}{n+s}\left(W_{t_k}(z) + \sum\nolimits_{t=t_k}^{t_{k+1}-1}\mathrm{Dirac}(X_{t+1})\mathbf{1}\{Z_t = z\}\right) - \frac{1}{n}W_{t_k}(z) && \text{(A.29)}\\
&= \frac{1}{n+s}\left(-\frac{s}{n}W_{t_k}(z) + \sum\nolimits_{t=t_k}^{t_{k+1}-1}\mathrm{Dirac}(X_{t+1})\mathbf{1}\{Z_t = z\}\right) && \text{(A.30)}
\end{aligned}
$$

Now, on the good event $\mathcal{E}(T)$,

$$
\begin{aligned}
\left\|W_{t_k}(z) - N_{t_k}(z)P(z)\right\|_1 &\leq \sqrt{4nS \log(4SAT^3/\delta)}, \quad \text{and} && \text{(A.31)}\\
\left\|\sum\nolimits_{t=t_k}^{t_{k+1}-1}\mathrm{Dirac}(X_{t+1})\mathbf{1}\{Z_t = z\} - sP(z)\right\|_1 &\leq \sqrt{4sS \log(4SAT^3/\delta)}. && \text{(A.32)}
\end{aligned}
$$

Injecting (A.31) and (A.32) in (A.30), we get that on the good event,

$$\left\|\hat{P}_{t_{k+1}}(z) - \hat{P}_{t_k}(z)\right\|_1 \leq \frac{1}{n+s}\left(\frac{s}{n}\sqrt{4nS\log(4SAT^3/\delta)} + \sqrt{2sS\log(4SAT^3/\delta)}\right) \tag{A.33}$$

$$\leq \sqrt{4S\log(4SAT^3/\delta)}\frac{\sqrt{s}\left(1+\sqrt{s/n}\right)}{n+s} \tag{A.34}$$

$$\leq 4\sqrt{S\log(4SAT^3/\delta)}\frac{\sqrt{s}}{n} \tag{A.35}$$

where (A.36) is obtained by using $s \leq n$. Overall, on $\mathcal{E}(T)$ and for all $z \in \mathcal{Z}$,

$$\left\|\hat{P}_{t_{k+1}}(z) - \hat{P}_{t_k}(z)\right\|_1 \leq 4\sqrt{S\log(4SAT^3/\delta)}\frac{\sqrt{N_{t_{k+1}}(z) - N_{t_k}(z)}}{N_{t_k}(z)} \tag{A.36}$$

□

▸ STEP 3: Upper bounding the variations of $\xi_t(z)$.

Lemma 4 (Step 3). *For all $z \in \mathcal{Z}$,*

$$\left|\xi_{t_{k+1}}(z) - \xi_{t_k}(z)\right| \leq \frac{\sqrt{\kappa_1\log(\kappa_2 T)}\,(N_{t_{k+1}}(z) - N_{t_k}(z))}{\max\{1, N_{t_k}(z)^{3/2}\}} + \frac{(t_{k+1}-t_k)\sqrt{\kappa_1}}{2t_k} \tag{A.37}$$

Proof. Fix $z \in \mathcal{Z}$. Recall that $\xi_t(z) = \sqrt{\kappa_1\log(\kappa_2 t)/N_t(z)}$. We have

$$\xi_{t_{k+1}}(z) - \xi_{t_k}(z) = \sqrt{\kappa_1\log(\kappa_2 t_{k+1})}\left(\frac{1}{\sqrt{N_{t_{k+1}}(z)}} - \frac{1}{\sqrt{N_{t_k}(z)}}\right) \qquad \text{(term ①)}$$

$$+ \sqrt{\frac{\kappa_1}{N_{t_k}(z)}}\left(\sqrt{\log(\kappa_2 t_{k+1}} - \sqrt{\log(\kappa_2 t_k)}\right) \qquad \text{(term ②)}$$

Because $\left|\sqrt{1/(n+s)} - \sqrt{1/n}\right| \leq s/n^{3/2}$, term ① is bounded as

$$\left|\sqrt{\kappa_1\log(\kappa_2 t_{k+1})}\left(\frac{1}{\sqrt{N_{t_{k+1}}(z)}} - \frac{1}{\sqrt{N_{t_k}(z)}}\right)\right| \leq \frac{\sqrt{\kappa_1\log(\kappa_2 T)}\,(N_{t_{k+1}}(z) - N_{t_k}(z))}{N_{t_k}(z)^{3/2}} \tag{A.38}$$

For term ②, let $\phi(t) := \sqrt{\log(\kappa_2 t)}$. Its derivative is $\phi'(t) = \frac{1}{2t}\log^{-1/2}(\kappa_2 t)$ which is decreasing in t. Hence $\phi(t+h) \leq \phi(t) + h\phi'(t) = \phi(t) + \frac{h}{2t}\log^{-1/2}(\kappa_2 t)$. Then we get

$$\left|\sqrt{\frac{\kappa_1}{N_{t_k}(z)}}\left(\sqrt{\log(\kappa_2 t_{k+1}} - \sqrt{\log(\kappa_2 t_k)}\right)\right| \leq \frac{(t_{k+1}-t_k)\sqrt{\kappa_1}}{2t_k\sqrt{N_{t_k}(z)\log(\kappa_2 t_k)}} \leq \frac{(t_{k+1}-t_k)\sqrt{\kappa_1}}{2t_k} \tag{A.39}$$

Putting everything together,

$$\left|\xi_{t_{k+1}}(z) - \xi_{t_k}(z)\right| \leq \frac{\sqrt{\kappa_1\log(\kappa_2 T)}\,(N_{t_{k+1}}(z) - N_{t_k}(z))}{N_{t_k}(z)^{3/2}} + \frac{(t_{k+1}-t_k)\sqrt{\kappa_1}}{2t_k} \tag{A.40}$$

□

▸ STEP 4: New epoch means increase of visit counts. By using the explicit variations of empirical kernels (Lemma 3) and of bonuses (Lemma 4) in (A.13) (see Lemma 2), we see that if $k \in \mathcal{K}_0$, then on the good event $\mathcal{E}(T)$, there must be $z \in \mathcal{Z}$ such that one of the following holds

$$\frac{1}{2^4 D\,\mathrm{sp}(r)}\sqrt{\frac{\alpha\log(T)}{T}} \leq 4\sqrt{S\log(4SAT^3/\delta)}\frac{\sqrt{N_{t_{k+1}}(z) - N_{t_k}(z)}}{\max\{1, N_{t_k}(z)\}}, \qquad \text{or} \qquad \text{(type A)}$$

$$\frac{1}{2^4 D\,\mathrm{sp}(r)}\sqrt{\frac{\alpha\log(T)}{T}} \leq \sqrt{\kappa_1\log(\kappa_2 T)}\frac{N_{t_{k+1}}(z) - N_{t_k}(z)}{\max\{1, N_{t_k}(z)^{3/2}\}}, \qquad \text{or} \qquad \text{(type B)}$$

$$\frac{1}{2^4 D\,\mathrm{sp}(r)}\sqrt{\frac{\alpha\log(T)}{T}} \leq \frac{(t_{k+1}-t_k)\sqrt{\kappa_1}}{2t_k}. \qquad \text{(type C)}$$

There are at most SA episodes such that the z achieving one of the conditions above has never been visited yet, i.e., such that $N_{t_k}(z) = 0$. Such episodes belong to $\mathcal{K} \setminus \mathcal{K}_0$, so can be ignored by assumption. Therefore, we can change $\max\left\{1, N_{t_k}(z)^\lambda\right\}$ to the simpler $N_{t_k}(z)^\lambda$. For episodes of (type A), solving in $N_{t_k}(z)$ yields

$$N_{t_{k+1}}(z) \geq \max\left\{N_{t_k}(z)\left(1 + \frac{\alpha \log(T) \cdot N_{t_k}(z)}{2^{10} D^2 \operatorname{sp}(r)^2 S T \log(4SAT^3/\delta)}\right), 1\right\} \qquad \text{(characterization of type A)}$$

with $N_{t_k}(z) \geq 1$. Similarly, for episodes of (type B), we have

$$N_{t_{k+1}}(z) \geq \max\left\{N_{t_k}(z)\left(1 + \sqrt{\frac{\alpha \log(T) \cdot N_{t_k}(z)}{2^8 D^2 \operatorname{sp}(r)^2 T \kappa_1 \log(\kappa_2 T)}}\right), 1\right\} \qquad \text{(characterization of type B)}$$

with also $N_{t_k}(z) \geq 1$. Finally, for episodes of (type C), we get

$$t_{k+1} \geq t_k \left(1 + \frac{\sqrt{\alpha \log T}}{2^3 D \operatorname{sp}(r) \sqrt{T \kappa_1}}\right). \qquad \text{(characterization of type C)}$$

▸ **STEP 5: Finally counting the number of episodes.** The episodes are partitioned into standard episodes $\mathcal{K}_0$ and non-standard episodes $\mathcal{K}_0 \setminus \mathcal{K}$. The elements of $\mathcal{K}_0$ of (type A), (type B) and (type C) are respectively denoted $\mathcal{K}_{0,\text{A}}, \mathcal{K}_{0,\text{B}}$ and $\mathcal{K}_{0,\text{C}}$. There respective cardinalities are K_A, K_B and K_C. We show that on the good event $\mathcal{E}(T)$, the total number of episodes is bounded above as

$$K \leq \left(\frac{128 D S^{3/2} A \operatorname{sp}(r)}{\sqrt{\alpha}} + 2^4 D \operatorname{sp}(r) \sqrt{\frac{\kappa_1}{\alpha}}\right) \sqrt{T} \log\left(4SAT^3/\delta\right) + O\left(T^{1/3} \log(T)\right) \tag{A.41}$$

K is bounded by $|\mathcal{K} \setminus \mathcal{K}_0| + K_\text{A} + K_\text{B} + K_\text{C}$. The computation are detailed below.

Upper bound of $\mathcal{K} \setminus \mathcal{K}_0$. The number of non-standard episodes is small. If an episode is non-standard, either (a) there is $z \in \mathcal{Z}$ visited on $[t_k, t_{k+1} - 1]$ such that $N_{t_k}(z) = 0$ or (b) there is $z \in \mathcal{Z}$ such that $N_{t_k}(z) \geq 1$ and $N_{t_{k+1}}(z) \geq 2N_{t_k}(z)$. There is some z that accounts for $n \geq \frac{1}{SA}|\mathcal{K} \setminus \mathcal{K}_0|$ episodes of $\mathcal{K} \setminus \mathcal{K}_0$. Let $k_1, k_2, \ldots k_n$ the associated episodes. As $N_{t_{k_{i+1}}}(z) \geq N_{t_{k_i}+1}(z) \geq \max\{1, 2N_{t_{k_i}}(z)\}$ we deduce that $N_{t_{k_{n=1}}} \geq 2^{n-1}$. Since $T \geq N_{t_{k_n}+1}(z)$, we obtain

$$|\mathcal{K} \setminus \mathcal{K}_0| \leq SA(1 + \log_2(T)) = O(SA \log(T)). \tag{A.42}$$

Upper bound of K_A. For episodes of (type A), there is some $z \in \mathcal{Z}$ that accounts for $n \geq \frac{1}{SA} K_\text{A}$ elements of $\mathcal{K}_{0,\text{A}}$. Let $k_1, k_2, \ldots, k_n$ the respective episodes. We have by (characterization of type A)

$$\forall i < n, \quad N_{t_{k_{i+1}}}(z) \geq N_{t_{k_i}+1}(z) \geq N_{t_{k_i}}(z)\left(1 + \frac{\alpha \log(T) \cdot N_{t_{k_i}}(z)}{2^{10} D^2 \operatorname{sp}(r)^2 S T \log(4SAT^3/\delta)}\right) \tag{A.43}$$

with $N_{t_{k_2}}(z) \geq 1$. Setting $u_i := N_{t_{k_{i+1}}}(z)$, we can apply Lemma 13 with λ picked as $\alpha \log(T) 2^{-10} D^{-2} \operatorname{sp}(r)^{-2} S^{-1} \log^{-1}(4SAT^3/\delta)$ and $\omega = 1$ to get, since $u_{n-1} = N_{t_{k_n}}(z) \leq T$, that

$$n - 1 \leq 3 \cdot 2^5 D \operatorname{sp}(r) \sqrt{S T \alpha^{-1} \log(4SAT^3/\delta) \log(T)} \tag{A.44}$$

Using that $n \geq \frac{1}{SA} K_\text{A}$ and solving in K_A, we obtain

$$K_\text{A} \leq SA + 2^7 \cdot \frac{D S^{3/2} A \operatorname{sp}(r)}{\sqrt{\alpha}} \sqrt{T \log(4SAT^3/\delta) \log(T)} \tag{A.45}$$

Upper bound of K_B. The number of episodes of (type B) is bounded similarly. There is some $z \in \mathcal{Z}$ that accounts for $n \geq \frac{1}{SA} K_\text{B}$ elements of $\mathcal{K}_{0,\text{B}}$. Let $k_1, k_2, \ldots, k_n$ the respective episodes. We have by (characterization of type B)

$$\forall i < n, \quad N_{t_{k_{i+1}}}(z) \geq N_{t_{k_i}+1}(z) \geq N_{t_{k_i}}(z)\left(1 + \sqrt{\frac{\alpha \log(T) \cdot N_{t_{k_i}}(z)}{2^8 D^2 \operatorname{sp}(r)^2 T \kappa_1 \log(\kappa_2 T)}}\right), \tag{A.46}$$

with $N_{t_{k_2}}(z) \geq 1$. Setting $u_i := N_{t_{k_{i+1}}}(z)$, we can apply Lemma 13 with λ picked as $\alpha \log(T) 2^{-8} D^{-2} \mathrm{sp}(r)^{-2} \kappa_1^{-1} \log^{-1}(\kappa_2 T)$ and $\omega = \frac{1}{2}$ to get, since $u_{n-1} = N_{t_{k_n}}(z) \leq T$, that

$$n - 1 \leq 3 \cdot 2^{4/3} D^{2/3}\ \mathrm{sp}(r)^{2/3} \left(T \kappa_1 \alpha^{-1} \log(\kappa_2 T)\right)^{1/3} \log^{2/3}(T) \tag{A.47}$$

Using that $n \geq \frac{1}{SA} K_{\mathrm{B}}$ and solving in K_{B}, we obtain

$$K_{\mathrm{B}} \leq SA + 2^3 \cdot \frac{D^{2/3} \kappa^{1/3} SA\ \mathrm{sp}(r)}{\alpha^{1/3}} T^{1/3} \log^{1/3}(\kappa_2 T) \log^{2/3}(T) = O\left(T^{1/3} \log(T)\right) \tag{A.48}$$

Upper bound of K_{C}. Denote $n = K_{\mathrm{C}}$ and introduce $k_1, k_2, \ldots, k_n$ the elements of $\mathcal{K}_{0,\mathrm{C}}$. By (characterization of type C), we have

$$t_{k_{i+1}} \geq t_{k_i + 1} \geq t_{k_i} \left(1 + \frac{\sqrt{\alpha \log T}}{2^3 D\ \mathrm{sp}(r) \sqrt{T \kappa_1}}\right). \tag{A.49}$$

By induction on i, we show

$$t_{k_i} \geq \left(1 + \frac{\sqrt{\alpha \log T}}{2^3 D\ \mathrm{sp}(r) \sqrt{T \kappa_1}}\right)^{i-1} \tag{A.50}$$

Since $t_{k_n} \leq T$, we deduce that

$$(n-1) \log \left(1 + \frac{\sqrt{\alpha \log T}}{2^3 D\ \mathrm{sp}(r) \sqrt{T \kappa_1}}\right) \leq \log(T). \tag{A.51}$$

Unfold $n = K_{\mathrm{C}}$ and solve in K_{C}. Also rely on $\log(1 + x) \geq \frac{1}{2} x$ which holds for $x \leq 1$. We finally obtain, for $T \geq \frac{\alpha \log T}{2^6 D^2 \mathrm{sp}(r) \kappa_1}$,

$$K_{\mathrm{C}} \leq 1 + 2^4 D\ \mathrm{sp}(r) \sqrt{\frac{\kappa_1}{\alpha}} \sqrt{T} \log(T) \tag{A.52}$$

A.3. Analysis of the Regret of UCRL-PT

The decomposition of the regret is standard (see (Auer et al., 2009)). The end points of episode k are denoted $[t_k, t_{k+1} - 1]$. Let K the (random) number of episodes up to T, i.e., the maximal k such that $t_k \leq T$ and t_{K+1} is truncated to T. The regret is first decomposed into episodes:

$$\mathrm{Reg}(T) = \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \left[g^* - r(Z_t)\right]. \tag{A.53}$$

Then expanded as:

$$\mathrm{Reg}(T) = \sum_{k=1}^{K} \left[\underbrace{\left(\sum_{t=t_k}^{t_{k+1}-1} \left(g^*_M - \tilde{g}^*_{t_k}\right) \right)}_{\textcircled{1_k}} + \underbrace{\left(\sum_{t=t_k}^{t_{k+1}-1} \left(\tilde{g}^*_{t_k} - \tilde{g}^{\pi^k}_{t_k}\right) \right)}_{\textcircled{2_k}} + \underbrace{\left(\sum_{t=t_k}^{t_{k+1}-1} \left(\tilde{g}^{\pi^k}_{t_k} - r(Z_t)\right) \right)}_{\textcircled{3_k}} \right] \tag{A.54}$$

Upper bounding term $\textcircled{1_k}$. On the good event $\mathcal{E}(T)$, $M \in \widetilde{\mathcal{M}}_{t_k}$. It follows that $g^*(M) \leq \sup_{\tilde{M} \in \widetilde{\mathcal{M}}_t} g^*(\tilde{M}) = \tilde{g}^*_{t_k}$. Therefore, on $\mathcal{E}(T)$,

$$\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \left(g^*(M) - \tilde{g}^*_{t_k}\right) \leq 0. \tag{A.55}$$

Upper bounding term $\textcircled{2_k}$. Simply by definition of π^k, $\tilde{g}^*_{t_k} = \tilde{g}_{t_k}(\pi^k)$ and $\textcircled{2_k} = 0$.

A.4. Main Component of the Regret

The term $\textcircled{3_k}$ is also rather standard. Recall that if π is a policy, g^π its gain vector, r^π its reward vector, h^π its bias vector and P^π its transition matrix, with have the vectorial identity

$$g^\pi - r^\pi = (P^\pi - I)h^\pi \tag{A.56}$$

where I is the identity matrix. Let $\tilde{M}_t^\pi$ the optimistic model (at time t) associated to $\pi \in \Pi$ and denote $\tilde{P}_t^\pi$ the associated transition matrix. Applying (A.56) to π^k under $\tilde{M}_{t_k}^{\pi^k}$, we have

$$\begin{aligned}
\text{‘}\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\left(\tilde{g}_{t_k}(\pi^k) - r(Z_t)\right) = \sum_{k=1}^{K}\Bigg[& \underbrace{\sum_{t=t_k}^{t_{k+1}-1}\left(\left(\tilde{P}_{t_k}^{\pi^k} - P^{\pi^k}\right)\tilde{h}_{t_k}^{\pi^k}\right)(X_t)}_{\textcircled{4_k}} && \text{(A.57)}\\
&+ \underbrace{\sum_{t=t_k}^{t_{k+1}-1}\left(\left(P^{\pi^k}\tilde{h}_{t_k}^{\pi^k}\right)(X_t) - \tilde{h}_{t_k}^{\pi^k}(X_{t+1})\right)}_{\textcircled{5_k}} && \text{(A.58)}\\
&+ \underbrace{\sum_{t=t_k}^{t_{k+1}-1}\left(\tilde{h}_{t_k}^{\pi^k}(X_{t+1}) - \tilde{h}_{t_k}^{\pi^k}(X_t)\right)}_{\textcircled{6_k}}\Bigg] && \text{(A.59)}
\end{aligned}$$

In (A.57), the term $\textcircled{4_k}$ is bounded on the good event by the width of the confidence intervals, the term $\textcircled{5_k}$ is a well-known martingale difference sequence and the telescopic $\textcircled{6_k}$ is proportional to the number of episodes. An important point to note here is a remarkable property of EVI, stating that on the good event $\mathcal{E}(T)$, $\mathrm{sp}(\tilde{h}_{t_k}^{\pi^k}) \le \mathrm{sp}(r)D$, see (Auer et al., 2009, Section 4.3.1).

Bound of the term $\textcircled{4_k}$. Check that for all $x \in \mathcal{S}$, on the good event,

$$\begin{aligned}
\left(\left(\tilde{P}_{t_k}^{\pi^k} - P^{\pi^k}\right)\tilde{h}_{t_k}^{\pi^k}\right)(x) &\le \left\|\tilde{P}_{t_k}^{\pi^k}(x) - P^{\pi^k}(x)\right\|_1 \mathrm{sp}(\tilde{h}_{t_k}^{\pi^k})\\
&\le \xi_{t_k}(x, \pi^k(x))D\,\mathrm{sp}(r)
\end{aligned}$$

where the first inequality is Hölder's and the second is by properties of the good event. Instantiating to $x = X_t$ and summing, we get, on the good event

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\left(\left(\tilde{P}_{t_k}^{\pi^k} - P^{\pi^k}\right)\tilde{h}_{t_k}^{\pi^k}\right)(X_t) &\le D\,\mathrm{sp}(r)\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{\kappa_1\log(\kappa_2 t_k)}{\max\{1, N_{t_k}(Z_t)\}}} && \text{(A.60)}\\
&= D\,\mathrm{sp}(r)\sum_{k=1}^{K}\sum_{z}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{\kappa_1\log(\kappa_2 T)}{\max\{1, N_{t_k}(z)\}}}\mathbf{1}\{Z_t = z\} && \text{(A.61)}\\
&\le D\,\mathrm{sp}(r)\sqrt{2\kappa_1\log(\kappa_2 T)}\sum_{k=1}^{K}\sum_{z}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{\mathbf{1}\{Z_t = z\}}{\max\{1, N_t(Z_t)\}}} && \text{(A.62)}\\
&\le D\,\mathrm{sp}(r)\sqrt{2\kappa_1\log(\kappa_2 T)}\sum_{z}\sum_{n=0}^{N_T(z)}\sqrt{\frac{1}{\max\{1, n\}}} && \text{(A.63)}
\end{aligned}$$

where (A.60) is because we are on the good event $\mathcal{E}(T)$, (A.62) follows from the doubling trick and (A.63) is just rewriting. Now, for $U \ge 1$, a straight-forward series-integral comparison establishes $\sum_{u=1}^{U}\sqrt{1/u} \le 2\sqrt{U} - 1$, so for $U \ge 0$, $\sum_{u=1}^{U}\sqrt{1/u} \le 2\sqrt{U}$. Hence, for $U \ge 0$,

$$\sum_{u=0}^{U}\sqrt{\frac{1}{\max\{1, u\}}} \le 2\sqrt{U} + 1.$$

Using that in (A.63), we get that, on the good event,

$$\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\left(\left(\tilde{P}_{t_k}^{\pi^k} - P^{\pi^k}\right)\tilde{h}_{t_k}^{\pi^k}\right)(X_t) \leq D\ \mathrm{sp}(r)\sqrt{2\kappa_1\log(\kappa_2 T)}\left(SA + 2\sum_{z}\sqrt{N_T(z)}\right) \tag{A.64}$$

$$\begin{aligned} &\leq DSA\ \mathrm{sp}(r)\sqrt{2\kappa_1\log(\kappa_2 T)} \\ &+ 2D\ \mathrm{sp}(r)\sqrt{2SA\kappa_1 T\log(\kappa_2 T)}. \end{aligned} \tag{A.65}$$

Here, (A.65) follows from $\sum_z \sqrt{N_T(z)} \leq \sqrt{SAT}$ which is a consequence of Cauchy-Schwartz's inequality.

Bound of the term $\textcircled{5_k}$. First, observe that on the good event $\mathcal{E}(T)$,

$$\left(\tilde{g}_{t_k}^* - \tilde{g}_{t_k}(\pi^k)\right) = \left(\tilde{g}_{t_k}^* - \tilde{g}_{t_k}(\pi^k)\right)\mathbf{1}\left\{\mathcal{E}(t_k)\right\} \tag{A.66}$$

We obtain a standard martingale difference sequence (MDS). Denote $\{\mathcal{F}_t\}$ the filtration induced by the history of play. Then, for $t \geq t_k$, the Markov property of system guarantees that

$$\mathbb{E}\left[\tilde{h}_{t_k}^{\pi^k}(X_{t+1})\mathbf{1}\left\{\mathcal{E}(t_k)\right\} \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}\left[\sum_y \mathbf{1}\left\{\mathcal{E}(t_k)\right\}\tilde{h}_{t_k}^{\pi^k}(y)\mathbf{1}\left\{X_{t+1} = y\right\} \,\middle|\, \mathcal{F}_t\right] \tag{A.67}$$

$$= \mathbf{1}\left\{\mathcal{E}(t_k)\right\}\sum_y \tilde{h}_{t_k}^{\pi^k}(y)\mathbb{E}\left[\mathbf{1}\left\{X_{t+1} = y\right\} \mid \mathcal{F}_t\right] \tag{A.68}$$

$$= \mathbf{1}\left\{\mathcal{E}(t_k)\right\}\sum_y P^{\pi^k}(y|X_t)\tilde{h}_{t_k}^{\pi^k}(y) = \mathbf{1}\left\{\mathcal{E}(t_k)\right\}\left(P^{\pi^k}\tilde{h}_{t_k}^{\pi^k}\right)(X_t). \tag{A.69}$$

Above, (A.68) is obtained by $\mathcal{F}_{t_k}$-measurability of $\mathbf{1}\left\{\mathcal{E}(t_k)\right\}$ and $\tilde{h}_{t_k}^{\pi^k}$. Now, the MDS has terms almost surely bounded by

$$\mathrm{sp}\left(\mathbf{1}\left\{\mathcal{E}(t_k)\right\}\tilde{h}_{t_k}^{\pi^k}\right) \leq D\ \mathrm{sp}(r). \tag{A.70}$$

By Azuma-Hoeffding's inequality Lemma 15, we have

$$\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}\left\{\mathcal{E}(t_k)\right\}\left(\left(P^{\pi^k}\tilde{h}_{t_k}^{\pi^k}\right)(X_t) - \tilde{h}_{t_k}^{\pi^k}(X_{t+1})\right) \leq D\ \mathrm{sp}(r)\sqrt{2T\log(1/\delta')} \tag{A.71}$$

with probability $1-\delta'$. Therefore, for all $\delta' > 0$, there is an event $\mathcal{E}_{\delta'}$ of probability at least $1-\delta'$ such that, on $\mathcal{E}(T)\cap\mathcal{E}_{\delta'}$, we have

$$\sum_{k=1}^{K'}\sum_{t=t'_k}^{t'_{k+1}-1}\left(\left(P^{\pi^k}\tilde{h}_{t_k}^{\pi^k}\right)(X_t) - \tilde{h}_{t_k}^{\pi^k}(X_{t+1})\right) \leq D\ \mathrm{sp}(r)\sqrt{2T\log(1/\delta')} \tag{A.72}$$

Bound of the term $\textcircled{6_k}$. We have, on the good event,

$$\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\left(\tilde{h}_{t_k}^{\pi^k}(X_{t+1}) - \tilde{h}_{t_k}^{\pi^k}(X_t)\right) = \sum_{k=1}^{K}\left(\tilde{h}_{t_k}^{\pi^k}(X_{t_{k+1}}) - \tilde{h}_{t_k}^{\pi^k}(X_{t_k})\right) \tag{A.73}$$

$$\leq \sum_{k=1}^{K}\mathrm{sp}\left(\tilde{h}_{t_k}^{\pi^k}\right) \tag{A.74}$$

$$\leq KD\ \mathrm{sp}(r) \tag{A.75}$$

where (A.75) is because on the good event $\mathcal{E}(T)$, $\mathrm{sp}(\tilde{h}_{t'_k}^{\pi^k}) \leq D\ \mathrm{sp}(r)$. We finally ready to put everything together. Under the good event, keeping only the terms in $\sqrt{T}$,

$$\mathrm{Reg}(T) \leq 2D\ \mathrm{sp}(r)\sqrt{2SA\kappa_1 T\log(\kappa_2 T)} + D\ \mathrm{sp}(r)\sqrt{2T\log(1/\delta')} + K'D\ \mathrm{sp}(r) + o(\sqrt{T}).$$

This finishes the proof of Theorem 3.

B. The Regret of Exploration

B.1. The Regret of Exploration of Classical Algorithms: Proof of Theorem 1

Proof. Let M be any ergodic MDP with at least one sub-optimal state-action pair. A sub-optimal state-action x, a is such that $Q^*(x, a) < g^*$. Because the algorithm $\mathcal{L}$ is consistent, it must use all sub-optimal state-actions infinitely often (see for example (Burnetas & Katehakis, 1997) that shows that each sub-optimal state-action is visited at least $K \log(T)$ times in expectation, where $K > 0$). Since M is ergodic, if a sub-optimal state-action is used at time t, then the current policy π_t is also sub-optimal. Therefore, $\mathcal{L}$ samples from sub-optimal policies for infinitely many episodes, so the number of exploration episodes is infinite. Let $\Delta_g(M) := \min\{g^*(M) - g(\pi; M) : \pi \in \Pi \text{ s.t. } g(\pi; M) < g^*(M)\}$ the gain-gap of M and let $H := \max_\pi \mathrm{sp}(h(\pi; M))$ the worst bias span.

Recall that $k(n)$ denotes the n-th exploration episode. Fix $T \geq 1$ a horizon. We know by assumption on $\mathcal{L}$ that $\mathbb{P}\{t_{k(n)+1} - t_{k(n)} \to \infty\} = 1$, so there is a finite $n(T)$ such that

$$\mathbb{P}\{\forall m \geq n,\ t_{k(m)+1} - t_{k(m)} \geq T\} \geq 1 - \frac{1}{1+T}.$$

For short, denote $\mathcal{E}(m)$ the event above. Since $t_{k(m)}$ is a stopping time w.r.t. the natural filtration, we have

$$\begin{aligned}\mathbb{E}\left[\mathrm{SReg}(t_{k(m)}; t_{k(m)} + T)\right] &\geq \mathbb{E}\left[\sum\nolimits_{t=t_{k(m)}}^{t_{k(m)}+T-1} (g^*(M) - g(\pi_t; M) - \mathrm{sp}(h(\pi_t; M)\mathbf{1}\{\pi_t \neq \pi_{t-1}\}))\right] \\ &\geq \Delta_g \cdot \mathbb{E}\left[\min\{T, t_{k(m)+1} - t_{k(m)}\}\right] - H \cdot \mathbb{E}\left[1 + \max\{0, t_{k(m)} + T - t_{k(m)+1}\}\right].\end{aligned}$$

By choice of $n(T)$, it is direct to check that we obtain that for all $m \geq n(T)$,

$$\mathbb{E}\left[\mathrm{SReg}(t_{k(m)}; t_{k(m)} + T)\right] \geq \frac{T^2}{1+T}\Delta_g - 2H.$$

So, taking the limsup in m, we get $\mathrm{RegExp}(\mathcal{L}, M; T) \geq \frac{T^2}{1+T}\Delta_g - 2H = \Omega(T)$. □

B.2. The Regret of Exploration of UCRL2-PT: a Proof

Convention on the use of constants. In the following, unless it is specifically precised that a constant is *numerical*, all constants $C_{\text{something}}$ may depend on the DMDP; that is, on $S, A, g^*(M), h^*(M)$ etc.

Along the proof, if the proof of a lemma doesn't immediately follow its statement, it is proved separately further down.

The proof begins with the introduction of a good event which is different to the one used in the regret analysis. Because the regret of exploration is a statement on the regret truncated to random time-windows $[t_{k(n)}, t_{k(n)} + T]$, we need time-uniform confidence intervals. Although, when one requires regret guarantees in expectation, confidence bounds must be of the form $\sqrt{\log(t)/N_t(z)}$, the regret of exploration $\mathrm{RegExp}(T)$ accounts for time-windows of lengths T, hence we only need the good event to hold with probability $1 - \frac{1}{T}$. This means that we will be able to say that, with high probability, the confidence intervals that the algorithm relies on are much larger the practical deviations between empirical and true mean rewards.[4] We rely on $\log(\log(t)/\delta)$ confidence intervals, see (Kaufmann & Koolen, 2021) for e.g. Let $c, d > 0$ given by Lemma 16. That is, for all $\delta > 0$ and all $z \in \mathcal{Z}$, we have

$$\mathbb{P}\left\{\exists t \geq 1 : N_t(z)\left(\hat{r}_t(z) - r_t(z)\right)^2 > c\left(\log(d + \log(t)) + \log(1/\delta)\right)\right\} \leq \delta.$$

Consider the family of good events:

$$\mathcal{E}(T) := \left\{\forall z, \forall t \geq 1 : N_t(z)\left(\hat{r}_t(z) - r_t(z)\right)^2 > c\left(\log(d + \log(t)) + \log(SAT)\right).\right\} \tag{B.1}$$

Then it appears clearly that $\mathbb{P}(\mathcal{E}(T)) \geq 1 - \frac{1}{T}$. On the good event, we can control the visit counts: transitions $z \in \mathrm{supp}(\mu_{\pi^*})$ are sampled linearly often while, for $z \notin \mathrm{supp}(\mu_{\pi^*})$, $N_t(z) = \Theta(\log(t))$.

[4] Remember that to have regret guarantees in expectation, we set the confidence threshold of the algorithm as the time-dependent $\delta(t) = \frac{1}{t}$. So, morally, the confidence intervals of the algorithm are designed to hold with probability $1 - \frac{1}{t} \to 1$, while the required $1 - \frac{1}{T}$ is constant.

Lemma 5 (Visit counts)**.** *There is a time-threshold $t_{\text{visits}} : T \mapsto t_{\text{visits}}(T)$ and constants $C^*_{\text{visits}}, C^-_{\text{visits}}, C^+_{\text{visits}} > 0$ such that, provided that $t \geq t_{\text{visits}}(T)$, on the good event $\mathcal{E}(T)$,*

(1) $N_t(z) \geq C^*_{\text{visits}} t$ for $z \in \operatorname{supp}(\mu_{\pi^*})$;
(2) $N_t(z) \geq C^-_{\text{visits}} \log(t)$ for $z \notin \operatorname{supp}(\mu_{\pi^*})$; and
(3) $N_t(z) \leq C^+_{\text{visits}} \log(t)$ for $z \notin \operatorname{supp}(\mu_{\pi^*})$.

Lemma 5 is proved later in this paper. According to Lemma 5, all transitions are visited logarithmically often, hence optimistic quantity change at most at rate $O(\frac{1}{\log(t)})$. Following this observation, we show that at the beginning of an exploration episode $k(n)$, the optimistic value of the optimal policy π^* is close to $\tilde{g}^*_{t_{k(n)}}$ up to a factor of $\frac{1}{\log t}$ on the good event. That is, no policy have an optimistic value which is significantly larger than the optimistic value of π^*.

Lemma 6 (Good initialization of exploration episodes)**.** *There is a time-threshold $t_{\text{init}} : T \mapsto t_{\text{init}}(T)$ such that for all $T \geq 1$, for all exploration episode $k(n) \geq t_{\text{init}}(T)$ (i.e., $\pi^{k-1} = \pi^*$ and $\pi^k \neq \pi^*$) and on the good event $\mathcal{E}(T)$,*

$$\tilde{g}_{t_{k(n)}}(\pi^*) + \frac{C_{\text{init}}}{\log(t_{k(n)})} \geq \tilde{g}^*_{t_{k(n)}}.$$

Proof. We know that for some $t \geq t_{k(n)} - S$, $\tilde{g}_t(\pi^*) + \sqrt{\alpha \log(t)/t} \geq \tilde{g}^*_t$, hence we only have to control by how much the optimistic value $\tilde{g}_t(\pi^*)$ is subjected to change over S steps. We can further assume that transitions are visited at most once over $[t, t_k]$. Given that $t_{k(n)} \geq t_{\text{visits}}(T) + S$ (e.g., $k(n) \geq t_{\text{visits}}(T) + S$) and on the good event $\mathcal{E}(T)$, we have:

$$N_t(z) \geq C^-_{\text{visits}} \log(t)$$

for all z. We deduce that for all z,

$$\left|\hat{r}_t(z) - \hat{r}_{t_k}(z)\right| \leq \frac{2(N_{t_k}(z) - N_t(z))R_{\max}}{N_t(z)} \leq \frac{2R_{\max}}{C^-_{\text{visits}} \log(t)},$$

Similarly, the change in optimistic value $\left|\xi_t(z) - \xi_{t_k}(z)\right|$ can be shown to be at most $\frac{C}{\log(t)}$ for some constant $C > 0$. Accordingly, there exists a constant $C_{\text{init}} > 0$ such that on the good event $\mathcal{E}(T)$, we have

$$\forall \pi, \quad \left|\tilde{g}_t(\pi) - \tilde{g}_{t_k}(\pi)\right| \leq \frac{C_{\text{init}}}{3\log(t)}.$$

If in addition, we have $\sqrt{\alpha \log(t)/t} \leq \frac{C_{\text{init}}}{3\log(t)}$, we conclude that $\tilde{g}_t(\pi^*) + \frac{C_{\text{init}}}{\log(t)} \leq \tilde{g}^*_t$ by triangular inequality. This holds for t large w.r.t. T on the good event $\mathcal{E}(T)$. □

To some extent, Lemma 6 states that exploration episodes are "well-initialized". By definition of the (**PT**)-rule, UCRL2-PT only picks policies that are nearly-optimistically optimal. Therefore, if the difference of optimistic values $\tilde{\Delta}_t(\pi^*; \pi) := \tilde{g}_t(\pi^*) - \tilde{g}(\pi)$ increases quickly enough when sampling from the suboptimal policy π, one should expect UCRL2-PT to switch back quickly to π^* hence have small regret of exploration. This is the main line of the proof: On the time-window $[t_{k(n)}, t_{k(n)} + T]$, there is a threshold on the visits of transitions $z \notin \operatorname{supp}(\mu_{\pi^*})$ beyond which a policy π such that $z \in \operatorname{supp}(\mu_\pi)$ won't be used by UCRL2-PT. This is formally described by Lemma 8, whose proof relies on the technical Lemma 7.

Below, $\mathcal{D}_h$ is the h-step difference operator on functions $\mathbb{N} \to \mathbb{R}$, i.e., for $f : \mathbb{N} \to \mathbb{R}$, $\mathcal{D}_h f$ is the function given by

$$(\mathcal{D}_h f)(t) := f(t+h) - f(t). \tag{B.2}$$

Lemma 7. *There exist constants $C_{\text{noise}}, C_{\text{drift}}, C_{\pi^*} > 0$ and a non-decreasing function $t_0 : T \mapsto t_0(T)$ such that, on $\mathcal{E}(T)$, for all $T \geq 1$ and all $t \geq t_0(T)$, we have*

$$\begin{aligned}\forall \pi \neq \pi^*, \forall h \leq T, \quad \mathcal{D}_h\left[\tilde{\Delta}_t(\pi^*; \pi)\right](t) \geq &+ C_{\text{drift}} \sum_{\substack{z \in \operatorname{supp}(\mu_\pi) \\ z \notin \operatorname{supp}(\mu_{\pi^*})}} \frac{N_{t+h}(z) - N_t(z)}{N_t(z)} \\ &- C_{\text{noise}} \sum_{\substack{z \in \operatorname{supp}(\mu_\pi) \\ z \notin \operatorname{supp}(\mu_{\pi^*})}} \frac{\sqrt{(N_{t+h}(z) - N_t(z)) \log(T)}}{N_t(z)} \\ &- C_{\pi^*} \cdot \frac{h}{\sqrt{t}},\end{aligned}$$

provided that $\forall z, \forall h \leq T$, $\left|\sum_{i=0}^{h} (R_{t+i} - r(z)) \mathbf{1}\{Z_{t+i} = z\}\right| \leq \sqrt{\frac{1}{2}(N_{t+h}(z) - N_t(z)) \log(SAT^3)}$.

Refer to the dedicated section for a proof of Lemma 7. In the lower bound of $\mathcal{D}_h[\tilde{\Delta}_t(\pi^*;\pi)](t)$, the dominant terms are a *drift* term and a *noise* term. The drift term is obtained by the decrease of the confidence bonus $\xi_t(z)$ associated to $z \in \operatorname{supp}(\mu_\pi) \setminus \operatorname{supp}(\mu_{\pi^*})$ that has to compensate for a noise term that originates from the rewards gathered within the time range $[t, t+h]$. Interestingly, for small time range h, the drift term may be small in front of its noisy counterpart. When $N_{t+h}(z) - N_t(z)$ grows large however, the noise eventually becomes small in front of the drift and $\Delta_t(\pi^*;\pi)$ overall increases over time. This is be observed experimentally, see Figure 4.

From this lower bound, we can derive the following more practical result.

Lemma 8. *There is a time-threshold* $t_{\exp} : T \mapsto t_{\exp}(T)$ *and a constant* $C_{\exp} > 0$ *satisfying the following property: For all* $T \geq 1$ *and all* $n \geq t_{\exp}(T)$*, there exists* $\mathcal{E}_n(T)$ *a probability* $1 - \frac{2}{T}$ *event on which, for all* $h \in [0, T]$*, if* $\pi_{t_{k(n)}+h} \neq \pi^*$*,*

$$\exists z \in \operatorname{supp}(\mu_\pi) \setminus \operatorname{supp}(\mu_{\pi^*}), \quad N_{t_{k(n)}+h}(z) - N_{t_{k(n)}}(z) \leq C_{\exp}(1 + \log(T)).$$

Proof. To start off, remark that Lemma 7 ask for the condition on collected rewards:

$$\forall z, \forall h \leq T, \quad \left| \sum\nolimits_{i=0}^{h} (R_{t_{k(n)}+i} - r(z)) \mathbf{1}\left\{Z_{t_{k(n)}+i} = z\right\} \right| \leq \sqrt{\frac{1}{2}(N_{t+h}(z) - N_t(z)) \log(SAT^3)}. \tag{B.3}$$

If $t_{k(n)}$ where to be changed to a constant t, the above would hold with probability $1 - \frac{1}{T}$ following a standard union bound on z, on $h \leq T$, and on the possible values for $N_{t+h}(z) - N_t(z)$, then applying Azuma-Hoeffding's inequality. But because $t_{k(n)}$ is a stopping time and (B.3) only depends on what happens *after* $t_{k(n)}$, the Markov property guarantees that (B.3) holds with probability $1 - \frac{1}{T}$, even though $t_{k(n)}$ is random.

Now, we set

$$\mathcal{E}_n(T) := \mathcal{E}(T) \cap \text{(B.3)} \tag{B.4}$$

which is of probability at least $1 - \frac{2}{T}$.

We are searching for a sufficient condition under which $(*)$ $\mathcal{D}_h[\tilde{\Delta}_t(\pi^*;\pi)](t) \geq \frac{2C_{\text{init}}}{\log(t)}$ for a given $\pi \neq \pi^*$. Fix $T \geq 1$ and restrict the attention to the good event $\mathcal{E}_n(T)$. Pick $t \geq t_0(T)$ and denote $\mathcal{X}_\pi := \operatorname{supp}(\mu_\pi) \setminus \operatorname{supp}(\mu_{\pi^*})$ for short. By Lemma 7, to achieve $(*)$, it is enough to have

$$C_{\text{drift}} \sum_{z \in \mathcal{X}_\pi} \frac{N_{t+h}(z) - N_t(z)}{N_t(z)} \geq C_{\text{noise}} \sum_{z \in \mathcal{X}_\pi} \frac{\sqrt{(N_{t+h}(z) - N_t(z)) \log(T)}}{N_t(z)} + \frac{2C_{\text{init}}}{\log(t)} + \frac{C_{\pi^*} h}{\sqrt{t}} \tag{B.5}$$

By taking t large in front of $h \leq T$, we have $C_{\pi^*} h / \sqrt{t} \leq C_{\text{init}} / \log(t)$, hence we can simplify the condition to the sufficient one:

$$C_{\text{drift}} \sum_{z \in \mathcal{X}_\pi} \frac{N_{t+h}(z) - N_t(z)}{N_t(z)} \geq C_{\text{noise}} \sum_{z \in \mathcal{X}_\pi} \frac{\sqrt{(N_{t+h}(z) - N_t(z)) \log(T)}}{N_t(z)} + \frac{3C_{\text{init}}}{\log(t)} \tag{B.6}$$

Now, we replace the visit counts by the estimates given by Lemma 5. We derive the sufficient condition:

$$C_{\text{drift}} \sum_{z \in \mathcal{X}_\pi} \frac{N_{t+h}(z) - N_t(z)}{C^+_{\text{visits}} \log(t)} \geq C_{\text{noise}} \sum_{z \in \mathcal{X}_\pi} \frac{\sqrt{(N_{t+h}(z) - N_t(z)) \log(T)}}{C^-_{\text{visits}} \log(t)} + \frac{3C_{\text{init}}}{\log(t)} \tag{B.7}$$

By Cauchy-Schwartz's inequality, we have

$$\sum_{z \in \mathcal{X}_\pi} \sqrt{N_{t+h}(z) - N_t(z)} \leq \sqrt{S \sum_{z \in \mathcal{X}_\pi} (N_{t+h}(z) - N_t(z))}, \tag{B.8}$$

Therefore, rearranging terms, a sufficient condition to $(*)$ is

$$\sum_{z \in \mathcal{X}_\pi} (N_{t+h}(z) - N_t(z)) \geq \frac{C_{\text{noise}} C^+_{\text{visits}} \sqrt{S} \log(T)}{C_{\text{drift}} C^-_{\text{visits}}} \sqrt{\sum_{z \in \mathcal{X}_\pi} (N_{t+h}(z) - N_t(z))} + \frac{3C_{\text{init}} C^+_{\text{visits}}}{C_{\text{drift}}} \tag{B.9}$$

Again, it is enough for $\sum_{z\in\mathcal{X}_\pi}(N_{t+h}(z) - N_t(z))$ to be greater than twice of each of the right-hand terms. Simple algebra leads to the sufficient condition:

$$\sum_{z\in\mathcal{X}_\pi}(N_{t+h}(z) - N_t(z)) \geq \left(\frac{2C_{\text{noise}}C^{+}_{\text{visits}}}{C_{\text{drift}}C^{-}_{\text{visits}}}\right)^2 S\,\log(T) + \left(\frac{6C_{\text{init}}C^{+}_{\text{visits}}}{C_{\text{drift}}}\right)^2. \tag{B.10}$$

Taking the worst of the two constants above, we get a condition of the form:

$$\sum_{z\in\mathcal{X}_\pi}(N_{t+h}(z) - N_t(z)) \geq C'_{\text{exp}}\,(1 + \log(T)). \tag{B.11}$$

To conclude the proof, we couple this sufficient condition with the good-initialization property of Lemma 6. Choose $t = t_{k(n)}$ the start-time of an exploration episode with n large enough. On the good event $\mathcal{E}(T)$, we know that $\tilde{\Delta}_{t_{k(n)}}(\pi^*;\pi) \geq -\frac{C_{\text{init}}}{\log(t_{k(n)})}$. So if addition (B.11) holds, then

$$\tilde{\Delta}_{t_{k(n)}+h}(\pi^*;\pi) \geq \frac{C_{\text{init}}}{\log(t_{k(n)})} \tag{B.12}$$

which is larger than $\sqrt{\alpha\log(t_{k(n)})/t_{k(n)}}$ if $t_{k(n)}$ is large enough. Therefore, we deduce that on the good event $\mathcal{E}_n(T)$, for $h \in [0,T]$ with $t_{k(n)}$ large enough, if $\pi_{t_{k(n)}+h} \neq \pi^*$ then, there must be a $h - S \leq h' \leq h$ such that $\tilde{\Delta}_{t_{k(n)}+h'}(\pi^*;\pi) \leq \sqrt{\alpha\log(t_{k(n)})/t_{k(n)}} < C_{\text{init}}/\log(t_{k(n)})$, hence (B.11) cannot hold, i.e.,

$$\sum_{z\in\mathcal{X}_\pi}\left(N_{t_{k(n)}+h}(z) - N_{t_{k(n)}}(z)\right) \leq S + C'_{\text{exp}}(1 + \log(T)). \tag{B.13}$$

Setting $C_{\text{exp}} = S + C'_{\text{exp}}$, we get that, there must exists $z \in \text{supp}(\mu_\pi)$ such that

$$N_{t+h}(z) - N_t(z) \leq C_{\text{exp}}(1 + \log(T)). \tag{B.14}$$

This proves the claim. □

This lemma states that, starting from an exploration episode $k(n)$, iterating from a suboptimal policy necessarily implies that one of its recurrent transition has been limitedly sampled since $t_{k(n)}$. This result is key to bound the number of times suboptimal policies can be sampled from in the time-window $[t_{k(n)}, t_{k(n)} + T]$. The issue is that iterating a policy doesn't implies that we immediately sample from the sub-sampled transition z in the above lemma. To address that, we link the number of times a transition belongs to the optimal cycle of the exploration policy to its actual visit count:

Lemma 9 (Strong Laziness)**.** *Let z a transition and denote $m_h(z)$ the number of times over $[t_{k(n)}, t_{k(n)} + h]$ when z is a recurrent transition of the exploration policy, i.e., $m_h(z)$ is the cardinal of $\{i \in [t_{k(n)}, t_{k(n)} + h] : z \in \text{supp}(\mu_{\pi_i})\}$. Then*

$$N_{t_{k(n)}+h}(z) - N_{t_{k(n)}}(z) \geq \frac{m_h(z)}{2S} - 1. \tag{B.15}$$

We can finally establish the result on the regret of exploration, starting by bounding the number of times the exploration policy can be suboptimal over $[t_{k(n)}, t_{k(n)} + T]$. Applying Lemma 8 and Lemma 10 in tandem, we see that for $h \in [0,T]$, if $\pi_{t_{k(n)}+h} \neq \pi^*$, then there is some $z \in \text{supp}(\mu_\pi)$ such that

$$m_h(z) \leq 2S(1 + C_{\text{exp}}\log(T)). \tag{B.16}$$

Now, introduce the *variant* quantity:

$$\varphi(h) := \sum_{z\notin\text{supp}(\mu_{\pi^*})} \left|1 + m_h(z) - 2S(2 + S + C\log(T)\right|_+. \tag{B.17}$$

On the good event $\mathcal{E}_n(T)$, (1) $\varphi > 0$ on $[0,T]$, and (2) if $\pi_{t+h} \neq \pi^*$, $\varphi(h+1) \leq \varphi(h) - 1$. We deduce that the number of times UCRL2-PT can sample from a suboptimal policy in $[0,T]$ is bounded above by $\varphi(0)$. Specifically,

$$\left|\left\{h \in [0,T] : \pi_{t_{k(n)}+h} \neq \pi^*\right\}\right| \leq 2S^2A(1 + C_{\text{exp}}\log(T)). = O(\log(T)). \tag{B.18}$$

To link the visit counts to the expected regret, we rely on the notion of *gap-regret*. Introducing the Bellman-gaps $\Delta^*(x, a) := h^*(x) + g^*(x) - r(x, a) - \langle p(x, a), h^* \rangle$, the gap-regret from $t = a$ to b is given by

$$\operatorname{GapReg}(a; b) := \sum_{i=a}^{b-1} \Delta^*(Z_i) = \sum_{z \in \mathcal{Z}} (N_b(z) - N_a(z))\, \Delta^*(z). \tag{B.19}$$

The name is given by the fact that $\Delta^*(x, a)$ can be interpreted as by how much picking action a from state x is suboptimal. We can actually show that for all stopping time τ and $T \geq 1$, we have $\mathbb{E}[\operatorname{Reg}(\tau; \tau + T)] \leq \operatorname{sp}(h^*) + \mathbb{E}[\operatorname{GapReg}(\tau; \tau + T)]$. Now, on $\mathcal{E}_n(T)$,

$$\begin{aligned}
\operatorname{GapReg}\left(t_{k(n)}; t_{k(n)} + T\right) &\leq \sum_{z \notin \operatorname{supp}(\mu_{\pi^*})} \left(N_{t_{k(n)}+T}(z) - N_{t_{k(n)}}(z)\right) \Delta^*(z) && \text{(B.20)} \\
&\leq \left|\left\{h \in [0, T] : \pi_{t_{k(n)}+h} \neq \pi^*\right\}\right| \cdot \max_z \Delta^*(z) && \text{(B.21)} \\
&\leq 2S^2 A(1 + C_{\exp} \log(T)) \cdot \max_z \Delta^*(z). && \text{(B.22)}
\end{aligned}$$

Moreover, because the MDP is deterministic with at most S states, we have $\Delta^*(z) \leq S$ for all z, that can be used to bound $\max_z \Delta^*(z)$ in the above. Also, it implies that $\operatorname{GapReg}(t_{k(n)}; t_{k(n)} + T) \leq ST$ a.s.. Overall $\mathbb{E}[\operatorname{GapReg}(t_{k(n)}; t_{k(n)} + T)] \leq S + 2S^3 A(1 + C_{\exp} \log(T))$. So, because $t_{k(n)}$ is a stopping time,

$$\begin{aligned}
\mathbb{E}\left[\operatorname{Reg}(t_{k(n)}; t_{k(n)} + T)\right] &\leq S + \mathbb{E}\left[\operatorname{GapReg}(t_{k(n)}; t_{k(n)} + T)\right] && \text{(B.23)} \\
&= O(\log(T)). && \text{(B.24)}
\end{aligned}$$

This holds for all n large enough w.r.t. T. Hence $\operatorname{RegExp}(T) = O(\log(T))$.

Remark 2. The proof actually provides a high probability upper bound of the regret on $[t_{k(n)}, t_{k(n)} + T]$ when n grows large.

B.3. Laziness and Proof of Lemma 9

Lemma 10 (Weak Laziness)**.** *Let π a policy. If π is iterated m times over a time-window $[t, t + h]$ (i.e., $\{i \in [t, t + h], \pi_i = \pi\}$ is of cardinality m), then*

$$z \in \operatorname{supp}(\mu_\pi), \quad N_{t+h}(z) - N_t(z) \geq \frac{m}{2S} - 1. \tag{B.25}$$

Proof. Let $K(\pi)$ the episodes over $[t, t + h - 1]$ where the current policy is π. In abuse of notations, denote $t_k, t_{k+1} - 1$ the respective beginning and ending time-instants of episode k truncated to $[t, t + h]$. Let $z \in \operatorname{supp}(\mu_\pi)$. We have:

$$\begin{aligned}
N_{t+h}(z) - N_t(z) &= \sum\nolimits_{i=t}^{t+h-1} \mathbf{1}\{Z_i = z\} && \text{(B.26)} \\
&\geq \sum\nolimits_{k \in K(\pi)} \sum\nolimits_{i=t_k}^{t_{k+1}-1} \mathbf{1}\{Z_i = z\} && \text{(B.27)} \\
&\geq \sum\nolimits_{k \in K(\pi)} \frac{1}{2} \max_{z' \in \operatorname{supp}(\mu_\pi)} \sum\nolimits_{t=t_k}^{t_{k+1}-1} \mathbf{1}\{Z_i = z'\} - 1 && \text{(B.28)} \\
&\geq \sum\nolimits_{k \in K(\pi)} \frac{t_{k+1} - t_k}{S} - 1 && \text{(B.29)} \\
&= \frac{m}{2S} - 1 && \text{(B.30)}
\end{aligned}$$

where the second inequality relies on the laziness rule. This proves the result. □

Lemma 9 (Strong Laziness) *Let z a transition and denote $m_h(z)$ the number of times over $[t_{k(n)}, t_{k(n)} + h]$ when z is a recurrent transition of the exploration policy, i.e., $m_h(z)$ is the cardinal of $\{i \in [t_{k(n)}, t_{k(n)} + h] : z \in \operatorname{supp}(\mu_{\pi_i})\}$. Then*

$$N_{t_{k(n)}+h}(z) - N_{t_{k(n)}}(z) \geq \frac{m_h(z)}{2S} - 1. \tag{B.31}$$

Proof. The proof is mostly the same as the one of Lemma 10. For short, denote $t = t_{k(n)}$. Let $K(z)$ the episodes over $[t, t+h-1]$ where $z \in \mathrm{supp}(\mu_{\pi^\ell})$. In abuse of notations, denote $t_\ell, t_{\ell+1} - 1$ the respective beginning and ending time-instants of episode ℓ truncated to $[t, t+h]$. Let $z \in \mathrm{supp}(\mu_\pi)$. We have:

$$N_{t+h}(z) - N_t(z) = \sum_{i=t}^{t+h-1} \mathbf{1}\{Z_i = z\} \tag{B.32}$$

$$\geq \sum_{\ell \in K(z)} \sum_{i=t_\ell}^{t_{\ell+1}-1} \mathbf{1}\{Z_i = z\} \tag{B.33}$$

$$\geq \sum_{\ell \in K(z)} \frac{1}{2} \max_{z' \in \mathrm{supp}(\mu_{\pi^\ell})} \sum_{t=t_\ell}^{t_{\ell+1}-1} \mathbf{1}\{Z_i = z'\} - 1 \tag{B.34}$$

$$\geq \sum_{\ell \in K(z)} \frac{t_{\ell+1} - t_\ell}{S} - 1 \tag{B.35}$$

$$= \frac{m_h(z)}{2S} - 1 \tag{B.36}$$

where the second inequality relies on the laziness rule. □

B.4. Proof of Lemma 5

We establish the following claim:

Lemma 5 *There is a time-threshold $t_{\mathrm{visits}} : T \mapsto t_{\mathrm{visits}}(T)$ such that, provided that $t \geq t_{\mathrm{visits}}(T)$, on the good event $\mathcal{E}(T)$,*

(1) $N_t(z) \geq C^*_{\mathrm{visits}} t$ for $z \in \mathrm{supp}(\mu_{\pi^*})$;
(2) $N_t(z) \geq C^-_{\mathrm{visits}} \log(t)$ for $z \notin \mathrm{supp}(\mu_{\pi^*})$; and
(3) $N_t(z) \leq C^+_{\mathrm{visits}} \log(t)$ for $z \notin \mathrm{supp}(\mu_{\pi^*})$.

These are not established in order; We start with (3), then use it to prove (1) and finally establish (2). Remark first that $\mathcal{E}(T)$ provides deviations of $\hat{r}_t(z)$ of order $\sqrt{\log(\log(t))/N_t(z)}$ that are asymptotically smaller than the optimism $\xi_t(z)$ given by $\sqrt{\kappa_1 \log(\kappa_2 t)/N_t(z)}$. Therefore, there is a time-threshold $t_{\mathrm{opt}} : T \mapsto T_{\mathrm{opt}}(T)$ such that, on $\mathcal{E}(T)$,

$$\forall t \geq t_{\mathrm{opt}}(T), \forall \pi, \quad \tilde{g}_t(\pi) \geq g(\pi). \tag{B.37}$$

In other word, beyond $t_{\mathrm{opt}}(T)$, optimistic values are indeed optimistic.

Now, fix $T \geq 1$.

Proof of (3) Step 1: *If $\pi_t \neq \pi^*$, then $N_t(z) \leq C^+_{\mathrm{visits}} \log(t)$.*

We show first that there are constants C_1, C_2 such that, on the good event $\mathcal{E}(T)$, if π_t is suboptimal policy π, then there is $z \in \mathrm{supp}(\mu_\pi)$ such that $N_t(z) \leq C_1 + C_2 \log(t)$. So, let π a suboptimal policy and assume that $\pi_t = \pi$ for some $t \geq t_{\mathrm{opt}}(T) + S$. Then, there is $t' = t - h$ with $h \leq S$, (**PT**) holds, i.e., we have $\tilde{g}_{t'}(\pi) \geq \tilde{g}_{t'}(\pi^*) - \sqrt{\alpha \log(t')/t'}$. So $\tilde{g}_{t'}(\pi) - g(\pi) \geq \tilde{g}_{t'}(\pi^*) - g(\pi) - \sqrt{\alpha \log(t')/t'}$. But on the good event $\mathcal{E}(T)$, $\tilde{g}_{t'}(\pi^*) \geq g_{\pi^*}$ and $\tilde{g}_{t'}(\pi) \geq g_\pi$, see (B.37). From it follows that, on $\mathcal{E}(T)$,

$$\tilde{g}_{t'}(\pi) - g(\pi) \geq \Delta(\pi^*; \pi) - \sqrt{\frac{\alpha \log t'}{t'}}. \tag{B.38}$$

For t large enough (for example $t \geq \frac{16\alpha^2}{\Delta(\pi^*;\pi)^4} + S$), holds that $\sqrt{\alpha \log(t')/t'} \leq \frac{1}{2}\Delta(\pi^*; \pi)$ for all suboptimal π. Also, using that $\tilde{r}_{t'} \leq r + 2\xi_{t'}$ on the $\mathcal{E}(T)$, the above equation takes the alternative form:

$$\frac{1}{2}\Delta(\pi^*; \pi). \leq \langle \mu_\pi, \tilde{r}_{t'} - r \rangle \leq \langle \mu_\pi, r + 2\xi_{t'} - r \rangle = 2 \langle \mu_\pi, \xi_{t'} \rangle \tag{B.39}$$

The size of the support of μ_π is at most S. So there is $z \in \mathrm{supp}(\mu_\pi)$ such that $\frac{1}{4S}\Delta(\pi^*; \pi) \leq \mu_\pi(z)\xi_{t'}(z) \leq \xi_{t'}(z)$. Unfolding the expression of $\xi_{t'}$, we get

$$\sqrt{\frac{\kappa_1 \log(\kappa_2 t')}{N_{t'}(z)}} \geq \frac{\Delta(\pi^*; \pi)}{4S}$$

Rearranging terms yields that, on the good event, for $t \geq \frac{16\alpha^2}{\Delta(\pi^*;\pi)^4} + S$,

$$N_{t'}(z) \leq \frac{16S^2\kappa_1}{\Delta(\pi^*;\pi)^2} \cdot \log(\kappa_2 t') \leq \frac{16S^2\kappa_1}{\Delta_{\min}^2} \cdot \log(\kappa_2 t')$$

The derivative of $t \mapsto \log(\kappa_2 t)$ vanishes and $t - S \leq t' \leq t$, so up to choosing t large enough, we can change the bound to $C_1 + 16S^2\kappa_1\Delta_{\min}^{-2}\log(\kappa_2 t)$.

We have established the following: There exists $C_1, C_2 > 0$ such that, provided that t is large enough (w.r.t. T), if $\pi_t \neq \pi^*$, then there is $z \in \operatorname{supp}(\mu_{\pi_t})$ such that, on $\mathcal{E}(T)$,

$$N_t(z) \leq C_1 + C_2 \log(t). \tag{B.40}$$

Up to increasing C_1, we can assume that this holds for all t.

Proof of (3) Step 2: *If $z \notin \pi^*$, then $N_t(z) \leq C^+_{\text{visits}} \log(t)$.*

Continuing with the proof, introduce the quantity:

$$\psi_t(i) := \sum\nolimits_{z \in \mathcal{Z}} \left|[1 + C_1 + C_2 \log(t)] - N_i(z)\right|_+ \tag{B.41}$$

We see that if $\pi_i \neq \pi^*$ for $i \leq t$, then $\psi_t(i) > 0$. Denote $\Lambda_-(t) := \{i \leq t : \pi_i \neq \pi^*\}$ the (ordered) collection of timesteps where the current policy is suboptimal. The key observation is that every $2S$ consecutive elements of $\Lambda_-(t)$, we complete a turn of a suboptimal policy, hence $\psi_t(\cdot)$ decreases by at least one. Therefore, $|\Lambda_-(t)| \leq 2S \cdot SA \cdot (1 + C_1 + C_2 \log(t))$ on the good event $\mathcal{E}(T)$. Hence, for all $z \notin \pi^*$, on $\mathcal{E}(T)$,

$$N_t(z) \leq |\Lambda_-(t)| \leq 2S^2A(1 + C_1) + 2S^2AC_2 \log(t). \tag{B.42}$$

Proof of (3) Step 3: *If $z \notin \operatorname{supp}(\mu_{\pi^*})$, then $N_t(z) \leq C^+_{\text{visits}} \log(t)$.*

We want to extend this to a bound on $N_t(z)$ for $z \in \pi^* \setminus \operatorname{supp}(\mu_{\pi^*})$ as well. By definition of the rule of a change of episode, the optimal policy changes from an episode to another Because every consecutive optimal episodes must be intertwined with at least one suboptimal episode, the total number of episodes K is bounded by $2|\Lambda_-(t)|$. Now, fix $z \in \pi^* \setminus \operatorname{supp}\mu_{\pi^*}$. Fix $t \geq T(\alpha)$ and denote K the set of episodes in $[1, t]$. Let $K_+ := \left\{k \in K : \pi^k = \pi^*\right\}$ and $K_-(z) := \left\{k \in K : z \in \pi^k \neq \pi^*\right\}$. For all $k \in K_-(z)$, there is some $z_k \in \operatorname{supp}\pi^k$ such that $N_{t_{k+1}}(z_k) - N_{t_k}(z_k) \geq 1 + N_{t_{k+1}}(z) - N_{t_k}(z)$. Now,

$$\begin{aligned}
N_t(z) &= \sum\nolimits_{k \in K_+} \sum\nolimits_{i=t_k}^{t_{k+1}-1} \mathbf{1}\{Z_i = z\} + \sum\nolimits_{k \in K_-(z)} \sum\nolimits_{i=t_k}^{t_{k+1}-1} \mathbf{1}\{Z_i = z\} && \text{(B.43)}\\
&\leq \sum\nolimits_{k \in K_+} 1 + \sum\nolimits_{k \in K_-(z)} \left(1 + \sum\nolimits_{i=t_k}^{t_{k+1}-1} \mathbf{1}\{Z_i = z_k\}\right) && \text{(B.44)}\\
&\leq |K| + \sum\nolimits_{k \in K_-(z)} \sum\nolimits_{z' \notin \pi^*} \sum\nolimits_{i=t_k}^{t_{k+1}-1} \mathbf{1}\{Z_i = z'\} && \text{(B.45)}\\
&\leq |K| + \sum_{z' \notin \pi^*} N_t(z') && \text{(B.46)}
\end{aligned}$$

that we bound accordingly by

$$N_t(z) \leq 3|\Lambda_-(z)| \leq 6S^2A(1 + C_1) + 6S^2AC_2 \log(t) \tag{B.47}$$

where $C_2 = 16\Delta_{\min}^{-2}S^2\kappa_1$. By choosing t large enough, the bound is converted to $N_t(z) \leq 12S^2AC_2 \log(t)$.

Proof of (1)

In the previous (3), we have shown that $\Lambda_-(t) := \{i \leq t : \pi_i \neq \pi^*\}$ is of cardinality at most $2S^2A(1 + C_1 + C_2 \log(t))$ on the good event $\mathcal{E}(T)$. By choosing t large enough, we deduce that $\Lambda_-(t)$ is of cardinality at most $4S^2AC_2 \log(t) \leq \frac{1}{2}t$ on $\mathcal{E}(T)$. Therefore, π^* is iterated at least $\frac{1}{2}t$ times over $[1, t]$. By Lemma 10,

$$\forall z \in \operatorname{supp}(\mu_{\pi^*}), \quad N_t(z) \geq \tfrac{1}{4S}t. \tag{B.48}$$

Proof of (2)

Assume that we are on the good event $\mathcal{E}(T)$ with t large. Following what have been said in the proof of (1), π^* is used at least $\frac{1}{2}t$ times over $[1, t]$. So, there exists $t' \geq \frac{1}{2}t$ such that $\pi_{t'} = \pi^*$. By definition, (**PT**) holds for some $t' - h$ with $h \leq S$, i.e.,

$$\tilde{g}^*_{t'-h} - \tilde{g}_{t'-h}(\pi^*) \leq \sqrt{\frac{\alpha \log(t' - h)}{t' - h}}. \tag{B.49}$$

Up to choosing t large enough, we can assume that $t'' := t' - h \geq \frac{1}{3}t$. We can t'' large as well, since t is. So, for all $z \in \operatorname{supp}(\mu_{\pi^*})$, $N_{t''}(z) \geq \frac{1}{12S}t$. So at time t'', the optimistic value of π^* (which is also close to $\tilde{g}^*_t$) is close to the true value $g(\pi^*)$. More precisely,

$$\tilde{g}^*_{t''} \leq \tilde{g}_{t''}(\pi^*) + \sqrt{\tfrac{\alpha \log(t'')}{t''}} = \langle \mu_{\pi^*}, \hat{r}_{t''} + \xi_{t''} \rangle + \sqrt{\tfrac{\alpha \log(t'')}{t''}} \tag{B.50}$$

$$\leq \langle \mu_{\pi^*}, r \rangle + 2 \langle \mu_{\pi^*}, \xi_{t''} \rangle + \sqrt{\tfrac{\alpha \log(t'')}{t''}} \tag{B.51}$$

$$\leq g(\pi^*) + 2 \sum\nolimits_{z \in \operatorname{supp}(\mu_{\pi^*})} \mu_{\pi^*}(z) \sqrt{\frac{\kappa_1 \log(\kappa_2 t'')}{N_{t''}(z)}} + \sqrt{\tfrac{\alpha \log(t'')}{t''}} \tag{B.52}$$

$$\leq g(\pi^*) + 6 \sqrt{\tfrac{S}{t''} \kappa_1 \log(\kappa_2 t'')} + \sqrt{\tfrac{\alpha \log(t'')}{t''}} \tag{B.53}$$

so $\tilde{g}^*_{t''} \leq g(\pi^*) + 1$ when t'' (so t) is large enough. Now, we use the fact that $\xi_t(z)$ is way larger that it needs to be on $\mathcal{E}(T)$. Let $z \notin \operatorname{supp}(\mu_{\pi^*})$. There exists π such that $z \in \operatorname{supp}(\mu_\pi)$ by communicativity of the DMDP. Moreover, on $\mathcal{E}(T)$,

$$\tilde{g}_{t''}(\pi) = \langle \mu_\pi, \tilde{r}_{t''} \rangle \tag{B.54}$$

$$\geq g(\pi) + \sum\nolimits_{z'} \mu_\pi(z') \left| \hat{r}_{t''}(z') - r(z') \right| + \sum\nolimits_{z'} \mu_\pi(z') \xi_{t''}(z') \tag{B.55}$$

$$\geq g(\pi) - \sum\nolimits_{z'} \mu_\pi(z') \sqrt{\frac{c \log(SAT(d + \log(t'')))}{N_{t''}(z')}} + \sum\nolimits_{z'} \mu_\pi(z') \sqrt{\frac{\kappa_1 \log(\kappa_2 t'')}{N_{t''}(z')}} \tag{B.56}$$

$$\geq g(\pi) + \frac{1}{2} \sum\nolimits_{z'} \mu_\pi(z') \sqrt{\frac{\kappa_1 \log(\kappa_2 t'')}{N_{t''}(z')}} \tag{B.57}$$

where the last inequality is holds when $\log(\kappa_2 t'')$ is large in front of $\log(SAT(d + \log(t'')))$. Together, we get

$$\mu_\pi(z) \sqrt{\frac{\kappa_1 \log(\kappa_2 t'')}{N_{t''}(z)}} \leq 2 + 2\Delta(\pi). \tag{B.58}$$

Solve in $N_{t''}(z)$, we end up with

$$N_t(z) \geq N_{t''}(z) \geq \frac{\kappa_1 \log(\kappa_2 t'')}{4 \mu_\pi(z)^2 (1 + \Delta(\pi))^2} \geq C + \frac{\kappa_1 \log(\kappa_2 t)}{8S^2 (1 + \Delta(\pi))^2}$$

for some constant C, independent of t. Therefore, on the good event, visit counts are $\Omega(\log(t))$, finally proving the claim.

B.5. Proof of Lemma 7

This section is dedicated to the proof of:

Lemma 7 *There exist constants $C_{\text{noise}}, C_{\text{drift}}, C_{\pi^*} > 0$ and a non-decreasing function $t_0 : T \mapsto t_0(T)$ such that, on $\mathcal{E}(T)$, for all $T \geq 1$ and all $t \geq t_0(T)$, we have*

$$\begin{aligned}
\forall \pi \neq \pi^*, \forall h \leq T, \quad \mathcal{D}_h\left[\tilde{\Delta}_t(\pi^*; \pi)\right](t) \geq &+ C_{\text{drift}} \sum\nolimits_{\substack{z \in \operatorname{supp}(\mu_\pi) \\ z \notin \operatorname{supp}(\mu_{\pi^*})}} \frac{N_{t+h}(z) - N_t(z)}{N_t(z)} \\
&- C_{\text{noise}} \sum\nolimits_{\substack{z \in \operatorname{supp}(\mu_\pi) \\ z \notin \operatorname{supp}(\mu_{\pi^*})}} \frac{\sqrt{(N_{t+h}(z) - N_t(z)) \log(T)}}{N_t(z)} \\
&- C_{\pi^*} \cdot \frac{h}{\sqrt{t}},
\end{aligned}$$

provided that $\forall z, \forall h \leq T$, $\left|\sum_{i=0}^{h}(R_{t+i} - r(z))\mathbf{1}\{Z_{t+i} = z\}\right| \leq \sqrt{\frac{1}{2}(N_{t+h}(z) - N_t(z))\log(SAT^3)}$.

This result is based on an explicit lower bound on the optimistic gap deviations.

Lemma 11 (Optimistic Gap Deviations)**.** *For all policy π, the h-step deviations of the optimistic gap $\tilde{\Delta}_t(\pi^*; \pi) := \tilde{g}_t(\pi^*) - \tilde{g}_t(\pi)$, that is, $[\mathcal{D}_h \tilde{\Delta}_t(\pi^*; \pi)](t)$, is lower bounded by:*

$$\begin{array}{lcc}
+\sum_{z\in\operatorname{supp}\mu_{\pi^*}} (\mu_{\pi^*}(z) - \mu_\pi(z)) \dfrac{\sum_{i=1}^{h}(R_{t+i} - r(z))\mathbf{1}\{Z_{t+i} = z\}}{N_{t+h}(z)} & \text{(I)} & \begin{array}{c}\textsc{Stochastic Revision}\\ \textsc{on Optimal Cycle}\end{array} \\
-\sum_{z\in\operatorname{supp}\mu_{\pi^*}} |\mu_{\pi^*}(z) - \mu_\pi(z)| \dfrac{(N_{t+h}(z) - N_t(z))\,|\hat{r}_t(z) - r(z)|}{N_t(z)} & \text{(II)} & \begin{array}{c}\textsc{Initial Error Drift}\\ \textsc{on Optimal Cycle}\end{array} \\
-\sum_{z\in\operatorname{supp}\mu_{\pi^*}} |\mu_{\pi^*}(z) - \mu_\pi(z)| \dfrac{(N_{t+h}(z) - N_t(z))\sqrt{\kappa_1 \log(\kappa_2(t+h))}}{2N_t(z)^{3/2}} & \text{(III)} & \begin{array}{c}\textsc{Optimism Drift}\\ \textsc{on Optimal Cycle}\end{array} \\
+\sum_{z\notin\operatorname{supp}\mu_{\pi^*}} \mu_\pi(z) \dfrac{\sum_{i=1}^{h}(r(z) - R_{t+i})\mathbf{1}\{Z_{t+i} = z\}}{N_{t+h}(z)} & \text{(IV)} & \begin{array}{c}\textsc{Stochastic Revision}\\ \textsc{on Subopt. Transitions}\end{array} \\
+\sum_{z\notin\operatorname{supp}\mu_{\pi^*}} \mu_\pi(z) \dfrac{(N_{t+h}(z) - N_t(z))\,|\hat{r}_t(z) - r(z)|}{N_t(z)} & \text{(V)} & \begin{array}{c}\textsc{Initial Error Drift}\\ \textsc{on Subopt. Transitions}\end{array} \\
+\sum_{z\in\operatorname{supp}\mu_{\pi^*}} \mu_\pi(z) \dfrac{(N_{t+h}(z) - N_t(z))\sqrt{\kappa_1 \log(\kappa_2(t+h))}}{2N_t(z)^{3/2}} & \text{(VI)} & \begin{array}{c}\textsc{Optimism Drift}\\ \textsc{on Subopt. Transitions}\end{array} \\
-\sum_{z} \mu_\pi(z) \sqrt{\dfrac{\kappa_1}{N_t(z)}} \cdot \dfrac{h}{2t\sqrt{\log(\kappa_2 t)}} & \text{(VII)} & \textsc{Logarithmic Drift}
\end{array}$$

Proof of Lemma 11. Recall that $\tilde{g}_t(\pi) = \langle \mu(\pi), \hat{r}_t + \xi_t \rangle$, hence we have

$$\tilde{\Delta}_t(\pi^*; \pi) = \Delta(\pi^*; \pi) + \underbrace{[\hat{g}_t(\pi^*) - g(\pi^*)]}_{\textcircled{1}} + \underbrace{[g(\pi) - \hat{g}_t(\pi)]}_{\textcircled{2}} + \underbrace{\langle \mu(\pi^*), \xi_t \rangle}_{\textcircled{3}} + \underbrace{\langle \mu(\pi), -\xi_t \rangle}_{\textcircled{4}} \tag{B.59}$$

It is easy to see that $\mathcal{D}_h$ is a linear operator on $\mathbb{N} \to \mathbb{R}$, i.e., commute with addition and scalar multiplication. Therefore,

$$[\mathcal{D}_h \tilde{\Delta}_t(\pi^*; \pi)](t) = \left[\mathcal{D}_h \textcircled{1}\right](t) + \left[\mathcal{D}_h \textcircled{2}\right](t) + \left[\mathcal{D}_h \textcircled{3}\right](t) + \left[\mathcal{D}_h \textcircled{4}\right](t). \tag{B.60}$$

Rewriting $\textcircled{1}$. First, remark that if $\{U_i\}$ are random variables with sums $S_n := U_1 + \cdots + U_n$ and empirical means $\hat{U}_n := \frac{1}{n} S_n$, then we have the identity $\hat{U}_{n+m} - \hat{U}_n = \frac{1}{n+m}\left(\sum_{i=1}^{m} U_{n+i} - m\hat{U}_n\right)$. We will rely on this to rewrite $\hat{r}_{t+h}(z) - \hat{r}_t(z)$.

By linearity of $\mathcal{D}_h$,

$$\begin{aligned}
\left[\mathcal{D}_h \textcircled{1}\right](t) &:= [\mathcal{D}_h(\hat{g}_t(\pi^*) - g(\pi^*))](t) \\
&= \langle \mu(\pi^*), [\mathcal{D}_h(\hat{r}_t - r)](t) \rangle \\
&= \sum_z \mu_{\pi^*}(z) [\mathcal{D}_h(\hat{r}_t(z) - r(z))](t) \\
&= \sum_z \mu_{\pi^*}(z) \left(\frac{\sum_{i=1}^{h}(R_{t+i} - r(z))\mathbf{1}\{Z_{t+i} = z\}}{N_{t+h}(z)} - \frac{(N_{t+h}(z) - N_t(z))(\hat{r}_t(z) - r(z))}{N_{t+h}(z)} \right) \\
&\geq \sum_z \mu_{\pi^*}(z) \left(\frac{\sum_{i=1}^{h}(R_{t+i} - r(z))\mathbf{1}\{Z_{t+i} = z\}}{N_{t+h}(z)} - \frac{(N_{t+h}(z) - N_t(z))\,|\hat{r}_t(z) - r(z)|}{N_t(z)} \right).
\end{aligned}$$

Rewriting $\textcircled{2}$. Here, also notice that if $\{U_i\}$ are random variables with sums $S_n := U_1 + \cdots + U_n$ and empirical means $\hat{U}_n := \frac{1}{n} S_n$, then we have the other identity $\hat{U}_{n+m} - \hat{U}_n = \frac{1}{n}\left(\sum_{i=1}^{m} U_{n+i} - m\hat{U}_{n+m}\right)$. We will rely on this to rewrite $\hat{r}_{t+h}(z) - \hat{r}_t(z)$.

With the same calculations,

$$\begin{aligned}\left[\mathcal{D}_h \textcircled{2}\right](t) &:= \left[\mathcal{D}_h(g(\pi) - \hat{g}_t(\pi))\right](t) \\ &\geq \sum_z \mu_\pi(z) \left(\frac{\sum_{i=1}^h (r(z) - R_{t+i}) \mathbf{1}\{Z_{t+i} = z\}}{N_{t+h}(z)} - \frac{(N_{t+h}(z) - N_t(z)) \left| r(z) - \hat{r}_t(z) \right|}{N_{t+h}(z)} \right).\end{aligned}$$

Rewriting $\textcircled{3}$. Recall that $\xi_t(z) := \sqrt{\kappa_1 \log(\kappa_2 t)/N_t(z)}$. Therefore,

$$\begin{aligned}\left[\mathcal{D}_h \textcircled{3}\right](t) = \langle \mu(\pi^*), \xi_{t+h} - \xi_t \rangle &= \sum\nolimits_z \mu_{\pi^*}(z) \left(\sqrt{\frac{\kappa_1 \log(\kappa_2(t+h))}{N_{t+h}(z)}} - \sqrt{\frac{\kappa_1 \log(\kappa_2 t)}{N_t(z)}} \right) \\ &= \sum\nolimits_z \mu_{\pi^*}(z) \Bigg(\sqrt{\kappa_1 \log(\kappa_2(t+h))} \left(\frac{1}{\sqrt{N_{t+h}(z)}} - \frac{1}{\sqrt{N_t(z)}} \right) \\ &\qquad + \sqrt{\frac{\kappa_1}{N_t(z)}} \left(\sqrt{\log(\kappa_2(t+h))} - \sqrt{\log(\kappa_2 t)} \right) \Bigg) \\ &\geq \sum\nolimits_z \mu_{\pi^*}(z) \sqrt{\kappa_1 \log(\kappa_2(t+h))} \left(\frac{1}{\sqrt{N_{t+h}(z)}} - \frac{1}{\sqrt{N_t(z)}} \right) \\ &\geq -\sum\nolimits_z \mu_{\pi^*}(z) \sqrt{\kappa_1 \log(\kappa_2(t+h))} \cdot \frac{N_{t+h}(z) - N_t(z)}{2 N_t(z)^{3/2}}\end{aligned}$$

where for the last inequality, we've used that $[\mathcal{D}_h t^{-1/2}](n) \geq -\frac{1}{2} h n^{-3/2}$.

Rewriting $\textcircled{4}$. Similarly,

$$\begin{aligned}\left[\mathcal{D}_h \textcircled{4}\right](t) = \langle \mu(\pi^*), \xi_t - \xi_{t+h} \rangle &= \sum\nolimits_z \mu_\pi(z) \left(\sqrt{\frac{\kappa_1 \log(\kappa_2 t)}{N_t(z)}} - \sqrt{\frac{\kappa_1 \log(\kappa_2(t+h))}{N_{t+h}(z)}} \right) \\ &= \sum\nolimits_z \mu_\pi(z) \Bigg(\sqrt{\kappa_1 \log(\kappa_2(t+h))} \left(\frac{1}{\sqrt{N_t(z)}} - \frac{1}{\sqrt{N_{t+h}(z)}} \right) \\ &\qquad + \sqrt{\frac{\kappa_1}{N_t(z)}} \left(\sqrt{\log(\kappa_2 t)} - \sqrt{\log(\kappa_2(t+h))} \right) \Bigg) \\ &\geq \sum\nolimits_z \mu_\pi(z) \left(\sqrt{\kappa_1 \log(\kappa_2(t+h))} \cdot \frac{N_{t+h}(z) - N_t(z)}{2 N_{t+h}(z)^{3/2}} - \sqrt{\frac{\kappa_1}{N_t(z)}} \cdot \frac{h}{2t \sqrt{\log(\kappa_2 t)}} \right)\end{aligned}$$

To end the proof of Lemma 11, group terms within two groups: when $z \in \operatorname{supp}(\mu_{\pi^*})$ and when $z \in \operatorname{supp}(\mu_\pi) \setminus \operatorname{supp}(\mu_{\pi^*})$. □

Proof of Lemma 7. Fix $T \geq 1$. Let t large enough, and introduce

$$\mathcal{E}_t := \left\{ \forall z, \forall h \leq T : \left| \sum\nolimits_{i=0}^h (R_{t+i} - r(z)) \mathbf{1}\{Z_{t+i} = z\} \right| \leq \sqrt{\frac{1}{2}(N_{t+h}(z) - N_t(z)) \log(SAT^3)} \right\} \tag{B.61}$$

Focusing on what happens on $\mathcal{E}(T)$, we see that if $t \geq t_{\text{visits}}(T)$, then for all $z \in \operatorname{supp}(\mu_{\pi^*})$, $N_t(z) \geq C^*_{\text{visits}} t$, see Lemma 5. Hence, we immediately see that terms $\textcircled{I}$ $\textcircled{II}$ $\textcircled{III}$ and $\textcircled{VII}$ are of order

$$O\left(\frac{\sqrt{h \log(T)}}{t} + \frac{h}{t} \right) = O\left(\frac{h}{\sqrt{t}} \right)$$

on $\mathcal{E}_t \cap \mathcal{E}(T)$ when t is large w.r.t. T. This accounts for the term $C_{\pi^*} \frac{h}{\sqrt{t}}$ in the final bound.

We are left with the stochastic revision, the initial error drift and the optimism drift on suboptimal transitions.

On $\mathcal{E}_t$, the stochastic revision term is upper bounded as

$$\textcircled{IV} \geq -\sum\nolimits_{z \notin \mathrm{supp}(\mu_{\pi^*})} \mu_\pi(z) \left| \frac{\sum_{i=1}^{h} (r(z) - R_{t+i}) \mathbf{1}\{Z_{t+i} = z\}}{N_{t+h}(z)} \right| \tag{B.62}$$

$$\geq -\sum\nolimits_{z \notin \mathrm{supp}(\mu_{\pi^*})} \mu_\pi(z) \frac{\sqrt{\frac{1}{2}(N_{t+h}(z) - N_t(z)) \log(SAT^3)}}{N_t(z)} \tag{B.63}$$

$$\geq -C'_{\mathrm{noise}} \sum_{\substack{z \in \mathrm{supp}(\mu_\pi) \\ z \notin \mathrm{supp}(\mu_{\pi^*})}} \frac{\sqrt{(N_{t+h}(z) - N_t(z)) \log(T)}}{N_t(z)} \tag{B.64}$$

for some constant $C_{\mathrm{noise}} > 0$.

Then there is the optimism drift term, that we lower bound using the visit count lemma Lemma 5 to upper bound $N_t(z)$ when $z \notin \mathrm{supp}(\mu_{\pi^*})$. Specifically, on $\mathcal{E}(T)$, we have

$$\textcircled{V} \geq \sum\nolimits_{z \notin \mathrm{supp}(\mu_{\pi^*})} \mu_\pi(z) \frac{(N_{t+h}(z) - N_t(z)) \sqrt{\kappa_1 \log(\kappa_2(t+h))}}{2N_t(z)^{3/2}} \tag{B.65}$$

$$\geq \sum\nolimits_{z \notin \mathrm{supp}(\mu_{\pi^*})} \mu_\pi(z) \frac{(N_{t+h}(z) - N_t(z)) \sqrt{\kappa_1 \log(\kappa_2(t+h))}}{2N_t(z) \sqrt{C^+_{\mathrm{visits}} \log(t)}} \tag{B.66}$$

$$\geq C'_{\mathrm{drift}} \sum_{\substack{z \in \mathrm{supp}(\mu_\pi) \\ z \notin \mathrm{supp}(\mu_{\pi^*})}} \frac{N_{t+h}(z) - N_t(z)}{N_t(z)} \tag{B.67}$$

for some constant $C'_{\mathrm{drift}} > 0$. And finally there is the initial error drift term, that we show to be negligible on the good event $\mathcal{E}(T)$. That is:

$$\textcircled{V} \geq -\sum\nolimits_{z \notin \mathrm{supp}(\mu_{\pi^*})} \mu_\pi(z) \frac{(N_{t+h}(z) - N_t(z)) \left| \hat{r}_t(z) - r(z) \right|}{N_t(z)} \tag{B.68}$$

$$\geq -\sum\nolimits_{z \notin \mathrm{supp}(\mu_{\pi^*})} \mu_\pi(z) \frac{(N_{t+h}(z) - N_t(z)) \sqrt{c \log(T(d + \log(t)))}}{N_t(z) \sqrt{N_t(z)}} \tag{B.69}$$

$$\geq -\sum\nolimits_{z \notin \mathrm{supp}(\mu_{\pi^*})} \mu_\pi(z) \frac{(N_{t+h}(z) - N_t(z)) \sqrt{c \log(T(d + \log(t)))}}{N_t(z) \sqrt{C^+_{\mathrm{visits}} \log(t)}} \tag{B.70}$$

$$\geq -\frac{1}{2} C'_{\mathrm{drift}} \sum_{\substack{z \in \mathrm{supp}(\mu_\pi) \\ z \notin \mathrm{supp}(\mu_{\pi^*})}} \frac{N_{t+h}(z) - N_t(z)}{N_t(z)}. \tag{B.71}$$

The last inequality follows from the fact that $c \log(T(d + \log(t)))$ is negligible in front of $C^+_{\mathrm{visits}} \log(t)$ when t is large enough. Setting $C_{\mathrm{drift}} := C'_{\mathrm{drift}}/2$, we obtain the claim. □

C. Auxiliary Results

C.1. Controling Gain Variations

Lemma 12. *Let P_1, P_2 two transition kernels over $\mathcal{S}$, let $r_1, r_2 \in \mathbb{R}^{\mathcal{S}}$ two reward vectors. Denote $g_i := \lim \frac{1}{T} \sum_{t=0}^{T-1} P_i^t r_i$ the associated gain, and $h_i := \lim \sum_{t=0}^{T-1} P_i^t (r_i - g_i)$ the associated bias. If g_1 is a constant vector, then for all $x \in \mathcal{S}$,*

$$|g_1(x) - g_2(x)| \leq \|r_1 - r_2\|_\infty + 2 \,\mathrm{sp}(h_1) \,\|P_1 - P_2\|_1 . \tag{C.1}$$

Proof. Denote e_x the vector of $\mathbb{R}^{\mathcal{S}}$ such that $e_x(y) = \mathbf{1}\{x = y\}$. Unfolding the definition of the gain, we have

$$g_1(x) - g_2(x) = \lim_{T \to \infty} \left[e_x \cdot \frac{1}{T} \sum_{t=0}^{T-1} \left(P_1^t - P_2^t \right) r_\pi \right] + \lim_{T \to \infty} \left[e_x \cdot \frac{1}{T} \sum_{t=0}^{T-1} P_2^t (r_1 - r_2) \right]. \tag{C.2}$$

Following Hölder's inequality, the right-hand term is bounded as

$$\left| \lim_{T \to \infty} \left[e_x \cdot \frac{1}{T} \sum_{t=0}^{T-1} P_2^t (r_1 - r_2) \right] \right| \leq \left\| e_x \cdot \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} P_2^t \right\|_1 \|r_1 - r_2\|_\infty \leq \|r_1 - r_2\|_\infty . \tag{C.3}$$

We are left with the left term $\lim[e_x \cdot \frac{1}{T}\sum_{t=0}^{T-1}(P_1^t - P_2^t)r_1]$. Introduce $b \in \mathbb{R}^{\mathcal{S}}$ the bonus value defined as follows:

$$b(x) := 2\ \mathrm{sp}(h_2)\, \|P_1(\cdot|x) - P_2(\cdot|x)\|_1\,. \tag{C.4}$$

Let us show by induction on T that for all $T \geq 0$,

$$\sum_{t=0}^{T-1} P_1^t r_1 \leq \sum_{t=0}^{T-1} P_2^t (r_1 + b) \tag{C.5}$$

] where the inequality is meant componentwise. This is obvious for $T = 0$. For the induction case, one have

$$\begin{aligned}
\sum_{t=0}^{T} P_1^t r_1 - \sum_{t=0}^{T} P_2^t(r_1+b) \quad &= -b + P_1 \sum_{t=0}^{T-1} P_1^t r_1 - P_2 \sum_{t=0}^{T-1} P_2^t (r_1 + b) && \text{(C.6)}\\
&= -b + P_2 \left(\sum_{t=0}^{T-1} P_1^t r_1 - \sum_{t=0}^{T-1} P_2^t(r_1+b) \right) + (P_1 - P_2) \sum_{t=0}^{T-1} P_1^t r_1 && \text{(C.7)}\\
&\leq -b + (P_1 - P_2) \sum_{t=0}^{T-1} P_1^t r_1 && \text{(C.8)}
\end{aligned}$$

where the last inequality follows by induction. Let $\mathrm{J}_1(T)$ the vector which x-th coordinate is $\mathrm{J}_1(x,T) := e_x \cdot \sum_{t=0}^{T-1} P_1^t r_1$. Because $P_1(\cdot|x)$ and $P_2(\cdot|x)$ are probability vectors, for all scalar $\lambda \in \mathbb{R}$ and writing e the vector whose components are all 1s, we have

$$(P_1(\cdot|x) - P_2(\cdot|x))\,(\mathrm{J}_1(T) + \lambda e) = (P_1(\cdot|x) - P_2(\cdot|x))\,\mathrm{J}_1(T). \tag{C.9}$$

Thus, choosing $\lambda := -\min_{y\in\mathcal{S}} \mathrm{J}_1(y,T)$, we get

$$\begin{aligned}
\left| e_x \cdot (P_1 - P_2) \sum_{t=0}^{T-1} P_1^t r_1 \right| &= |(P_1(\cdot|x) - P_2(\cdot|x))\,(\mathrm{J}_1(T) + \lambda e)| && \text{(C.10)}\\
&\leq \|P_1(\cdot|x) - P_2(\cdot|x)\|_1 \cdot \|\mathrm{J}_1(T) + \lambda e\|_\infty && \text{(C.11)}\\
&= \|P_1(\cdot|x) - P_2(\cdot|x)\|_1 \cdot \mathrm{sp}(\mathrm{J}_1(T)). && \text{(C.12)}
\end{aligned}$$

We then link the span of the cumulative score $\mathrm{J}_1(T)$ to the span of the bias. By Bellman's identity, for all state x, $r_1(x) = g_1(x) + h_1(x) - P_1(\cdot|x)h_1$. By induction, one checks that

$$\sum_{t=0}^{T-1} P_1^t r_1 = T g^\pi + (I - P_1^T) h^\pi. \tag{C.13}$$

As $g_1(x) = g_1(y)$ this yields $\mathrm{J}_1(x,T) - \mathrm{J}_1(y,T) = (P_1^T(\cdot|y) - P_1^T(\cdot|x))h_1 + h_1(x) - h_1(y)$. Therefore, $\mathrm{J}_1(x,T) - \mathrm{J}_1(y,T) \leq 2\ \mathrm{sp}(h_1)$, i.e., $\mathrm{sp}(\mathrm{J}_1(T)) \leq 2\ \mathrm{sp}(h_1)$ and

$$e_x \cdot (P_1 - P_2) \sum_{t=0}^{T-1} P_1^t r_1 \leq 2\, \|P_1(\cdot|x) - P_2(\cdot|x)\|_1\, \mathrm{sp}(h_1) =: b(x).$$

This concludes the induction. So, for all $T \geq 1$,

$$e_x \cdot \frac{1}{T} \sum_{t=0}^{T-1} P_1^t r_1 \leq e_x \cdot \frac{1}{T} \sum_{t=0}^{T-1} P_2^t r_1 + 2\ \mathrm{sp}(h^\pi) \max_{y\in\mathcal{S}} \|P_1(\cdot|y) - P_2(\cdot|y)\|_1\,. \tag{C.14}$$

Going in the limit when T goes to infinity in (C.14), together with (C.2) and (C.3), we obtain

$$g_1(x) - g_2(x) \leq \|r_1 - r_2\|_\infty + 2\ \mathrm{sp}(h^\pi)\, \|P_1 - P_2\|_1\,. \tag{C.15}$$

The other direction is proved similarly by using reward penalties $-b(x)$ instead. □

C.2. Numerical Lemma for the Upper Bound of Episodes

Lemma 13 is crucial in the proof of the upper bound of the number of episodes, see Theorem 2.

Lemma 13. *Let $T \geq 3$ and fix $\lambda \leq T$. Let $\omega \in (0, 1]$ and $(x_n \mid n \geq 1)$ an integer-valued sequence with $x_1 := 1$ such that*

$$x_{n+1} \geq \left(1 + \left(\frac{\lambda x_n}{T}\right)^{\omega}\right) x_n.$$

If n is such that $x_n \leq T$, then $n \leq 3\left(\frac{T}{\lambda}\right)^{\frac{\omega}{\omega+1}} \log(T)$.

We further conjecture that the $\log(T)$ is not necessary, i.e., that if $x_n \leq T$ then $n = O((T/\lambda)^{\omega/(\omega+1)})$.

Proof. Define the integer valued sequence $x_{n+1} = \lceil (1 + (\lambda x_n/T)^{\omega}) x_n \rceil$ initialized to $x_1 = 1$ and analyze the increments of (x_n). Observe that $x_{n+1} > x_n$, so $x_{n+1} \geq x_n + 1$ and the sequence diverges to infinity. Setting $\beta := \frac{1}{\omega+1} \in (0, 1)$, for $k \geq 1$, we get

$$x_{n+1} = x_n + k \iff x_n \in \left(\left(\tfrac{T}{\lambda}\right)^{1-\beta} (k-1)^{\beta}, \left(\tfrac{T}{\lambda}\right)^{1-\beta} k^{\beta}\right] =: I_k.$$

The length of I_k is decreasing with k and in particular $\mathrm{Leb}(I_k) \leq \mathrm{Leb}(I_1) = (\frac{T}{\lambda})^{1-\beta}$. Accordingly, the integer-valued sequence (y_n) with $y_1 = 1$ defined by its increments

$$y_{n+1} = y_n + k \iff y_n \in \left(\left(\tfrac{T}{\lambda}\right)^{1-\beta} (k-1), \left(\tfrac{T}{\lambda}\right)^{1-\beta} k\right] \tag{C.16}$$

satisfies: $\forall n \geq 1, y_n \leq x_n$. Moreover,

$$\begin{aligned} \forall n \geq 1, \quad y_{n+1} &= y_n + \left\lceil y_n \left(\tfrac{\lambda}{T}\right)^{1-\beta} \right\rceil && \text{(C.17)} \\ &\geq y_n \left(1 + \left(\tfrac{\lambda}{T}\right)^{1-\beta}\right) && \text{(C.18)} \\ &\geq \dots && \text{(C.19)} \\ &\geq \left(1 + \left(\tfrac{\lambda}{T}\right)^{1-\beta}\right)^n. && \text{(C.20)} \end{aligned}$$

Let $n \geq 1$ such that $x_n \leq T$. Then $y_n \leq T$, hence $(1 + (\lambda/T)^{1-\beta})^{n-1} \leq T$. Thus

$$(n-1) \log\left(1 + \left(\tfrac{\lambda}{T}\right)^{1-\beta}\right) \leq \log(T). \tag{C.21}$$

Since $\lambda \leq T$, we have $(\lambda/T)^{1-\beta} \in (0, 1]$ so $\log(1 + (\lambda/T)^{1-\beta}) \geq \frac{1}{2}(\lambda/T)^{1-\beta}$. We obtain:

$$n \leq 1 + 2\left(\tfrac{T}{\lambda}\right)^{1-\beta} \log(T) \leq 3\left(\tfrac{T}{\lambda}\right)^{1-\beta} \log(T) \tag{C.22}$$

and as $1 - \beta = \frac{\omega}{\omega+1}$, this proves the claim. □

C.3. Standard Concentration Inequalities

Lemma 14 (Weissman's inequality). *Let $p(\cdot)$ a probability distribution on $\mathcal{S}$ and let $\{Y_i\}$ i.i.d. random variables with $Y_i \sim p(\cdot)$. Denote $\hat{p}_n(\cdot) := \frac{1}{n}\sum_{i=1}^n \mathrm{Dirac}(Y_i)$ the empirical distribution. For all $\delta > 0$ and $n > 0$, we have*

$$\mathbb{P}\left\{ \|\hat{p}_n(\cdot) - p(\cdot)\|_1 \geq 2\sqrt{\frac{S \log(1/\delta)}{n}} \right\} \leq \delta. \tag{C.23}$$

Lemma 15 (Azuma-Hoeffding inequality). *Let $\{U_n\}$ be a martingale with $|U_i| \leq c$ almost surely. Then, given a confidence level $\delta > 0$, we have*

$$\mathbb{P}\left\{ \left|\sum\nolimits_{i=1}^n U_i\right| \geq c\sqrt{2n \log\left(\tfrac{1}{\delta}\right)} \right\} \leq \delta. \tag{C.24}$$

Lemma 16 ((Kaufmann & Koolen, 2021)). *Assume that $q(z)$ is a Bernoulli distribution. There exist constants $c, d > 0$ such that for all $\delta > 0$,*

$$\mathbb{P}\left\{ \exists t \geq 1 : N_t(z) \left(\hat{r}_t(z) - r_t(z)\right)^2 > c\left(\log(d + \log(t)) + \log(1/\delta)\right) \right\} \leq \delta.$$