Adaptive reward-free exploration

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, Michal Valko

Adaptive Reward-Free Exploration

Emilie Kaufmann
CNRS & ULille (CRIStAL), Inria SequeL
emilie.kaufmann@univ-lille.fr

Pierre Ménard
Inria Lille, SequeL team
pierre.menard@inria.fr

Omar Darwiche Domingues
Inria Lille, SequeL team
omar.darwiche-domingues@inria.fr

Anders Jonsson
Universitat Pompeu Fabra
anders.jonsson@upf.edu

Edouard Leurent
Renault & Inria Lille, SequeL team
edouard.leurent@inria.fr

Michal Valko
DeepMind Paris
valkom@deepmind.com

Abstract

Reward-free exploration is a reinforcement learning setting recently studied by Jin et al. [17], who address it by running several algorithms with regret guarantees in parallel. In our work, we instead propose a more *adaptive* approach for reward-free exploration which directly reduces upper bounds on the maximum MDP estimation error. We show that, interestingly, our reward-free UCRL algorithm can be seen as a variant of an algorithm of Fiechter from 1994 [11], originally proposed for a different objective that we call *best-policy identification*. We prove that RF-UCRL needs $\mathcal{O}\left((SAH^4/\varepsilon^2)\log(1/\delta)\right)$ episodes to output, with probability $1-\delta$, an ε-approximation of the optimal policy for *any* reward function. We empirically compare it to oracle strategies using a generative model.

1 Introduction

Reinforcement learning problems are related to learning and/or acting with a good policy in an unknown, stochastic environment, which requires to perform the right amount of *exploration*. In this work, we consider the discounted episodic setting with discount $\gamma \in (0,1]$ and horizon H and model the environment as a Markov Decision Process (MDP) with finite state space $\mathcal{S}$ of size S and finite action space $\mathcal{A}$ of size $A \geq 2$, transition kernels $P = (p_h(\cdot|s,a))_{h,s,a}$ and reward function $r = (r_h(s,a))_{h,s,a}$ for $h \in [H]$[1], $(s,a) \in \mathcal{S} \times \mathcal{A}$. The value of a policy $\pi = (\pi_1, \dots, \pi_H)$ in step $h \in [H]$ is given by

$$V_h^\pi(s_h; r) \triangleq \mathbb{E}^\pi\Big[\sum_{\ell=h}^{H} \gamma^{h-1} r_h(s_\ell, \pi_\ell(s_\ell))\Big| s_h\Big].$$

In this definition we explicitly materialize the dependency in the reward function r, but the expectation also depends on the transition kernel: for all $\ell \in [H]$, $s_{\ell+1} \sim p_\ell(\cdot|s_\ell, \pi_\ell(s_\ell))$ and a reward with expectation $r_\ell(s_\ell, \pi_\ell(s_\ell))$ is generated. We denote by $\pi^\star$ the optimal policy, such that in every step $h \in [H], V_h^{\pi^\star}(s; r) \geq V_h^\pi(s; r)$ for any policy π, and by $V^\star$ its value function.

An online reinforcement learning algorithm successively generates trajectories of length H in the MDP, starting from an initial state s_1 drawn from some distribution P_0. The t-th trajectory is generated under a policy π^t which may depend on the data collected in the $(t-1)$ previous episodes.

[1] We use the shorthand $[n] = \{1, \dots, n\}$ for every integer $n \in \mathbb{N}^*$

Preprint. Under review.

Given a fixed reward function r, several objective have been considered in the literature: maximizing the total reward accumulated during learning, or minimizing some notion of regret [2], proposing a guess for a good policy after a sufficiently large number of episodes [11] or guarantee that the policies used during learning are most of the time ε-optimal [7], see Section 2 for a precise description.

Yet in applications, the reward function r is often handcrafted to incentivize some behavior from the RL agent, and its design can be hard, so that we may end up successively learning optimal policies for different reward functions. This is the motivation given by Jin et al. [17] for the *reward-free exploration problem*, in which the goal is to be able to approximate the optimal policy under *any* reward function after a single exploration phase. More precisely, an algorithm for reward-free exploration should generate a dataset $\mathcal{D}_N$ of N reward-free trajectories —with N as small as possible— such that, letting $\hat{\pi}_{N,r}$ be the optimal policy in the MDP $(\hat{P}_N, r)$ (where $\hat{P}_N$ is the empirical transition matrix based on the trajectories in $\mathcal{D}_N$), one has

$$\mathbb{P}\left(\text{ for all reward functions } r,\ \mathbb{E}_{s_1\sim P_1}\left[V_1^\star(s;r) - V_1^{\hat{\pi}_{N,r}}(s;r)\right] \leq \varepsilon\right) \geq 1-\delta. \tag{1}$$

The solution proposed by [17] builds on an algorithm proposed for the different *regret minimization* objective. In order to generate $\mathcal{D}_N$, their algorithm first run, for each (s,h), N_0 episodes of the Euler algorithm of [32] for the MDP $(P, r^{(s,h)})$ where $r^{(s,h)}$ is a reward function that gives 1 at step h if state s is visited, and 0 otherwise. For each (s,h), after the corresponding Euler has been executed, the N_0 policies used in the N_0 episodes of Euler are added to a policy buffer Φ. Once this policy buffer Φ (which contains $S\times H\times N_0$ policies) is complete , the $\mathcal{D}_N$ database is obtained by generating N episodes under N policies picked uniformly at random in Φ (with replacement). [17] provide a calibration of N and N_0 for which (1) holds, leading to a *sampling complexity*, i.e. a total number of exploration episodes, of $\mathcal{O}(\frac{S^2AH^5}{\varepsilon^2}\log\left(\frac{SAH}{\delta\varepsilon}\right) + \frac{S^4AH^7}{\varepsilon^2}\log^3\left(\frac{SAH}{\delta\varepsilon}\right))$.

In this paper, we propose an alternative, more natural approach to reward free exploration, that does not rely on any regret minimizing algorithm. We show that (a variant of) an algorithm proposed by Fiechter in 1994 [11] for Best Policy Identification (BPI) —a setting described in details in Section 2— can be used for reward-free exploration. We give a new, simple, sample complexity analysis for this algorithm which improves over that of [17]. This (new) algorithm can be seen as a reward-free variant of UCRL [15], and is designed to uniformly reduce the estimation error of the Q-value function of any policy under any reward function, which is instrumental to prove (1), as already noted by [17].

Related work Realistic reinforcement-learning applications often face a challenge of a sparse rewards which at the beginning provides no signal for decision-making. Numerous attempts were made to guide the exploration in the beginning, motivated by curiosity [26, 27], intrinsic motivation [23, 4], exploration bonuses [30, 25] mutual information [24] and many of its approximations, for instance with variational autoencoders [23].

Nonetheless, it is even more challenging to analyze the exploration and provide guarantees for it. A typical take is to consider a well defined proxy for exploration and analyze that. For example, [22, 13] cast the skill discovery as an ability to reach any state within L hops. Another example is to look for policies finding the stochastic shortest path [31, 5] or aiming for the maximum entropy [14]. In our work, we provide an adaptive counterpart to [17] for a *reward-free exploration*, described next.

2 From Best Policy Identification to Reward-Free Exploration

In this section, we formally present the reward free exploration problem, which is a particular PAC (Probability Approximately Correct) learning problem. We then contrast it with several other PAC reinforcement learning frameworks that have been studied in the literature.

Reward-free exploration An algorithm for Reward-Free Exploration (RFE) sequentially collects a database of trajectories in the following way. In each time step t, a policy $\pi^t = (\pi_h^t)_{h=1}^H$ is computed based on data from the $t-1$ previous episodes, a *reward-free episode* $z_t = (s_1^t, a_1^t, s_2^t, a_2^t, \ldots, s_H^t, a_H^t)$ is generated under the policy π^t in the MDP starting from a first state $s_1^t \sim P_0$: for all $h\in[H]$, $s_h^t \sim p_h(s_{h-1}^t, \pi^t(s_{h-1}^t)$ and the new trajectory is added to the database: $\mathcal{D}_t = \mathcal{D}_{t-1}\bigcup\{z_t\}$. At the end of each episode, the algorithm can decide to stop collecting data (we denote by τ its random stopping time) and outputs the dataset $\mathcal{D}_\tau$.

A RFE algorithm is therefore made of a triple $((\pi^t)_{t\in\mathbb{N}}, \tau, \mathcal{D}_\tau)$. The goal is to build an (ε, δ)-PAC algorithm according to the following definition, for which the *sample complexity*, that is the number of exploration episodes τ is as small as possible.

Definition 1 (PAC algorithm for RFE)**.** *An algorithm if (ε, δ)-PAC for reward free exploration if*

$$\mathbb{P}\left(\textit{for all reward function } r, \left|\mathbb{E}_{s_1\sim P_0}\left[V_1^\star(s_1;r) - V_1^{\hat{\pi}^\star_{\tau,r}}(s_1;r)\right]\right| \le \varepsilon\right) \ge 1-\delta,$$

where $\hat{\pi}^\star_{\tau,r}$ is the optimal policy in the MDP parameterized by $(\hat{P}^\tau, r)$, with $\hat{P}^\tau$ being the empirical transition kernel estimated from the dataset $\mathcal{D}_\tau$.

Sample complexity in RL For the discounted episodic setting that is our focus in this paper, in which learning proceeds by a sequence of episodes, the first formal PAC RL model was proposed by Fiechter in 1994 [11]. As in this framework a RL algorithm should also *output a guess for a near-optimal policy*, we refer to it as Best Policy Identification (BPI), formally defined below.

Definition 2 (PAC algorithm for BPI)**.** *A RL algorithm $(\pi^t)_{t\in\mathbb{N}}$ is (ε, δ)-PAC for best policy identification if it* returns a policy $\hat{\pi}_\tau$ *after some number of episodes τ that satisfies*

$$\mathbb{P}\left(\mathbb{E}_{s_1\sim P_0}\left[V_1^\star(s_1) - V_1^{\hat{\pi}_\tau}(s_1)\right] \le \varepsilon\right) \ge 1-\delta.$$

In the discounted setting that is the focus of [11], choosing a horizon $H = (1-\gamma)^{-1}\log((\varepsilon(1-\gamma))^{-1})$ the policy $\hat{\pi}_\tau$[2] outputed by an (ε, δ)-PAC algorithm for BPI with horizon H is 2ε-optimal in terms of the infinite horizon discounted value function. Yet, the assumption of a "restart button" that allows to learn a good policy for a discounted MDP based on successive episodes was presented as a limitation in subsequent works, which converged on a different notion of PAC algorithm. While the E^3 algorithm [21] stops in some state s_τ and outputs a policy $\hat{\pi}_\tau$ that needs to be ε-optimal in that state, other algorithms [3, 28, 29] do not output a policy, but are proved to be PAC-MDP according to a definition formalized by [19] for discounted or average reward MDPs. Under an (ε, δ)-PAC MDP algorithm generating a trajectory $(s_t)_{t\in\mathbb{N}}$, there is a polynomial number of time steps t in which $V^\star(s_t) - V^{\mathcal{A}_t}(s_t) > \varepsilon$ where $\mathcal{A}_t$ is the policy used in the future steps of the algorithms.

The notion of PAC-MDP algorithm was later transposed to the (discounted) episodic setting [7, 8] as an algorithm such that, with probability $1-\delta$, $\sum_{t=1}^\infty \mathbb{1}(V_1^\star(s_1^t) - V_1^{\pi^t}(s_1^t) > \varepsilon)$ is upper bounded by a polynomial in $S, A, 1/\varepsilon, 1/\delta$ and H. PAC-MDP seems to be the most studied PAC reinforcement learning framework these days. However, reward free exploration is closer to the BPI framework: in the latter, an algorithm should stop and output a guess for the optimal policy associated to a particular reward function r (possibly unknown and observed through samples), while in the former it should stop and be able to estimate the optimal policy associated to *any* reward function. In the next section, we show that a variant of the first algorithm proposed by Fiechter for BPI [11], that we call RF-UCRL can actually be used for the (harder ?) reward free exploration problem and provide a new sample complexity analysis for it. We discuss further the link between RFE and BPI in Section 5.

Finally, sample complexity results have also been given for reinforcement learning based on a generative model, in which one can build a database of transitions performed in an arbitrary order (without the constrain to generate episodes). In the discounted setting, Azar et al. [1] propose an improved analysis of model-based Q-Value Iteration (MB-QVI) [20], which samples n transitions from every state-action pair and run value-iteration in the estimated MDP. They show that with a total sampling budget $T = nSA = \mathcal{O}\left((cSA/((1-\gamma)^3\varepsilon^2))\log(SA/\delta)\right)$, the optimal Q-value in the estimated MDP $\hat{Q}$ satisfies $\|\hat{Q} - Q^\star\|_\infty \le \varepsilon$ with probability larger than $1-\delta$.

3 Reward-Free UCRL

To ease the presentation of our algorithm, we assume that the first state distribution P_0 is supported on a single state s_1. Following an observation from [11], this is without loss of generality, as we may otherwise consider an alternative MDP with an extra initial state s_0 with a single action a_0 that yield a null reward and from which the transitions are $P_0(\cdot|s_0, a_0) = P_0$. Indeed, letting $\tilde{V}_0^{\tilde{\pi}}$ denote the value of policy $\tilde{\pi}$ such that $\tilde{\pi}_0 = a_0$ and $\tilde{\pi}_{1:H} = \pi$ for any episodic problem of horizon $H+1$ and discount sequence $(1, 1, \gamma, \gamma^2, \dots)$, it holds that $\tilde{V}_0^\star(s_0;r) - \tilde{V}_0^{\tilde{\pi}}(s_0;r) = \mathbb{E}_{s_1\sim P_1}\left[V_1^\star(s_1;r) - V_0^\pi(s_0;r)\right]$.

[2] extended to select random actions for $h > H$

Notation For all $h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we let $n_h^t(s,a) = \sum_{i=1}^t \mathbb{1}\left((s_h^i, a_h^i) = (s,a)\right)$ be the number of times the state action-pair (s,a) was visited in step h in the first t episodes and $n_h^t(s,a,s') = \sum_{i=1}^t \mathbb{1}\left((s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\right)$. This permits to define the empirical transitions

$$\hat{p}_h^t(s'|s,a) = \frac{n_h^t(s,a,s')}{n_h^t(s,a)} \text{ if } n_h^t(s,a) > 0, \text{ and } \hat{p}_h^t(s'|s,a) = \frac{1}{S} \text{ otherwise.}$$

We denote by $\hat{V}_h^{t,\pi}(s;r)$ (resp. $\hat{Q}_h^{t,\pi}(s,a;r)$) the value (resp. Q-values) functions in the empirical MDP with transition kernels $\hat{P}^t$ and reward function r, where we recall that the Q-value of a policy π in a MDP with transitions $p_h(s'|s,a)$ and mean reward $r_h(s,a)$ is defined by $Q_h^\pi(s,a;r) = r_h(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_h(s'|s,a) V_{h+1}^\pi(s')$. Finally, we let $\sigma_h = \sum_{i=0}^{h-1} \gamma^i$ and note that $\sigma_h \leq h$.

Error upper bounds RF-UCRL is based on an upper bound on the estimation error for each policy π (and each value function r). For every π, r, t, we define this error as

$$\hat{e}_h^{t,\pi}(s,a;r) := |\hat{Q}_h^{t,\pi}(s,a;r) - Q_h^\pi(s,a;r)|.$$

The algorithm relies on an "upper confidence bound" $\bar{E}_h^t(s,a)$ for the error defined recursively as follows: $\bar{E}_{H+1}^t(s,a) = 0$ for all (s,a) and, for all $h \in [H]$, with the convention $1/0 = +\infty$,

$$\bar{E}_h^t(s,a) = \min\left(\gamma\sigma_{H-h},\, \gamma\sigma_{H-h}\sqrt{\frac{2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \gamma \sum_{s'} \hat{p}_h^t(s'|s,a) \max_b \bar{E}_{h+1}^t(s',b)\right). \quad (2)$$

Although $\bar{E}_h^t(s,a)$ does not depend on a policy π or a reward function r, Lemma 1, proved in Appendix B, shows that it is a high-probability upper bound an the error $\hat{e}_h^{t,\pi}(s,a;r)$ for *any* π *and* r.

Lemma 1. *With* $\mathrm{KL}(p||q) = \sum_{s \in \mathcal{S}} p(s) \log \frac{p(s)}{q(s)}$ *the Kullback-Leibler divergence between two distributions over* $\mathcal{S}$*, on the event*

$$\mathcal{E} = \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a), \mathrm{KL}\left(\hat{p}_h^t(\cdot|(s,a)), p_h(\cdot|(s,a))\right) \leq \tfrac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\right\},$$

it holds that for any policy π *and reward function* r, $\hat{e}_h^{t,\pi}(s,a;r) \leq \bar{E}_h^t(s,a)$.

Sampling rule and stopping rule The idea of RF-UCRL is to uniformly reduce the estimation error all policies under all possible reward functions by being greedy with respect to the upper bounds $\bar{E}^t$ on these errors. RF-UCRL stops when the error in step 1 is smaller than $\varepsilon/2$:

- **sampling rule**: the policy π^{t+1} is the greedy policy with respect to $\bar{E}^t(s,a)$, that is

$$\forall s \in \mathcal{S}, \forall h \in [h], \quad \pi_h^{t+1}(s) = \mathrm{argmax}_a \bar{E}_h^t(s,a).$$

- **stopping rule**: $\tau = \inf\left\{t \in \mathbb{N} : \bar{E}_h^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2\right\}$.

This algorithm is very similar to the one originally proposed by Fiechter for Best Policy Identification in the discounted case [11]. The main difference is that the original algorithm additionally uses some scaling and rounding: the index used are integers, defined as $\tilde{E}_h^t(s,a) = \lfloor \bar{E}_h^t(s,a)/\eta \rfloor$ for some parameter $\eta > 0$, and the algorithm stops when $\tilde{E}_h^t(s,a)$ is smaller than a slightly different threshold. The reason for this discretization is the use of a combinatorial argument in the sample complexity analysis, which says that every m time steps (with m that is a function of S, A, δ and ε), at least one of the indices must decrease. The other difference is that the term $\sigma_{H-h}\sqrt{2\beta(n_h^t(s,a),\delta)/n_h^t(s,a)}$ in (2) is replaced by $\sigma_{H-h}\sqrt{2\log(2SAH/\delta)}$, which we believe is not enough guarantee the corresponding index to be a high-probability upper bounds on $\hat{e}^{t,\pi}(s,a;r)$[3].

In the next section, we propose a different analysis for RF-UCRL compared to the original analysis of [11], which yields an improved sample complexity in the more general discounted episodic setting.

[3] there are some missing union bounds in the concentration argument given by [11]

4 Theoretical Guarantees for RF-UCRL

We show that RF-UCRL is (ε, δ)-PAC for reward-free exploration and provide a high-probability upper bound on its sample complexity.

4.1 Correctness and Sample Complexity

First, for every reward function r, one can easily show (see Appendix D.1) that for all $t \in \mathbb{N}^*$,

$$\left\{\forall \pi, \left|\hat{V}_1^{t,\pi}(s_1; r) - V_1^{\pi}(s_1; r)\right| \leq \varepsilon/2\right\} \subseteq \left\{V_1^{\star}(s_1; r) - V_1^{\hat{\pi}_{t,r}^{\star}}(s_1; r) \leq \varepsilon\right\}. \tag{3}$$

This property is already used in Lemma 3.6 of [17], where an extra planning error is allowed, whereas we assume that the optimal policy $\hat{\pi}_{t,r}^{\star}$ in $(\hat{P}^t, r)$ is computed exactly (with backward induction). Hence, a sufficient condition to prove the correctness of RF-UCRL is to establish that, when it stops, the estimation errors for all policies *and all reward functions* is smaller than $\varepsilon/2$. But the stopping rule of RF-UCRL is precisely designed to achieve this property.

Lemma 2 (correctness)**.** *On the event $\mathcal{E}$, for any reward function r, $V_1^{\star}(s_1) - V_1^{\hat{\pi}_{\tau,r}^{\star}}(s_1) \leq \varepsilon$.*

Proof. By definition of the stopping rule, $\bar{E}_1^{\tau}(s_1, \pi_1^{\tau+1}(s_1)) \leq \varepsilon/2$. As $\pi^{\tau+1}$ is the greedy policy w.r.t. $\bar{E}^{\tau}$, this implies that for all $a \in \mathcal{A}$, $\bar{E}_1^{\tau}(s_1, a) \leq \varepsilon/2$. Hence, by Lemma 1 on the event $\mathcal{E}$, for all policy π, all reward function r, and all action a, $\hat{e}_1^{\tau,\pi}(s_1, a; r) \leq \varepsilon/2$. In particular, for all π and r, $|\hat{V}_1^{\tau,\pi}(s_1; r) - V_1^{\pi}(s_1; r)| \leq \varepsilon/2$, and the conclusion follows from the implication (3). □

We now state our main results for RF-UCRL. We prove that for a well-chosen calibration of the threshold $\beta(n, \delta)$, the algorithm is (ε, δ)-PAC for reward-free exploration and we provide a high-probability upper bound on its sample complexity.

Theorem 1. RF-UCRL *using threshold* $\beta(n, \delta) = \log\big(2SAH/\delta\big) + (S-1)\log\big(e(1+n/(S-1))\big)$ *is (ε, δ)-PAC for reward-free exploration. Moreover, with probability $1-\delta$,*

$$\tau \leq \frac{\mathcal{C}_H SA}{\varepsilon^2}\left[\log\left(\frac{2SAH}{\delta}\right) + 2(S-1)\log\left(\frac{\mathcal{C}_H SA}{\varepsilon^2}\left(\log\left(\frac{2SAH}{\delta}\right) + (S-1)\left(\sqrt{e} + \sqrt{\frac{e}{S-1}}\right)\right)\right) + (S-1)\right]$$

where $\mathcal{C}_H = 144(1+\sqrt{2})^2\sigma_H^4$.

From Theorem 1, the number of episodes of exploration needed is of order

$$\tfrac{SAH^4}{\varepsilon^2}\log\left(\tfrac{2SAH}{\delta}\right) + \tfrac{S^2AH^4}{\varepsilon^2}\log\left(\tfrac{SAH^4}{\varepsilon^2}\log\left(\tfrac{2SAH}{\delta}\right)\right),$$

up to (absolute) multiplicative constants. As explained in Appendix F, for stationary transitions, we can further replace H^4 by H^3 in this bound. We now examine the scaling of this bound when ε goes to zero and when δ goes to zero. In a regime of small ε, our $\tilde{\mathcal{O}}\left(S^2AH^4/\varepsilon^2\right)$[4] bound improves the dependency in H compared to the one given by [17] from H^5 to H^4 (and to H^3 for stationary transitions). This new bound is matching the lower bound of [17] up to a factor H^2 (and a factor H for stationary transitions). Then, in a regime small δ, the sample complexity of RF-UCRL scales in $\mathcal{O}\left((SAH^4/\varepsilon^2)\log(1/\delta)\right)$, which greatly improves over the $\mathcal{O}((S^4AH^7/\varepsilon)\left(\log(1/\delta)\right)^3)$ scaling of the algorithm of [17]. Finally, we note that our result also improves over the original sample complexity bound given by [11], which is in $\tilde{\mathcal{O}}\left((S^2A/(1-\gamma)^7\varepsilon^3))\log\left(SA/((1-\gamma)\delta)\right)\right)$ in the discounted setting (for which $H \sim 1/(1-\gamma)$).

4.2 Proof of Theorem 1

We first introduce a few notation. We let $p_h^{\pi}(s, a)$ be the probability that the state action pair (s, a) is reached in the h-th step of a trajectory generated under the policy π, and we use the shorthand $p_h^t(s, a) = p_h^{\pi^t}(s, a)$. We introduce the *pseudo-counts* $\bar{n}_h^t(s, a) = \sum_{i=1}^t p_h^i(s, a)$ and define

$$\mathcal{E}^{\text{cnt}} = \left\{\forall t \in \mathbb{N}^\star, \forall h \in [H], \forall (s, a) \in \mathcal{S} \times \mathcal{A} :\ n_h^t(s, a) \geq \frac{1}{2}\bar{n}_h^t(s, a) - \beta^{\text{cnt}}(\delta)\right\},$$

[4] the notion $\tilde{O}$ is neglecting logarithmic factors in the variable in which the asymptotics is taken

where $\beta^{\text{cnt}}(\delta) = \log\big(2SAH/\delta\big)$. Recalling the event $\mathcal{E}$ defined in Lemma 1, we let $\mathcal{F} = \mathcal{E} \cap \mathcal{E}^{\text{cnt}}$. Lemma 5 in Appendix C shows that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta/2$ and $\mathbb{P}(\mathcal{E}^{\text{cnt}}) \geq 1 - \delta/2$, which yields $\mathbb{P}(\mathcal{F}) \geq 1 - \delta$. From Lemma 2, on the event $\mathcal{F}$, it holds that $V_1^\star(s_1) - V_1^{\hat{\pi}^\star_\tau, r}(s_1) \leq \varepsilon$ for all reward function r, which proves that RF-UCRL is (ε, δ)-PAC.

We now upper bound the sample complexity of RF-UCRL on the event $\mathcal{F}$, postponing the proof of some intermediate lemmas to Appendix D. The first step is to introduce an average upper bound on the error at step h under policy π^{t+1} defined as

$$q_h^t = \sum_{(s,a)} p_h^{t+1}(s,a)\bar{E}_h^t(s,a).$$

The following crucial lemma permits to relate the errors at step h to that at step $h+1$.

Lemma 3. *On the event $\mathcal{E}$, for all $h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\bar{E}_h^t(s,a) \leq 3\sigma_{H-h}\left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right] + \gamma \sum_{s' \in \mathcal{S}} p_h(s'|s,a)\bar{E}_{h+1}^t(s', \pi^{t+1}(s'))\,.$$

Thanks to Lemma 3, the average errors can in turn be related as follows:

$$\begin{aligned} q_h^t &\leq 3\sigma_{H-h}\sum_{(s,a)} p_h^{t+1}(s,a)\left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right] + \gamma \sum_{(s,a)}\sum_{(s',a')} p_h^{t+1}(s,a)p_h(s'|s,a)\mathbb{1}(a' = \pi^{t+1}(s'))\overline{E}_{h+1}^t(s',a') \\ &\leq 3\sigma_{H-h}\sum_{(s,a)} p_h^{t+1}(s,a)\left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right] + \gamma q_{h+1}^t\,. \end{aligned} \tag{4}$$

For $h = 1$, observe that $p_h^{t+1}(s_1,a)\bar{E}_h^t(s_1,a) = \bar{E}_h^t(s_1, \pi_1^{t+1}(s_1))\mathbb{1}\left(\pi_1^{t+1}(s_1) = a\right)$, as the policy is deterministic. Now, if $t < \tau$, $\bar{E}_h^t(s_1, \pi_1^{t+1}(s_1)) \geq \varepsilon/2$ by definition of the stopping rule, hence

$$q_1^t = \sum_a p_1^{t+1}(s_1,a)\bar{E}_h^t(s_1,a) \geq (\varepsilon/2)\sum_{a \in \mathcal{A}} \mathbb{1}\left(\pi_1^{t+1}(s_1) = a\right) = \varepsilon/2.$$

Using (4) to upper bound q_1^t yields $\varepsilon/2 \leq 3\sum_{h=1}^H \sum_{(s,a)} \gamma^{h-1}\sigma_{H-h} p_h^{t+1}(s,a)\left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right]$ for $t < \tau$ and summing these inequalities for $t \in \{0, \dots, T\}$ where $T < \tau$ gives

$$(T+1)\varepsilon \leq 6\sum_{h=1}^H \gamma^{h-1}\sigma_{H-h}\sum_{(s,a)}\sum_{t=0}^T p_h^{t+1}(s,a)\left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right].$$

The next step is to relate the counts to the pseudo-counts using the fact that the event $\mathcal{E}^{\text{cnt}}$ holds. The proof of this lemma is very close to the proof of Lemma 9 in [18].

Lemma 4. *On the event $\mathcal{E}^{cnt}$, $\forall h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\forall t \in \mathbb{N}^*,\ \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1 \leq 4\frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a) \vee 1}\,.$$

Using Lemma 4, one can write that, on the event $\mathcal{F}$, for $T < \tau$,

$$\begin{aligned} (T+1)\varepsilon &\leq 12\sum_{h=1}^H \gamma^{h-1}\sigma_{H-h}\sum_{(s,a)}\sum_{t=0}^T p_h^{t+1}(s,a)\sqrt{\frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a) \vee 1}} \\ &\leq 12\sqrt{\beta(T+1,\delta)}\sum_{h=1}^H \gamma^{h-1}\sigma_{H-h}\sum_{(s,a)}\sum_{t=0}^T \frac{\bar{n}_h^{t+1}(s,a) - \bar{n}_h^t(s,a)}{\sqrt{\bar{n}_h^t(s,a) \vee 1}}, \end{aligned}$$

where we have used that by definition of the pseudo-counts $p_h^{t+1}(s,a) = \bar{n}_h^{t+1}(s,a) - \bar{n}_h^t(s,a)$. Using Lemma 19 of [15] (recalled in Appendix G) to upper bound the sum in t yields

$$(T+1)\varepsilon \leq 12(1+\sqrt{2})\sqrt{\beta(T+1,\delta)}\sum_{h=1}^{H}\gamma^{h-1}\sigma_{H-h}\sum_{(s,a)}\sqrt{n_h^{T+1}(s,a)}$$
$$\leq 12(1+\sqrt{2})\sqrt{\beta(T+1,\delta)}\sum_{h=1}^{H}\gamma^{h-1}\sigma_{H-h}\sqrt{SA}\sqrt{\sum_{s,a}n_h^{T+1}(s,a)}.$$

As $\sum_{s,a} n_h^{T+1}(s,a) = T+1$, one obtains, using further that $\sigma_{H-h} \leq \sigma_H$,

$$\varepsilon\sqrt{T+1} \leq 12(1+\sqrt{2})\sqrt{SA}\left(\sigma_H\sum_{h=1}^{H}\gamma^{h-1}\right)\sqrt{\beta(T+1,\delta)}.$$

For T large enough, this inequality cannot hold, as the left hand side is in $\sqrt{T}$ while the right hand-side is logarithmic. Hence τ is finite and satisfies (applying the inequality to $T = \tau - 1$)

$$\tau \leq \frac{\mathcal{C}_H SA}{\varepsilon^2}\beta(\tau,\delta),$$

where $\mathcal{C}_H = 144(1+\sqrt{2})^2\sigma_H^4$. The conclusion follows from Lemma 9 stated in Appendix G.

5 From Reward-Free Exploration to Best Policy Identification

While originally proposed for solving the Best Policy Identification problem (see Definition 2), we proved that RF-UCRL is (ε,δ)-PAC for Reward Free Exploration. In particular, RF-UCRL is also (ε,δ)-PAC for BPI given some deterministic reward function r. In this section, we investigate the difference in complexity between BPI and RFE, trying to answer the following question: could an algorithm specifically designed for BPI have a smaller sample complexity than RF-UCRL?

A lower bound for BPI can be found in the work of Dann et al. [7] (although this work in focused on the design of PAC-MDP algorithms). This worse-case lower bound says that for any (ε,δ)-PAC algorithm for BPI there exists an MDP with stationary transitions for which $\mathbb{E}[\tau] = \Omega\left(((SAH^2)/\varepsilon^2)\log(c_2/(\delta+c_3))\right)$ for some constant c_2 and c_3. This lower bound directly translates to a lower bound for RFE, showing that for a small δ the sample complexity of RF-UCRL is optimal up to a factor H for stationary rewards, for *both* BPI and RFE.

To the best of our knowledge, the BPI problem has not been studied a lot since the work of Fiechter [11]. [10] propose (ε,δ)-PAC algorithms that stop and output a guess for the optimal policy (in all states) for discounted MDPs, but no upper bound on their sample complexity is given. Then, another avenue to get an (ε,δ)-PAC algorithm for BPI is to use a regret minimization algorithm: [16] suggests to run a regret minimization algorithm for some well chosen number of episode K and to let $\hat{\pi}$ be a policy chosen at random among the K policies used. Taking as a sub-routine the minimax algorithm UCB-VI [2] that has $\mathcal{O}(\sqrt{H^2SAK})$ regret for stationary rewards, with $K = \mathcal{O}(H^2SA/(\varepsilon^2\delta^2))$ this conversion yields a (ε,δ)-PAC for BPI. Its sample complexity has a bad scaling in δ, but is optimal for BPI when ε is small. In this regime, we see in Figure 1 that RFE and BPI have a different complexity.

Small δ	UB	LB
RF	$\frac{SAH^3}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	$\frac{SAH^2}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$ [7]
BPI	$\frac{SAH^3}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	$\frac{SAH^2}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$ [7]

Small ε	UB	LB
RF	$\frac{S^2AH^3}{\varepsilon^2}$	$\frac{S^2AH^2}{\varepsilon}$ [17]
BPI	$\frac{SAH^2}{\varepsilon^2}$ [2] + [16]	$\frac{SAH^2}{\varepsilon^2}$ [7]

Figure 1: Available upper and lower bounds on the sample complexity for RFE and BPI for stationary transition kernels ($p_h(s'|s,a) = p(s'|s,a)$) in a regime of small δ (left) and small ε (right)

In Appendix E, we propose an algorithm specifically designed for BPI, called BPI-UCRL which, unlike RF-UCRL, leverages the knowledge of a specific reward function r. Its sampling rule is close to that of the KL-UCRL algorithm [12] while its stopping rule also features *lower confidence*

bounds on the optimal Q value. We prove in Theorem 2 that BPI-UCRL enjoys the same sample complexity guarantees than RF-UCRL, both in the small δ and the small ε regime. As illustrated in the next section, BPI-UCRL appears to perform more efficient exploration than RF-UCRL for a fixed reward function. We leave as an open question whether an improved analysis for BPI-UCRL could corroborate this improvement (besides the slightly smaller constant $\mathcal{C}_H$ in Theorem 2).

6 Numerical Illustration

In this section we report the results of some experiments on synthetic MDPs aimed at illustrating how RF-UCRL and BPI-UCRL perform exploration, compared to simple baselines: (i) exploration with a random policy (RP) agent, and (ii) a generative model (GM) agent, which samples a fixed number of transitions from each state-action pair. We perform the following experiment: each algorithm interacts with the environment until it gathers a total of n transitions (exploration phase). RP, GM and RF-UCRL use the gathered data to estimate a model $\hat{P}$ and, at the end, they are given a reward function r and compute $\widehat{V}_1^\star(s_1; r)$ (estimation phase). Only the BPI-UCRL agent is allowed to observe the rewards during exploration phase, after which it outputs a policy $\widehat{\pi}^\star$. For RF-UCRL and BPI-UCRL we use the threshold $\beta(n, \delta)$ specified in Theorems 1 and 2 with $\delta = 0.1$. For RF-UCRL we found out that removing the minimum with $\gamma\sigma_{H-h}$ in the definition of the error bound (2) (which still gives a valid high-probability upper bound) leads to better pratical performance, and we report results for this variant. Our study does not include the parallel regret minimization approach of [17] described in the Introduction as this algorithm is mostly theoretical: the parameters N_0 and N that ensure the (ε, δ)-PAC property are only given up to non-specified mutliplicative constants.

We consider a DoubleChain MDP, with states $\mathcal{S} = \{0, \dots, L-1\}$, where L is the length of the chain, and actions $\mathcal{A} = \{0, 1\}$, which correspond to a transition to the left (action 0) or to the right (action 1). When taking an action, there is a 0.1 probability of moving to the other direction. A single reward of 1 is placed at the rightmost state $s = L - 1$, and the agent starts at $s_1 = (L-1)/2$, which leaves two possible directions for exploration.

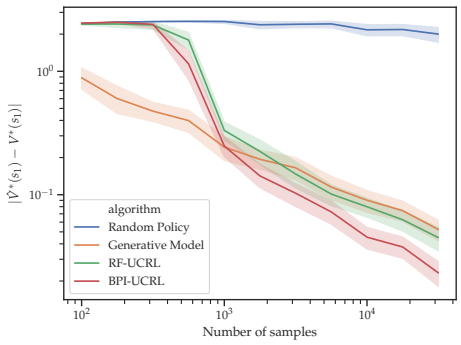
Figure 2a shows the estimation error $|\widehat{V}_1^\star(s_1; r) - V_1^\star(s_1; r)|$ as a function of n, estimated over $N = 48$ runs. As expected, this error decays the fastest under BPI-UCRL, since its exploration is guided by the observed rewards. Interestingly, the performance of RF-UCRL is close to the agent which has access to a generative model. Figure 2b shows the number of visits to each state during the exploration phase. We observe that the random policy is not able to reach the borders of the chain and, by design, the GM agent uniformly distributes its number of visits. RF-UCRL actively seeks to sample less visited states, and manage to fully explore the chain, whereas BPI-UCRL focuses its exploration on the part of the chain where the highest reward is placed. In Appendix A we report additional results of experiments in a GridWorld, where we observe the same behavior.

In Figure 2c and 2d we report estimates of the sample complexity τ of RF-UCRL and BPI-UCRL for different values of ε using $N = 48$ runs of $n = 10^8$ (resp. $n = 10^6$) sampled transitions, and checking the time needed for stopping at a level ε (if stopping occurs before n). BPI-UCRL has indeed a much smaller sample complexity than RF-UCRL.

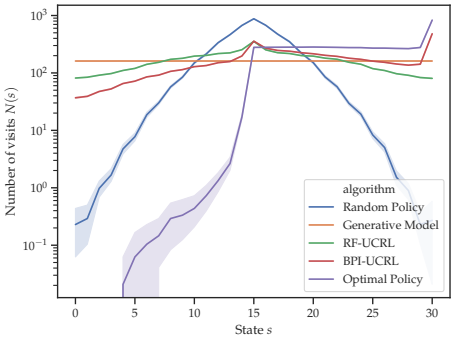
7 Conclusion

Inspired by the work of Fiechter from 1994, we proposed Reward-Free UCRL, a natural approach to Reward-Free Exploration. The improved sample complexity of this (ε, δ)-PAC algorithm is matching the existing lower bounds up to a factor H in both regimes of small ε and small δ. We also proposed BPI-UCRL for the related Best Policy Identification problem that is the initial focus of the work of Fiechter. Understanding the difference in complexity between RFE and BPI is an interesting open question, not fully solved in this paper. In particular, we will investigate in future work whether it is possible to design algorithms that are simultaneously optimal in the small δ and small ε regime for RFE or BPI. For BPI, we believe that one should look beyond the existing worse-case lower bound for problem dependent guarantees.

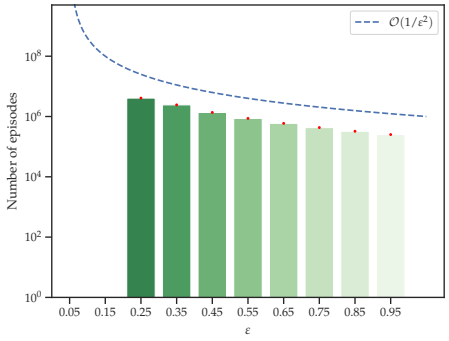
Acknowledgement We acknowledge the support of the European CHISTERA project DELTA. Anders Jonsson is partially supported by the Spanish grants TIN2015-67959 and PCIN-2017-082.



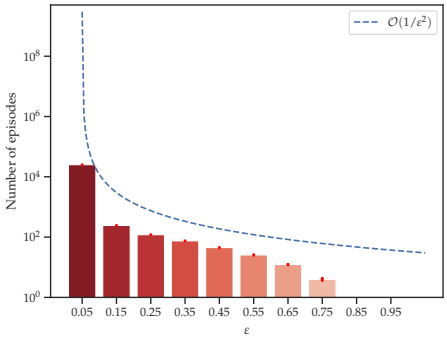
(a) Approximation error as a function of n



(b) Number of state visits for $n = 5000$



(c) Estimation of $\mathbb{E}[\tau|\tau < 10^8]$ for RF-UCRL



(d) Estimation of $\mathbb{E}[\tau|\tau < 10^6]$ for BPI-UCRL

Figure 2: RF-UCRL and BPI-UCRL on the DoubleChain MDP, with parameters $L = 31$, $H = 20$.

References

[1] Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[2] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, (ICML) 2017*, 2017.

[3] Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

[4] Nuttapong Chentanez, Andrew G. Barto, and Satinder P. Singh. Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1281–1288. MIT Press, 2005.

[5] Alon Cohen, Haim Kaplan, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret bounds for stochastic shortest path. *arXiv preprint arXiv:2002.09869*, 2020.

[6] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[7] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[8] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

[9] Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, pages 1902–1933, 2004.

[10] E. Even-Dar, S. Mannor, and Y. Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

[11] Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Proceedings of the Seventh Conference on Computational Learning Theory (COLT)*, 1994.

[12] S. Filippi, O. Cappé, and A. Garivier. Optimism in Reinforcement Learning and Kullback-Leibler Divergence. In *Allerton Conference on Communication, Control, and Computing*, 2010.

[13] Pratik Gajane, Ronald Ortner, Peter Auer, and Csaba Szepesvari. Autonomous exploration for navigating in non-stationary cmps, 2019.

[14] Elad Hazan, Sham M. Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. *arXiv preprint arXiv:1812.02690*, 2018.

[15] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[16] Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[17] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. *arXiv:2002.02794*, 2020.

[18] Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche-Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. *arXiv:2006.05879*, 2020.

[19] Sham Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.

[20] Michael J. Kearns and Satinder P. Singh. Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 996–1002, 1998.

[21] Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

[22] Shiau Hong Lim and Peter Auer. Autonomous exploration for navigating in mdps. In *Conference on Learning Theory*, pages 40–1, 2012.

[23] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2125–2133. Curran Associates, Inc., 2015.

[24] Guido Montufar, Keyan Ghazi-Zahedi, and Nihat Ay. Information theoretically aided reinforcement learning for embodied agents, 2016.

[25] Georg Ostrovski, Marc G Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2721–2730. JMLR. org, 2017.

[26] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

[27] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in biosciences = Theorie in den Biowissenschaften*, 131(3):139—148, September 2012.

[28] Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC model-free reinforcement learning. In *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML*, 2006.

[29] Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

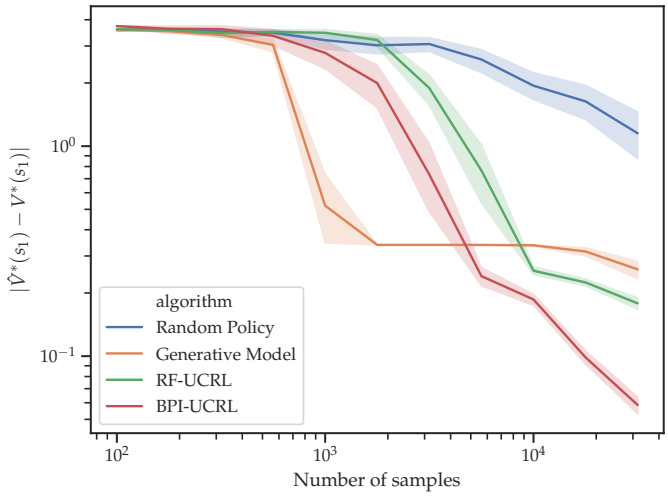
[30] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in neural information processing systems*, pages 2753–2762, 2017.

[31] Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. *arXiv preprint arXiv:1912.03517*, 2019.

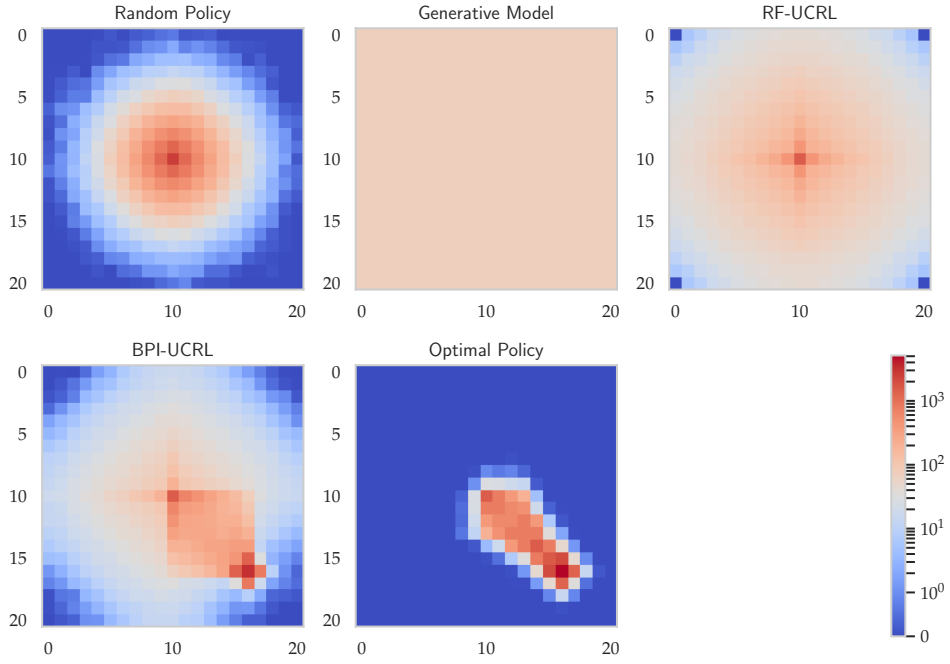
[32] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning, (ICML)*, 2019.

A Additional Experiments

Here, we consider a GridWorld environment, whose state space is a set of discrete points in a 21×21 grid. In each state, an agent can choose four actions: left, right, up or down, and it has a 5% probability of moving to the wrong direction. The reward is equal to 1 in state $(16, 16)$ and is 0 elsewhere. Figure 3a shows $\left|\widehat{V}_1^\star(s_1; r) - V_1^\star(s_1; r)\right|$ as a function of n and Figure 3b shows the number of visits to each state during the exploration phase. We observe the same behavior as explained for the DoubleChain: RF-UCRL seeks to sample from less visited states, whereas BPI-UCRL focuses its exploration near the rewarding state $(16, 16)$.



(a) Approximation error as a function of n



(b) Number of state visits for $n = 30000$

Figure 3: Comparison of several algorithms on the Gridworld MDP

B Proof of Lemma 1

From the Bellman equations in the empirical MDP and the true MDP,

$$\begin{aligned}\hat{Q}_h^{t,\pi}(s,a;r) &= r_h(s,a) + \gamma \sum_{s'} \hat{p}_h^t(s'|s,a) \hat{Q}_{h+1}^{t,\pi}(s',\pi(s');r) \\ \text{and } Q_h^{\pi}(s,a;r) &= r_h(s,a) + \gamma \sum_{s'} p_h(s'|s,a) Q_{h+1}^{\pi}(s',\pi(s');r)\,.\end{aligned}$$

Hence

$$\begin{aligned}\hat{Q}_h^{t,\pi}(s,a;r) - Q_h^{\pi}(s,a;r) &= \gamma \sum_{s'} \left(\hat{p}_h^t(s'|s,a) - p_h(s'|s,a)\right) Q_{h+1}^{\pi}(s',\pi(s');r) \\ &\quad + \gamma \sum_{s'} \hat{p}_h^t(s'|s,a) \left(\hat{Q}_{h+1}^{t,\pi}(s',\pi(s');r) - Q_{h+1}^{\pi}(s',\pi(s');r)\right).\end{aligned}$$

It follows that, for $n_h^t(s,a) > 0$, using successively that $Q_{h+1}^{\pi}(s',a';r) \le \sigma_{H-h}$, the definition of event $\mathcal{E}$ and the Pinsker's inequality,

$$\begin{aligned}\hat{e}_h^{t,\pi}(s,a;r) &\le \gamma \sum_{s'} \left|\hat{p}_h^t(s'|s,a) - p_h(s'|s,a)\right| Q_{h+1}^{\pi}(s',\pi(s');r) \\ &\qquad + \gamma \sum_{s'} \hat{p}_h^t(s'|s,a) \left|\hat{Q}_{h+1}^{t,\pi}(s',\pi(s');r) - Q_{h+1}^{\pi}(s',\pi(s');r)\right| \\ &\le \gamma\sigma_{H-h} \left\|\hat{p}_h^t(\cdot|s,a) - p_h(\cdot|s,a)\right\|_1 + \gamma \sum_{s'} \hat{p}_h^t(s'|s,a) \hat{e}_{h+1}^{t,\pi}(s',\pi(s');r) \\ &\le \gamma\sigma_{H-h} \sqrt{\frac{2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \gamma \sum_{s'} \hat{p}_h^t(s'|s,a) \hat{e}_{h+1}^{t,\pi}(s',\pi(s');r)\,.\end{aligned}$$

Then, noting that $\hat{e}_h^{t,\pi}(s,a;r) \le \gamma\sigma_{H-h}$, it holds for all $n_h^t(s,a) \ge 0$,

$$\hat{e}_h^{t,\pi}(s,a;r) \le \min\left(\gamma\sigma_{H-h}, \gamma\sigma_{H-h}\sqrt{\frac{2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \gamma \sum_{s'} \hat{p}_h^t(s'|s,a) \hat{e}_{h+1}^{t,\pi}(s',\pi(s');r)\right). \quad (5)$$

We can now prove the result by induction on h. The base case for $H+1$ is trivially true since $\hat{e}_{H+1}^{t,\pi}(s,a;r) = \bar{E}_{H+1}^t(s,a) = 0$ for all (s,a). Assume the result true for step $h+1$, using (5) we get for all (s,a),

$$\begin{aligned}\hat{e}_h^{t,\pi}(s,a;r) &\le \min\left(\gamma\sigma_{H-h},\, \gamma\sigma_{H-h}\sqrt{\frac{2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \gamma \sum_{s'} \hat{p}_h^t(s'|s,a) \hat{e}_{h+1}^{t,\pi}(s',\pi(s');r)\right) \\ &\le \min\left(\gamma\sigma_{H-h},\, \gamma\sigma_{H-h}\sqrt{\frac{2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \gamma \sum_{s'} \hat{p}_h^t(s'|s,a) \max_{b\in\mathcal{A}} \bar{E}_{h+1}^t(s',b)\right) \\ &= \bar{E}_{h+1}^t(s,a)\,.\end{aligned}$$

C High Probability Events

We recall that the event $\mathcal{F}$ introduced in the proof of Theorem 1 is the intersection of the two event

$$\begin{aligned}\mathcal{E} &= \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a), \mathrm{KL}\left(\hat{p}_h^t(\cdot|(s,a)), p_h(\cdot|(s,a))\right) \leq \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\right\} \\ \mathcal{E}^{\mathrm{cnt}} &= \left\{\forall t \in \mathbb{N}^\star, \forall h \in [H], \forall (s,a) \in \mathcal{S}\times\mathcal{A} :\ n_h^t(s,a) \geq \frac{1}{2}\bar{n}_h^t(s,a) - \log\left(\frac{2SAH}{\delta}\right)\right\}\end{aligned}$$

Using the the uniform concentration inequalities presented in the next sections and a union bound, we can prove the following. These concentration inequalities were already proved in [18], but we provide their proof in the next sections for the sake of completeness.

Lemma 5. *For $\beta(n,\delta) = \log\big(2SAH/\delta\big) + (S-1)\log\left(e(1+n/(S-1))\right)$, it holds that $\mathbb{P}\left(\mathcal{E}\right) \geq 1 - \frac{\delta}{2}$. Moreover, $\mathbb{P}\left(\mathcal{E}^{cnt}\right) \geq 1 - \frac{\delta}{2}$.*

Proof. Using Lemma 6 in Section C.1 yields

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}^c\right) &\leq \sum_{h=1}^H \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\left(\exists t \in \mathbb{N} : n_h^t(s,a)\,\mathrm{KL}\left(\hat{p}_h^t(\cdot|(s,a)), p_h(\cdot|(s,a))\right) \geq \beta(n_h^t(s,a),\delta)\right) \\ &\leq \sum_{h=1}^H \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\delta}{2SAH} = \frac{\delta}{2}.\end{aligned}$$

Then, using Lemma 8 in Section C.2 yields

$$\begin{aligned}\mathbb{P}\left(\left(\mathcal{E}^{\mathrm{cnt}}\right)^c\right) &\leq \sum_{h=1}^H \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\left(\exists t \in \mathbb{N} : n_h^t(s,a) \leq \frac{1}{2}\bar{n}_h^t(s,a) - \log\left(\frac{2SAH}{\delta}\right)\right) \\ &\leq \sum_{h=1}^H \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{i=1}^t \mathbb{1}\left((s_h^i, a_h^i) = (s,a)\right) \leq \frac{1}{2}\sum_{i=1}^t p_h^i(s,a) - \log\left(\frac{2SAH}{\delta}\right)\right) \\ &\leq \sum_{h=1}^H \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\delta}{2SAH} = \frac{\delta}{2}.\end{aligned}$$

□

C.1 Deviation Inequality for Categorical Distributions

Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. samples from a distribution supported over $\{1,\ldots,m\}$, of probabilities given by $p \in \Sigma_m$, where Σ_m is the probability simplex of dimension $m-1$. We denote by $\widehat{p}_n$ the empirical vector of probabilities, i.e. for all $k \in \{1,\ldots,m\}$

$$\widehat{p}_{n,k} = \frac{1}{n}\sum_{\ell=1}^n \mathbb{1}(X_\ell = k)\,.$$

Note that an element $p \in \Sigma_m$ will sometimes be seen as an element of $\mathbb{R}^{m-1}$ since $p_m = 1 - \sum_{k=1}^{m-1} p_k$. This should be clear from the context. We denote by $H(p)$ the (Shannon) entropy of $p \in \Sigma_m$,

$$H(p) = \sum_{k=1}^m p_k \log(1/p_k)\,.$$

Lemma 6. *For all $p \in \Sigma_m$, for all $\delta \in [0,1]$,*

$$\mathbb{P}\left(\exists n \in \mathbb{N}^*,\ n\,\mathrm{KL}(\widehat{p}_n, p) > \log(1/\delta) + (m-1)\log\left(e(1+n/(m-1))\right)\right) \leq \delta\,.$$

Proof. We apply the method of mixture with a Dirichlet prior on the mean parameter of the exponential family formed by the set of categorical distribution on $\{1, \dots, m\}$. Letting

$$\varphi_p(\lambda) = \log \mathbb{E}_{X\sim p}\left[e^{\lambda X}\right] = \log(p_m + \sum_{k=1}^{m-1} p_k e^{\lambda_k}),$$

be the log-partition function, the following quantity is a martingale:

$$M_n^\lambda = e^{n\langle \lambda, \widehat{p}_n\rangle - n\varphi_p(\lambda)}.$$

We set a Dirichlet prior $q \sim \mathcal{D}\mathrm{ir}(\alpha)$ with $\alpha \in {\mathbb{R}_+^*}^m$ and for $\lambda_q = (\nabla\varphi_p)^{-1}(q)$ and consider the integrated martingale

$$\begin{aligned}
M_n &= \int M_n^{\lambda_q} \frac{\Gamma\Big(\sum_{k=1}^m \alpha_k\Big)}{\prod_{k=1}^m \Gamma(\alpha_k)} q_k^{\alpha_k - 1}\,\mathrm{d}q \\
&= \int e^{n\left(\mathrm{KL}(\widehat{p}_n, p) - \mathrm{KL}(\widehat{p}_n, q)\right)} \frac{\Gamma\Big(\sum_{k=1}^m \alpha_k\Big)}{\prod_{k=1}^m \Gamma(\alpha_k)} q_k^{\alpha_k - 1}\,\mathrm{d}q \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} \int \frac{\Gamma\Big(\sum_{k=1}^m \alpha_k\Big)}{\prod_{k=1}^m \Gamma(\alpha_k)} q_k^{n\widehat{p}_{n,k} + \alpha_k - 1}\,\mathrm{d}q \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} \frac{\Gamma\Big(\sum_{k=1}^m \alpha_k\Big)}{\prod_{k=1}^m \Gamma(\alpha_k)} \frac{\prod_{k=1}^m \Gamma(\alpha_k + n\widehat{p}_{n,k})}{\Gamma\Big(\sum_{k=1}^m \alpha_k + n\Big)},
\end{aligned}$$

where in the second inequality we used Lemma 7. Now we choose the uniform prior $\alpha = (1, \dots, 1)$. Hence we get

$$\begin{aligned}
M_n &= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} (m-1)! \frac{\prod_{k=1}^m \Gamma(1 + n\widehat{p}_{n,k})}{\Gamma(m+n)} \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} (m-1)! \frac{\prod_{k=1}^m (n\widehat{p}_{n,k})!}{n!} \frac{n!}{(m+n-1)!} \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) + nH(\widehat{p}_n)} \frac{1}{\binom{n}{n\widehat{p}_n}} \frac{1}{\binom{m+n-1}{m-1}}.
\end{aligned}$$

Thanks to Theorem 11.1.3 by [6] we can upper bound the multinomial coefficient as follows: for $M \in \mathbb{N}^*$ and $x \in \{0, \dots, M\}^m$ such that $\sum_{k=1}^m x_k = M$ it holds

$$\binom{M}{x} = \frac{M!}{\prod_{k=1}^m x_k!} \le e^{MH(x/M)}.$$

Using this inequality we obtain

$$\begin{aligned}
M_n &\ge e^{n\,\mathrm{kl}(\widehat{p}_n, p) + nH(\widehat{p}_n) - nH(\widehat{p}_n) - (m+n-1)H\left((m-1)/(m+n-1)\right)} \\
&= e^{n\,\mathrm{KL}(\widehat{p}_n, p) - (m+n-1)H\left((m-1)/(m+n-1)\right)}.
\end{aligned}$$

It remains to upper-bound the entropic term

$$\begin{aligned}
(m+n-1)H\big((m-1)/(m+n-1)\big) &= (m-1)\log\frac{m+n-1}{m-1} + n\log\frac{m+n-1}{n} \\
&\le (m-1)\log\big(1 + n/(m-1)\big) + n\log(1 + (m-1)/n) \\
&\le (m-1)\log\big(1 + n/(m-1)\big) + (m-1).
\end{aligned}$$

Thus we can lower bound the martingale as follows

$$M_n \ge e^{n\,\mathrm{KL}(\widehat{p}_n, p)} \big(e(1 + n/(m-1))\big)^{m-1}.$$

Using the fact that, for any supermartingale it holds that

$$\mathbb{P}\left(\exists n \in \mathbb{N}^* : M_n > 1/\delta\right) \le \delta\mathbb{E}[M_1], \tag{6}$$

which is a well-known property used in the method of mixtures (see [9]), we conclude that

$$\mathbb{P}\left(\exists n \in \mathbb{N}^*,\, n\,\mathrm{KL}(\widehat{p}_n, p) > (m-1)\log\left(e(1+n/(m-1))\right) + \log(1/\delta)\right) \le \delta\,.$$

□

Lemma 7. *For $q, p \in \Sigma_m$ and $\lambda \in \mathbb{R}^{m-1}$,*

$$\langle \lambda, q\rangle - \varphi_p(\lambda) = \mathrm{KL}(q,p) - \mathrm{KL}(q, p^\lambda)\,,$$

where $\varphi_p(\lambda) = \log(p_m + \sum_{k=1}^{m-1} p_k e^{\lambda_k})$ and $p^\lambda = \nabla\varphi_{p_0}(\lambda)$.

Proof. There is a more general way than the ad hoc one below to prove the result. First note that

$$p_k^\lambda = \frac{p_k e^{\lambda_k}}{p_m + \sum_{\ell=1}^{m-1} p_\ell e^{\lambda_\ell}}\,,$$

which implies that

$$p_m + \sum_{k=1}^{m-1} p_k e^{\lambda_k} = \frac{p_m}{p_m^\lambda}, \qquad \lambda_k = \log\frac{p_k^\lambda}{p_k} + \log\frac{p_m}{p_m^\lambda}\,.$$

Therefore we get

$$\begin{aligned}
\langle \lambda, q\rangle - \varphi_p(\lambda) &= \sum_{k=1}^{m-1} q_k \log\left(\frac{p_k^\lambda}{p_k}\frac{p_m}{p_m^\lambda}\right) - \log\left(p_m + \sum_{k=1}^{m-1} p_k e^{\lambda_k}\right) \\
&= \sum_{k=1}^{m-1} q_k \log\frac{p_k^\lambda}{p_k} + (1-q_m)\log\frac{p_m}{p_m^\lambda} - \log\frac{p_m}{p_m^\lambda} \\
&= \sum_{k=1}^{m} q_k \log\frac{p_k^\lambda}{p_k} = \mathrm{KL}(q,p) - \mathrm{KL}(q,p^\lambda)\,.
\end{aligned}$$

□

C.2 Deviation Inequality for sequence of Bernoulli Random Variables

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of Bernoulli random variables adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. We restate here Lemma F.4. of [8].

Lemma 8. *If we denote $p_n = \mathbb{P}(X_n = 1|\mathcal{F}_{n-1})$, then for all $\delta \in (0,1]$*

$$\mathbb{P}\left(\exists n \in \mathbb{N}^* : \sum_{\ell=1}^{n} X_\ell < \sum_{\ell=1}^{n} p_\ell/2 - \log(1/\delta)\right) \le \delta\,.$$

D Proof of Auxiliary Lemmas for Theorem 1

D.1 From Values to Optimal Values

In this section, we prove the inclusion (3), namely that for all t

$$\left\{\forall \pi, \left|\hat{V}_1^{t,\pi}(s_1;r) - V_1^\pi(s_1;r)\right| \le \varepsilon/2\right\} \subseteq \left\{V_1^\star(s_1;r) - V_1^{\hat{\pi}_{t,r}^\star}(s_1;r) \le \varepsilon\right\}.$$

We denote by $\pi^\star$ the optimal policy in the MDP (P,r) and recall that $\hat{\pi}_{t,r}^\star$ is the optimal policy in the MDP $(\hat{P}_t, r)$. One can write

$$\begin{aligned}
V_1^\star(s_1;r) - V_1^{\hat{\pi}_{t,r}^\star}(s_1;r) \quad &= \quad V_1^{\pi^\star}(s_1;r) - \hat{V}_1^{t,\pi^\star}(s_1;r) + \underbrace{\hat{V}_1^{t,\pi^\star}(s_1;r) - \hat{V}_1^{t,\hat{\pi}_{t,r}^\star}(s_1;r)}_{\le 0} \\
&\quad + \hat{V}_1^{t,\hat{\pi}_{t,r}^\star}(s_1;r) - V_1^{\hat{\pi}_{t,r}^\star}(s_1;r).
\end{aligned}$$

The middle term is non-negative as $\hat{\pi}_{t,r}^\star$ is the optimal policy in the empirical MDP which yields

$$V_1^\star(s_1;r) - V_1^{\hat{\pi}_{t,r}^\star}(s_1;r) \quad \le \quad \left|V_1^{\pi^\star}(s_1;r) - \hat{V}_1^{t,\pi^\star}(s_1;r)\right| + \left|\hat{V}_1^{t,\hat{\pi}_{t,r}^\star}(s_1;r) - V_1^{\hat{\pi}_{t,r}^\star}(s_1;r)\right|$$

and easily yields the inclusion above.

D.2 Proof of Lemma 3

By definition of $\bar{E}_h^t(s,a)$ and the greedy policy π^{t+1}, if $n_h^t(s,a) > 0$,

$$\overline{E}_h^t(s,a) \quad \leq \quad \gamma\sigma_{H-h}\sqrt{\frac{2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \gamma \sum_{s'\in\mathcal{S}} \hat{p}_h^t(s'|s,a)\bar{E}_{h+1}^t(s',\pi^{t+1}(s'))\,. \tag{7}$$

From the definition of the event $\mathcal{E}$ and Pinsker's inequality, one can further upper bound

$$\begin{aligned}
\sum_{s'\in\mathcal{S}} \hat{p}_h^t(s'|s,a)\bar{E}_{h+1}^t(s',\pi^{t+1}(s')) &\leq \sum_{s'\in\mathcal{S}} p_h^t(s'|s,a)\bar{E}_{h+1}^t(s',\pi^{t+1}(s')) + \sum_{s'\in\mathcal{S}} \left(\hat{p}_h^t(s'|s,a) - p_h^t(s'|s,a)\right)\bar{E}_{h+1}^t(s',\pi^{t+1}(s')) \\
&\leq \sum_{s'\in\mathcal{S}} p_h^t(s'|s,a)\bar{E}_{h+1}^t(s',\pi^{t+1}(s')) + \|\hat{p}_h^t(\cdot|s,a) - p_h^t(\cdot|s,a)\|_1\sigma_{H-h} \\
&\leq \sigma_{H-h}\sqrt{2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \sum_{s'\in\mathcal{S}} p_h^t(s'|s,a)\bar{E}_{h+1}^t(s',\pi^{t+1}(s'))\,,
\end{aligned}$$

where we used that $\bar{E}_{h+1}^t(s',\pi^{t+1}(s')) \leq \gamma\sigma_{H-h-1} \leq \sigma_{H-h}$. Plugging this in (7), upper bounding γ by 1 and using $2\sqrt{2} \leq 3$ yields

$$\overline{E}_h^t(s,a) \leq 3\sigma_{H-h}\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \gamma \sum_{s'\in\mathcal{S}} p_h^t(s'|s,a)\bar{E}_{h+1}^t(s',\pi^{t+1}(s'))$$

The conclusion follows by noting that it also holds that

$$\overline{E}_h^t(s,a) \leq \gamma\sigma_{H-h} \leq 3\sigma_{H-h} \leq 3\sigma_{H-h} + \gamma \sum_{s'\in\mathcal{S}} p_h^t(s'|s,a)\bar{E}_{h+1}^t(s',\pi^{t+1}(s'))$$

and that the inequality is also true for $n_h^t(s,a) = 0$ with the convention $1/0 = +\infty$.

D.3 Proof of Lemma 4

As the event $\mathcal{E}^{\mathrm{cnt}}$ holds, we know that for all $t < \tau$,

$$n_\ell^t(s,a) \quad \geq \quad \frac{1}{2}\bar{n}_\ell^t(s,a) - \beta^{\mathrm{cnt}}(\delta).$$

We now distinguish two cases. First, if $\beta^{\mathrm{cnt}}(\delta) \leq \frac{1}{4}\bar{n}_\ell^t(s,a)$, then

$$\frac{\beta(n_\ell^t(s,a),\delta)}{n_\ell^t(s,a)} \wedge 1 \leq \frac{\beta(n_\ell^t(s,a),\delta)}{n_\ell^t(s,a)} \leq \frac{\beta\left(\frac{1}{4}\bar{n}_\ell^t(s,a),\delta\right)}{\frac{1}{4}\bar{n}_\ell^t(s,a)} \leq 4\frac{\beta\left(\bar{n}_\ell^t(s,a),\delta\right)}{\bar{n}_\ell^t(s,a)\vee 1},$$

where we use that $x \mapsto \beta(x,\delta)/x$ is non-increasing for $x \geq 1$, $x \mapsto \beta(x,\delta)$ is non-decreasing, and $\beta^{\mathrm{cnt}}(\delta) \geq 1$.

If $\beta^{\mathrm{cnt}}(\delta) > \frac{1}{4}\bar{n}_\ell^t(s,a)$, simple algebra shows that

$$\frac{\beta(n_\ell^t(s,a),\delta)}{n_\ell^t(s,a)} \wedge 1 \leq 1 < 4\frac{\beta^{\mathrm{cnt}}(\delta)}{\bar{n}_\ell^t(s,a)\vee 1} \leq 4\frac{\beta(\bar{n}_\ell^t(s,a),\delta)}{\bar{n}_\ell^t(s,a)\vee 1},$$

where we use that $1 \leq \beta^{\mathrm{cnt}}(\delta) \leq \beta(0,\delta)$ and $x \mapsto \beta(x,\delta)$ is non-decreasing.

In both cases, we have

$$\left[\frac{\beta(n_\ell^t(s,a),\delta)}{n_\ell^t(s,a)} \wedge 1\right] \leq 4\frac{\beta(\bar{n}_\ell^t(s,a),\delta)}{\bar{n}_\ell^t(s,a)\vee 1}.$$

E An Algorithm for Best Policy Identification

In this section, we describe the BPI-UCRL algorithm and analyze its sample complexity. BPI-UCRL aims at finding the optimal policy for a fixed reward function r, assumed to deterministic. Hence to ease the notation we drop the dependency in r in the value and Q-value functions.

Unlike RF-UCRL, who builds upper bound on the estimation errors, BPI-UCRL relies on confidence intervals on the Q-value function of a policy π. To define these confidence regions, we first introduce a confidence region on the transition probabilities

$$\mathcal{C}_h^t(s,a) = \left\{ p \in \Sigma_S : \mathrm{KL}\big(\widehat{p}_h^t(\cdot|s,a), p\big) \leq \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\}$$

and define, for each policy π the confidence regions after t episodes as

$$\begin{aligned}
\overline{Q}_h^{t,\pi}(s,a) &= (r_h + \gamma \max_{\overline{p}_h \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{V}_{h+1}^{t,\pi})(s,a) & \underline{Q}_h^{t,\pi}(s,a) &= (r_h + \gamma \min_{\underline{p}_h \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{V}_{h+1}^{t,\pi})(s,a) \\
\overline{V}_h^{t,\pi}(s) &= \pi \overline{Q}_h^{t,\pi}(s) & \underline{V}_h^{t,\pi}(s) &= \pi \underline{Q}_h^{t,\pi}(s) \\
\overline{V}_{H+1}^{t,\pi}(s) &= 0 & \underline{V}_{H+1}^{t,\pi}(s) &= 0 \\
\overline{p}_h^{t,\pi}(s,a) &\in \operatorname*{argmax}_{\overline{p} \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{V}_{h+1}^{t,\pi}(s,a) & \underline{p}_h^{t,\pi}(s,a) &\in \operatorname*{argmin}_{\underline{p} \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{V}_{h+1}^{t,\pi}(s,a)\,,
\end{aligned}$$

where we use the notation $p_h f(s,a) = \mathbb{E}_{s' \sim p_h(.|s,a)} f(s')$ for the expectation operator and $\pi g(s) = g\big(s, \pi(s)\big)$ for the application of a policy. We also define upper and lower confidence bounds on the optimal value and Q-value functions as

$$\begin{aligned}
\overline{Q}_h^{t}(s,a) &= (r_h + \gamma \max_{\overline{p}_h \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{V}_{h+1}^{t})(s,a) & \underline{Q}_h^{t}(s,a) &= (r_h + \gamma \min_{\underline{p}_h \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{V}_{h+1}^{t})(s,a) \\
\overline{V}_h^{t}(s) &= \max_a \overline{Q}_h^t(s,a) & \underline{V}_h^{t}(s) &= \max_a \underline{Q}_h^t(s,a) \\
\overline{V}_{H+1}^{t}(s) &= 0 & \underline{V}_{H+1}^{t}(s) &= 0 \\
\overline{p}_h^{t}(s,a) &\in \operatorname*{argmax}_{\overline{p} \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{V}_{h+1}^{t}(s,a) & \underline{p}_h^{t}(s,a) &\in \operatorname*{argmin}_{\underline{p} \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{V}_{h+1}^{t}(s,a) \\
\overline{\pi}_h^t(s,a) &\in \operatorname*{argmax}_a \overline{Q}_h^t(s,a) & \underline{\pi}_h^t(s,a) &\in \operatorname*{argmax}_a \underline{Q}_h^t(s,a)\,.
\end{aligned}$$

By definition of the event $\mathcal{E}$ in Lemma 1, note that for all h and (s,a) the true transition probability $p_h(\cdot|s,a)$ belongs to $\mathcal{C}_h^t(s,a)$ for all t, hence one can easily prove by induction that for all π, $\underline{Q}_h^{t,\pi}(s,a) \leq Q_h^\pi(s,a) \leq \overline{Q}_h^{t,\pi}(s,a)$ and $\underline{Q}_h^t(s,a) \leq Q_h^\star(s,a) \leq \overline{Q}_h^t(s,a)$.

BPI-UCRL We are now ready to present an algorithm for BPI based on the UCRL algorithm, named BPI-UCRL. It is defined by the three rules:

- **Sampling rule** The policy $\pi^{t+1} = \overline{\pi}^t$ acts greedily with respect to the upper-bounds on the optimal Q-value functions.
- **Stopping rule** $\tau = \inf\left\{ t \in \mathbb{N} : \overline{V}_1^t(s_1) - \underline{V}_1^t(s_1) \leq \varepsilon \right\}$.
- **Recommendation rule** The prediction $\widehat{\pi}_\tau = \underline{\pi}^\tau$ is the policy that acts greedily with respect to the lower-bounds on the optimal Q-value functions.

BPI-UCRL bears some similarities with the online algorithms proposed by [10] to identify the optimal policy in a discounted MDP. They also build (different) upper and lower confidence bounds on $Q^\star(s,a)$ for all (s,a) and output the greedy policy with respect to $\underline{Q}$. However the stopping rule waits until *for all* state s and all action a that has not been eliminated, $|\overline{Q}(s,a) - \underline{Q}(s,a)| < \varepsilon(1-\gamma)/2$, which may take a very long time if some states have a small probability to be reached. In our case, only the confidence interval on the optimal value in state s_1 needs to be small to trigger stopping. This different stopping rule is also due to a different objective: find an optimal policy in state s_1 (or when s_1 is drawn from some distribution P_0) as opposed to find an optimal policy in all states. In

our case, we are able to provide upper bound on the stopping rule, which are not given by [10], who analyze the sample complexity of a different algorithm that requires a generative model.

We now given an analysis of BPI-UCRL, which bears strong similarity with the proof of Theorem 1 and establishes similar sample complexity guarantees as those proved for RF-UCRL: a $\mathcal{O}\left((SAH^4/\varepsilon^2)\log(1/\delta)\right)$ bound in a regime of small δ and a $\mathcal{O}\left(S^2AH^4/\varepsilon^2\right)$ bound in a regime of small ε. For stationary transitions, the dependency in H in these bounds can be improved to H^3, following the same steps as in Appendix F.

Theorem 2. BPI-UCRL *using threshold*

$$\beta(n,\delta) = \log\left(2SAH/\delta\right) + (S-1)\log\left(e(1+n/(S-1))\right)$$

is (ε,δ)-correct for best policy identification. Moreover, with probability $1-\delta$, the number of trajectories τ collected satisfy

$$\tau \leq \frac{\mathcal{C}_H SA}{\varepsilon^2}\left[\log\left(\frac{2SAH}{\delta}\right) + 2(S-1)\log\left(\frac{\mathcal{C}_H SA}{\varepsilon^2}\left(\log\left(\frac{2SAH}{\delta}\right) + (S-1)\left(\sqrt{e} + \sqrt{\frac{e}{S-1}}\right)\right)\right) + (S-1)\right]$$

where $\mathcal{C}_H = 64(1+\sqrt{2})^2\sigma_H^4$.

Proof. The first part of the theorem is a direct consequence of the correctness of the confidence bounds. Indeed on the event $\mathcal{E}$ if the algorithm stops at time τ, then we know

$$V_1^{\pi^\tau}(s_1) \geq \underline{V}_1^{\tau,\pi^\tau}(s_1) = \underline{V}_1^{\tau}(s_1) \geq \overline{V}_1^{\tau}(s_1) - \varepsilon \geq V_1^\star(s_1) - \varepsilon\,.$$

The fact that the event $\mathcal{E}$ holds with probability at least $1-\delta$ (see Lemma 5) allows us to conclude that BPI-UCRL is (ε,δ)-correct.

The proof of upper bounds on the complexity is very close to a classical regret proof. Fix some $T < \tau$. Then we know that for all $t \leq T$ it holds

$$\varepsilon \leq \overline{V}_1^t(s_1) - \underline{V}_1^t(s_1) \leq \overline{V}_1^t(s_1) - \underline{V}_1^{t,\overline{\pi}^t}(s_1)\,.$$

On the event $\mathcal{F}$ for a state action (s,a), using the holder inequality, the fact that $\underline{V}_{h+1}^{t,\overline{\pi}^t}(s') \leq \sigma_{H-h}$ and $\overline{V}_{h+1}^t(s') \leq \sigma_{H-h}$ and the Pinsker's inequality we have

$$\begin{aligned}
\overline{Q}_h^t(s,a) - \underline{Q}_h^{t,\overline{\pi}^t}(s,a) &= \gamma(\overline{p}_h^t - p_h)\overline{V}_{h+1}^t(s,a) + \gamma(p_h - \underline{p}_h^{t,\overline{\pi}})\underline{V}_{h+1}^{t,\overline{\pi}^t}(s,a) + \gamma p_h(\overline{V}_{h+1}^t - \underline{V}_{h+1}^{t,\overline{\pi}^t})(s,a) \\
&\leq \|\overline{p}_h^t(\cdot|s,a) - p_h(\cdot|s,a)\|_1\sigma_{H-h} + \|p_h^t(\cdot|s,a) - \underline{p}_h^{t,\overline{\pi}^t}(\cdot|s,a)\|_1\sigma_{H-h} \\
&\qquad + \gamma p_h(\overline{V}_{h+1}^t - \underline{V}_{h+1}^{t,\overline{\pi}^t})(s,a) \\
&\leq 4\sigma_{H-h}\left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right] + \gamma p_h(\overline{V}_{h+1}^t - \underline{V}_{h+1}^{t,\overline{\pi}^t})(s,a)\,.
\end{aligned}$$

Thus, using that $\overline{\pi}^t = \pi^{t+1}$ and by definition that $\overline{V}_h^t(s) = \pi_h^{t+1}\overline{Q}_h^t(s)$ and $\underline{V}_h^{t,\overline{\pi}^t}(s) = \pi_h^{t+1}\overline{Q}_h^{t,\overline{\pi}^t}(s)$, we obtain a recursive formula for the difference of upper and lower bound on the value functions, which may be viewed as a counterpart of Lemma 3 in the proof of Theorem 1:

$$\overline{V}_h^t(s) - \underline{V}_h^{t,\overline{\pi}^t}(s) \leq 4\sigma_{H-h}\left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right] + \gamma p_h(\overline{V}_{h+1}^t - \underline{V}_{h+1}^{t,\overline{\pi}^t})(s,a)$$

Recalling that $p_h^t(s,a)$ denotes the probability that the state action pair (s,a) is visited at step h under the policy π^t used in the t-th episode, we can prove by induction with the previous formula that for all $t < \tau$,

$$\varepsilon \leq \overline{V}_1^t(s) - \underline{V}_1^{t,\overline{\pi}^t}(s) \leq 4\sum_{h=1}^H \gamma^{h-1}\sigma_{H-h}\sum_{s,a} p_h^{t+1}(s,a)\left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right]$$

Thus summing for all $0 \leq t \leq T < \tau$ leads to

$$(T+1)\varepsilon \leq \sum_{t=0}^T 4\sum_{h=1}^H \gamma^{h-1}\sigma_{H-h}\sum_{s,a} p_h^{t+1}(s,a)\left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right].$$

We can then conclude exactly as in the proof of Theorem 1, that is use Lemma 4 to relate the counts $n_h^t(s,a)$ to the pseudo-counts $\bar{n}_h^t(s,a)$ to upper bound the sample complexity on the event $\mathcal{F} = \mathcal{E} \cap \mathcal{E}^{\mathrm{cnt}}$, which holds with probability at least $1-\delta$. □

F Analysis in the stationary case

In the stationary case, the transition kernel doesn't depend on time, that is $p_h(s'|s,a) = p(s'|s,a)$ for all $h \in [H]$. In that case, the upper bounds used in the algorithms can take into account the number of visits to (s,a) in any step, $n^t(s,a) = \sum_{h\in[H]} n_h^t(s,a)$. More precisely, in that case, the $\bar{E}_h^t(s,a)$ are replaced by the tighter upper bounds

$$\tilde{E}_h^t(s,a) = \min\left\{\gamma\sigma_{H-h}; \gamma\sigma_{H-h}\sqrt{\frac{2\beta(n^t(s,a),\delta)}{n^t(s,a)}} + \gamma\sum_{s'}\hat{p}^t(s'|s,a)\max_b \tilde{E}_{h+1}^t(s',b)\right\}$$

where $\hat{p}^t(s'|s,a) = \frac{\sum_{i=1}^t\sum_{h=1}^H \mathbb{1}\left(s_h^i=s, a_h^i=a, s_{h+1}^i=s'\right)}{n^t(s,a)}$.

Letting $\tilde{\mathcal{E}}$ be the event

$$\tilde{\mathcal{E}} = \left(\forall t\in\mathbb{N}^*, \forall (s,a)\in\mathcal{S}\times\mathcal{A}, \|\hat{p}^t(\cdot|s,a) - p(\cdot|s,a)\|_1 \le \sqrt{\frac{2\beta(n^t(s,a),\delta)}{n^t(s,a)}}\right),$$

with Pinsker's inequality and Lemma 6, one can prove that $\mathbb{P}(\tilde{\mathcal{E}}) \ge 1-\delta/2$ for the choice $\beta(n,\delta) = \log\big(2SA/\delta\big) + (S-1)\log\big(e(1+n/(S-1))\big)$. RF-UCRL based on the alternative bounds $\tilde{E}_h^t(s,a)$ is correct on $\tilde{\mathcal{E}}$, and we can upper bound its sample complexity on the event $\tilde{\mathcal{E}}\cap\mathcal{E}^{\text{cnt}}$ following the same approach as before. Letting

$$\tilde{q}_h^t = \sum_{(s,a)} p_h^{t+1}(s,a)\tilde{E}_h^t(s,a),$$

one can establish a similar inductive relationship as that of Lemma 3 which yields

$$\tilde{q}_1^t \le 3\sum_{h=1}^H\sum_{(s,a)}\gamma^{h-1}\sigma_{H-h}p_h^{t+1}(s,a)\left[\sqrt{\frac{\beta(n^t(s,a),\delta)}{n^t(s,a)}}\wedge 1\right].$$

Hence, for every $T<\tau$, as $\tilde{q}_1^t \ge \varepsilon/2$ for all $t\le T$, one can write

$$\begin{aligned}
\varepsilon(T+1) &\le 6\sum_{t=0}^T\sum_{h=1}^H\sum_{(s,a)}\gamma^{h-1}\sigma_{H-h}p_h^{t+1}(s,a)\left[\sqrt{\frac{\beta(n^t(s,a),\delta)}{n^t(s,a)}}\wedge 1\right]\\
&\le 6\sigma_H\sum_{(s,a)}\sum_{t=0}^T\sum_{h=1}^H p_h^{t+1}(s,a)\left[\sqrt{\frac{\beta(n^t(s,a),\delta)}{n^t(s,a)}}\wedge 1\right]
\end{aligned}$$

Letting $\overline{n}^t(s,a) = \sum_{h=1}^H\sum_{i=1}^t p_h^i(s,a)$, observe that $\sum_{h=1}^H p_h^{t+1}(s,a) = \overline{n}^{t+1}(s,a) - \overline{n}^t(s,a)$ and a similar reasoning than that in the proof of Lemma 4 yields

$$\varepsilon(T+1) \le 12\sigma_H\sum_{(s,a)}\sum_{t=0}^T\left(\overline{n}^{t+1}(s,a)-\overline{n}^t(s,a)\right)\sqrt{\frac{\beta(\overline{n}^t(s,a),\delta)}{\overline{n}^t(s,a)\vee 1}}$$

Using Lemma 19 in [15] and the Cauchy-Schwarz inequality yields

$$\begin{aligned}
\varepsilon(T+1) &\le 12(1+\sqrt{2})\sigma_H\sum_{(s,a)}\sqrt{\overline{n}^{T+1}(s,a)}\sqrt{\beta(\overline{n}^{T+1}(s,a),\delta)}\\
&\le 12(1+\sqrt{2})\sigma_H\sqrt{\beta(H(T+1),\delta)}\sqrt{SA}\sqrt{\sum_{s,a}\overline{n}^{T+1}(s,a)}\\
&= 12(1+\sqrt{2})\sigma_H\sqrt{\beta(H(T+1),\delta)}\sqrt{SA}\sqrt{H(T+1)}.
\end{aligned}$$

Hence, τ is finite and satisfies

$$\tau \le \frac{\mathcal{C}SAH^3}{\varepsilon^2}\beta(H\tau,\delta)$$

for $\mathcal{C} = 144(1+\sqrt{2})^2$. By Lemma 9, it follows that

$$\tau \le \frac{\mathcal{C}SAH^3}{\varepsilon^2}\left[\log\left(\frac{2SA}{\delta}\right) + 2(S-1)\log\left(\frac{\mathcal{C}SAH^3\sqrt{H}}{\varepsilon^2}\left(\log\left(\frac{2SA}{\delta}\right) + (S-1)\left(\sqrt{e}+\sqrt{\frac{He}{S-1}}\right)\right)\right) + (S-1)\right].$$

G Technical Lemmas

Lemma 9. *Let $n \geq 1$ and $a, b, c, d > 0$. If $n\Delta^2 \leq a + b\log(c + dn)$ then*

$$n \leq \frac{1}{\Delta^2}\left[a + b\log\left(c + \frac{d}{\Delta^4}(a + b(\sqrt{c} + \sqrt{d}))^2\right)\right].$$

Proof. Since $\log(x) \leq \sqrt{x}$ and $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y > 0$, we have

$$\begin{aligned} & n\Delta^2 \leq a + b\sqrt{c + dn} \leq a + b\sqrt{c} + b\sqrt{d}\sqrt{n} \\ \implies & \sqrt{n}\Delta^2 \leq \frac{a + b\sqrt{c}}{\sqrt{n}} + b\sqrt{d} \leq a + b(\sqrt{c} + \sqrt{d}) \\ \implies & n \leq \frac{1}{\Delta^4}\left(a + b(\sqrt{c} + \sqrt{d})\right)^2. \end{aligned}$$

Hence,

$$\begin{aligned} & n\Delta^2 \leq a + b\log(c + dn) \\ \implies & n\Delta^2 \leq a + b\log(c + dn) \quad \text{and} \quad n \leq \frac{1}{\Delta^4}\left(a + b(\sqrt{c} + \sqrt{d})\right)^2 \\ \implies & n\Delta^2 \leq a + b\log\left(c + \frac{d}{\Delta^4}\left(a + b(\sqrt{c} + \sqrt{d})\right)^2\right). \end{aligned}$$

□

We recall for completeness Lemma 19 in [15] which is used in our sample complexity analysis.

Lemma 10 (Lemma 19 in [15])**.** *For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq Z_{k-1} = \max\left[1; \sum_{i=1}^{k-1} z_i\right]$*

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