Fast active learning for pure exploration in reinforcement learning

Pierre Ménard [1] **Omar Darwiche Domingues** [2] **Emilie Kaufmann** [2 3] **Anders Jonsson** [4] **Edouard Leurent** [2]
Michal Valko [2 3 5]

Abstract

Realistic environments often provide agents with very limited feedback. When the environment is initially unknown, the feedback, in the beginning, can be completely absent, and the agents may first choose to devote all their effort on *exploring efficiently.* The exploration remains a challenge while it has been addressed with many hand-tuned heuristics with different levels of generality on one side, and a few theoretically-backed exploration strategies on the other. Many of them are incarnated by *intrinsic motivation* and in particular *explorations bonuses*. A common choice is to use $1/\sqrt{n}$ bonus, where n is a number of times this particular state-action pair was visited. We show that, surprisingly, for a pure-exploration objective of *reward-free exploration*, bonuses that scale with $1/n$ bring faster learning rates, improving the known upper bounds with respect to the dependence on the horizon H. Furthermore, we show that with an improved analysis of the stopping time, we can improve by a factor H the sample complexity in the *best-policy identification* setting, which is another pure-exploration objective, where the environment provides rewards but the agent is not penalized for its behavior during the exploration phase.

1. Introduction

In reinforcement learning (RL), an agent learns how to act by interacting with an environment, which provides feedback in the form of reward signals. The agent's objective is to maximize the sum of rewards. In this work, we study how to explore efficiently. In particular we wish to compute near-optimal policies using the least possible amount of interactions with the environment (in the form of observed transitions). In general, we may be either interested in the performance of the agent during the learning phase or we may only care for the performance of the learned policy. In the first setting, we can measure the performance of the agent by its *cumulative regret* which is the difference between the total reward collected by an optimal policy and the total reward collected by the agent during the learning. Therefore, the agent is encouraged to *explore* new policies but also *exploit* its current knowledge (Bartlett & Tewari, 2009; Jaksch et al., 2010). Another performance measure related to the regret consists in counting the number of times during the learning that the value of the policy used by the agent is ε far from the optimal one. The minimization of this count is formalized in the PAC-MDP setting introduced by Kakade (2003), see also Dann & Brunskill (2015) and Dann et al. (2017). The second setting and our central focus in this paper is called *pure-exploration* where the agent is free to make mistakes during the learning and explore more vigorously (Fiechter, 1994; Kearns & Singh, 1998; Even-Dar et al., 2006). We provide results for two pure-exploration settings when the environment is an episodic *Markov decision process* (MDP): the *reward-free exploration* (RFE) and the *best-policy identification* (BPI).

Best-policy identification In BPI, an agent interacts with the MDP, observing *transitions* and *rewards*, to output an ε-optimal policy with probability at least $1-\delta$ (Fiechter, 1994). Most of the work on BPI assumes that the agent has access to a *generative model* (oracle, Kearns & Singh, 1998). Having an oracle access means that the agent can simulate a transition from any sate-action pair. In particular, Azar et al. (2013) show that the optimal rate of the sample complexity, defined in this case as the number n of oracle calls for getting an ε-optimal policy with probability at least $1-\delta$ is of order[1] $\widetilde{\mathcal{O}}\big(H^4SA\log(1/\delta)/\varepsilon^2\big)$ where S

[1]Otto von Guericke University [2]Inria [3]Université de Lille [4]Universitat Pompeu Fabra [5]DeepMind Paris. Correspondence to: Pierre Ménard <pierre.menard@ovgu.de>, Omar Darwiche Domingues <omar.darwiche-domingues@inria.fr>.

Proceedings of the 38^th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] Azar et al. (2012), express both the upper and the lower bounds in the **total number of calls to the generative model (instead of trajectories)** and prove them for the γ-discounted infinite horizon. They are of order $SA(1-\gamma)^{-3}\varepsilon^{-2}\log(1/\delta)$. We

is the size of the state space, A is the size of the action space, and H is the horizon (see Table 1 and also Agarwal et al., 2020, Sidford et al., 2018). The $\widetilde{\mathcal{O}}$ notation hides terms that are poly-log in H, S, A, ε, and $\log(1/\delta)$.

Even if the oracle access is reasonable in some situations (games, physics simulators, ...), we focus on the more challenging and practical setting where the agent has only access to a *forward model*, meaning that the agent can only sample *trajectories* from some predefined initial state. In this setting, the sample complexity τ is the number of trajectories that are necessary to output an ε-optimal policy with probability at least $1-\delta$ (which leads to $n = H\tau$ sampled transitions). A straightforward but indirect approach to BPI, suggested for example by (Jin et al., 2018), is to run a regret-minimizing algorithm (for instance, `UCBVI` of Azar et al., 2017) for a sufficiently large number K of episodes, and output a policy $\widehat{\pi}$ that is chosen uniformly at random among the K policies executed by the agent. Unfortunately, this indirect approach is sub-optimal with respect to the error probability δ. Indeed, the resulting sample complexity scales with $1/\delta^2$, instead of the expected $\log(1/\delta)$[2], as can be seen in Table 1.

Recently, Kaufmann et al. (2021) proposed `BPI-UCRL`, which adapts an episodic version of a `UCRL`-type algorithm (Jaksch et al., 2010) to best-policy identification. In essence, they replace the random choice of the predicted policy by a data-dependent choice. This algorithm enjoys the correct dependence on δ prescribed by the lower bound of Dann & Brunskill, 2015 (see also Domingues et al., 2021b), but suffers a sub-optimal dependence on S, the size of the state space, when ε is small, as well as a sub-optimal dependence on the horizon H (Table 1).

As an answer to the above sub-optimalities, we propose `BPI-UCBVI`, a new algorithm with a sample complexity of $\widetilde{\mathcal{O}}\big(SAH^3\log(1/\delta)/\varepsilon^2\big)$, which is optimal in terms of S, A, H, ε, and $\log(1/\delta)$ at the first order, according to the lower bound of Domingues et al. (2021b). `BPI-UCBVI` is based on `UCBVI` of Azar et al. (2017). It relies on a non-trivial upper bound on the *simple regret* of a `UCBVI`-type algorithm (Lemma 2), similar as Dann et al. (2019), that shaves the extra S factor of `RF-UCRL` while keeping the right dependence on δ. The main feature of this upper bound is that it can be computed in the empirical MDP and therefore is accessible to the agent.

Reward-free exploration Efficient exploration is especially difficult when the reward signals are sparse, as the agent needs to interact with the environment while receiving almost no feedback. To address such situations, we also study *reward-free exploration* introduced by Jin et al. (2020), where the interaction with the environment is split into two phases: (i) an *exploration* phase, in which the agent learns the transition model $\widehat{p}$ of the MDP by interacting with the environment for a given number of episodes (still with a forward model); and (ii) a *planning* phase, in which the agent receives a reward function r and computes the optimal policy for the MDP parameterized by $(r, \widehat{p})$. Given an accuracy parameter ε, we measure the performance of the agent by the number of trajectories required to compute a policy in the planning phase, that is ε-optimal *for any given reward function* r with probability at least $1-\delta$.

Our interest in RFE has two major reasons. First, in some applications, it is necessary to compute good policies for a wide range of reward functions. In such case, RFE allows to satisfy this need with only a single exploration phase. Second, RFE gives us good strategies for exploring the environment especially when the reward signal is very sparse or unknown.

One approach to pure exploration is to rely on known *cumulative-regret minimization* methods and their guarantees. This path is taken by `RF-RL-Explore` of Jin et al. (2020). More precisely, `RF-RL-Explore` builds upon the `EULER` algorithm by (Zanette & Brunskill, 2019) by running one instance of this algorithm for each state s and each episode step h with a reward function incentivizing the visit of state s in step h. The leading term in their sample complexity bound scales with $\widetilde{\mathcal{O}}\big(S^2AH^5\log(1/\delta)/\varepsilon^2\big)$ for MDPs with S states, A actions, and horizon H, which is sub-optimal in H (Table 1). Kaufmann et al. (2021) propose `RF-UCRL`, an alternative algorithm that is reminiscent of the original algorithm proposed by Fiechter (1994) for BPI with an improved sample complexity of $\widetilde{\mathcal{O}}\big(SAH^4(\log(1/\delta)+S)/\varepsilon^2\big)$. The main idea behind the algorithm of Fiechter (1994) is to build upper confidence bounds on the estimation error of the value function of *any* policy under *any* reward function, and then act greedily with respect to these upper bounds to minimize the estimation error. Using a similar approach, Wang et al. (2020) study reward-free exploration with a particular linear function approximation, providing an algorithm with a sample complexity of order $d^3H^6\log(1/\delta)/\varepsilon^2$, where d is the dimension of the feature space. Finally, Zhang et al. (2020) study a setting in which there are only N possible reward functions in the planning phase, for which they provide an algorithm whose sample complexity is $\widetilde{\mathcal{O}}\big(H^5SA\log(N)\log(1/\delta)/\varepsilon^2\big)$.

translate them to the episodic setting by replacing $(1-\gamma)^{-1}$ by the horizon H. In particular, the term $1/(1-\gamma)^3$ translates to H^3 and we include **an extra H factor in the upper bound** due to the non-stationary transitions, i.e., when the transition probabilities depend on the stage $h \in [H]$.

[2] See Appendix E for a discussion.

Algorithm	Setting	Upper bound	Lower bound
		(*non-stationary case*)	(*non-stationary case*)
Empirical QVI (Azar et al., 2012)[1]	gen. model	$\frac{H^4SA}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	$\frac{H^3SA}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$
UCBVI + random recomm.[2]	BPI	$\frac{H^3SA}{\varepsilon^2}\frac{\log(1/\delta)}{\delta^2}$	$\frac{H^3SA}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$
BPI-UCRL (Kaufmann et al., 2021)	BPI	$\frac{H^4SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)$	
BPI-UCBVI *(this work)*	BPI	$\frac{H^3SA}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	
UCBZero (Zhang et al., 2020)	RFE, N tasks	$\frac{H^5SA\log(N)}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	$\frac{H^2SA\log(N)}{\varepsilon^2}$
RF-RL-Explore (Jin et al., 2020)	RFE	$\frac{H^7S^2A}{\varepsilon}\log^3\left(\frac{1}{\delta}\right)+\frac{H^5S^2A}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	$\frac{SA}{\varepsilon^2}\left(H^3\log\left(\frac{1}{\delta}\right)+H^2S\right)$[3]
RF-UCRL (Kaufmann et al., 2021)	RFE	$\frac{H^4SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)$	
RF-Express *(this work)*	RFE	$\frac{H^3SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)$	

Table 1. Best-policy identification (BPI) and reward-free exploration (RFE) algorithms with their respective upper bounds on the sample complexity, expressed in terms of the number of trajectories required by the algorithms.[1] The factors and terms that are poly-log in S, A, H, ε, and $\log(1/\delta)$ are omitted.

In this work, we present RF-Express with sample complexity of $\widetilde{\mathcal{O}}\big(SAH^3(\log(1/\delta)+S)/\varepsilon^2\big)$, which improves the bound of Kaufmann et al. (2021) by a factor of H. In particular, up to poly-log terms, our rate matches the lower bound $\Omega(H^3SA\log(1/\delta)/\varepsilon^2)$ by Domingues et al. (2021b) when ε is fixed and δ goes to zero. Moreover we conjecture that the lower-bound by Jin et al. (2020), proved for the stationary setting (probability transition independent of h) becomes $\Omega\big(H^3SA\log(1/\delta)/\varepsilon^2\big)$ in our non-stationary setting (transition probabilities that could depend on the step h). Thus RF-Express would also match this lower-bound, effective when δ is fixed and ε goes to zero.[3]

A standard path to get such improved dependence is via confidence bonuses built using an *empirical Bernstein inequality* (Azar et al., 2012; 2017; Zanette & Brunskill, 2019) to make appear a variance term and then to sharply upper-bound these variance terms with a *Bellman type equation for the variances* (see Appendix F.1 or Azar et al., 2012). However, this standard path is far less clear for RFE as the agent does not observe the rewards and therefore cannot compute the empirical variance of the values! Therefore, one of our main technical contributions is to tackle this challenge by introducing a *new empirical Bernstein inequality* derived from a control of the transition probabilities (Appendix F.3) and applying the Bellman-type equation for the variances to construct exploration bonuses that do not require a computation of empirical variances. Surprisingly, the bonuses used in RF-Express scale with $1/n(s,a)$ where $n(s,a)$ is the number of times the state-action pair (s,a) was visited, instead of the usual $1/\sqrt{n(s,a)}$ bonus.

Contributions To sum up, we highlight our major contributions.

- BPI: we provide BPI-UCBVI, with a sample complexity of $\widetilde{\mathcal{O}}\big(H^3SA\log(1/\delta)\big)$ when ε is small enough. Up to poly-log terms, it matches the lower bound of Domingues et al. (2021b) and improves the dependence either on H, $1/\delta$ or S with respect to previous work.
- RFE: we provide RF-Express with a sample complexity of $\widetilde{\mathcal{O}}\big(H^3SA(\log(1/\delta)+S)/\varepsilon^2\big)$. Up to poly-log terms, our rate matches simultaneously the lower bound $\Omega(H^3SA\log(1/\delta)/\varepsilon^2)$ by Domingues et al. (2021b), effective when ε is fixed and δ goes to zero and, up to a factor H, the lower bound $\Omega(H^2S^2A/\varepsilon^2)$ by Jin et al. (2020), effective when δ is fixed and ε goes to zero.
- Due to the absence of the rewards in RFE, known techniques to get the optimal dependence in the horizon H (Azar et al., 2012; Zanette & Brunskill, 2019) do not apply. We therefore develop a new analysis that relies on the use of exploration bonuses scaling with $1/n$ instead of the standard $1/\sqrt{n}$.

2. Setting

We consider a finite episodic MDP $\big(\mathcal{S},\mathcal{A},H,\{p_h\}_{h\in[H]},\{r_h\}_{h\in[H]}\big)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, H is the number of steps in one episode, $p_h(s'|s,a)$ is the probability transition

[3] In Table 1, we combined the $\Omega\big(H^2S^2/\varepsilon^2\big)$ result of Jin et al. (2020) and the $\Omega\big(H^3SA\log(1/\delta)/\varepsilon^2\big)$ result of Domingues et al. (2021b). Note, that we can expect the lower bound by Jin et al. (2020) proved for the stationary setting (transition probabilities independent of h) to be $\Omega\big(H^3S^2/\varepsilon^2\big)$ in our non-stationary stetting (transition probabilities that could depend on h, see Section 2).

from state s to state s' by taking the action a at step h, and $r_h(s,a) \in [0,1]$ is the bounded deterministic reward received after taking the action a in state s at step h. Note that we consider the general case of rewards and transition functions that are possibly non-stationary, i.e., that are allowed to depend on the decision step h in the episode. We denote by S and A the number of states and actions, respectively.

Learning problem The agent, to which the transitions are *unknown*, interacts with the environment in episodes of length H, with a *fixed* initial state s_1.[4] At each step $h \in [H]$, the agent observes a state $s_h \in \mathcal{S}$, takes an action $a_h \in \mathcal{A}$ and makes a transition to a new state s_{h+1} according to the probability distribution $p_h(s_h, a_h)$. In BPI, the agent receives a deterministic reward $r_h(s_h, a_h)$ at each step h, for *fixed* reward functions $r \triangleq \{r_h\}_{h\in[H]}$, and it is required to output an ε-optimal policy *with respect to* r. In RFE, no rewards are observed during exploration, and the agent is required to output an estimate of the transition probabilities which can be used afterwards to compute an ε-optimal policy for *any reward function*.

Policy & value functions A *deterministic* policy π is a collection of functions $\pi_h : \mathcal{S} \mapsto \mathcal{A}$ for all $h \in [H]$, where every π_h maps each state to a *single* action. The value functions of π, denoted by V_h^π, are defined as

$$V_h^\pi(s) \triangleq \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \,\middle|\, s_h = s\right],$$

where $a_{h'} \triangleq \pi_{h'}(s_{h'})$ and $s_{h'+1} \sim p_{h'}(s_{h'}, a_{h'})$ for $h \in [H]$. The optimal value functions are defined as $V_h^\star(s) \triangleq \max_\pi V_h^\pi(s)$. Both V_h^π and $V_h^\star$ satisfy the Bellman equations (Puterman, 1994), that are expressed using the Q-value functions Q_h and $Q_h^\star$ in the following way,

$$V_h^\pi(s) = \pi_h Q_h^\pi(s), \quad Q_h^\pi(s,a) \triangleq r_h(s,a) + p_h V_{h+1}^\pi(s,a),$$
$$V_h^\star(s) = \max_a Q_h^\star(s,a), \quad Q_h^\star(s,a) \triangleq r_h(s,a) + p_h V_{h+1}^\star(s,a),$$

where by definition, $V_{H+1}^\star \triangleq V_{H+1}^\pi \triangleq 0$. Furthermore, $p_h f(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}[f(s')]$ denotes the expectation operator with respect to the transition probabilities p_h and $\pi_h g(s) \triangleq g(s, \pi_h(s))$ denotes the composition with the policy π at step h.

Empirical MDP Let $(s_h^i, a_h^i, s_{h+1}^i)$ be the state, the action, and the next state observed by an algorithm at step h of episode i. For any step $h \in [H]$ and any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we let $n_h^t(s,a) \triangleq \sum_{i=1}^t \mathbb{1}\big\{(s_h^i, a_h^i) = (s,a)\big\}$ be the number of times the state action-pair (s,a) was visited in step h in the first t episodes and $n_h^t(s,a,s') \triangleq \sum_{i=1}^t \mathbb{1}\big\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\big\}$. These definitions permit us to define the empirical transitions as

$$\widehat{p}_h^t(s'|s,a) \triangleq \frac{n_h^t(s,a,s')}{n_h^t(s,a)} \text{ if } n_h^t(s,a) > 0$$

and $\widehat{p}_h^t(s'|s,a) \triangleq 1/S$ otherwise. Based on the empirical transitions and on exploration bonuses, we introduce various data-dependent quantities that are useful for designing algorithms for either the BPI or the RFE objective. While the former allows the agent to access the reward function r during exploration, the latter does not. Therefore, in all data-dependent quantities introduced in Section 3 to design a RFE algorithm, we always materialize a possible dependency on r. In particular, we denote by $\widehat{V}_h^{t,\pi}(s; r)$ and $\widehat{Q}_h^{t,\pi}(s,a; r)$ the value and the action-value functions of a policy π in the MDP with transition kernels $\widehat{p}^t$ and reward function r. In Section 4, in which the reward function is fixed, we drop the dependency on r and use the simpler notation $\widehat{V}_h^{t,\pi}(s)$ and $\widehat{Q}_h^{t,\pi}(s,a)$.

3. Reward-free exploration

In this section, we consider reward-free exploration (RFE) where the agent *does not observe the rewards* during the exploration phase. Again, as the value functions defined in Section 2 depend on a reward function r, we sometimes use the notation $V_h(s; r)$ and $Q_h(s,a; r)$ instead of $V_h(s)$ and $Q_h(s,a)$.

Reward-free exploration In RFE , the agent interacts with the MDP in the following way. At the beginning of the episode t, the agent decides to follow a policy π^t, called the sampling rule, based only on the data collected up to episode $t-1$. Then, a *reward-free* episode $z^t \triangleq (s_1^t, a_1^t, s_2^t, a_2^t, \ldots, s_H^t, a_H^t)$ is generated starting from the the initial state $s_1^t \triangleq s_1$ by taking actions $a_h^t = \pi_h^t(s_h^t)$ and, for $h > 1$, observing next-states according to $s_h^t \sim p_h(s_{h-1}^t, a_{h-1}^t)$. This new trajectory is added to the dataset $\mathcal{D}_t \triangleq \mathcal{D}_{t-1} \cup \{z_t\}$. At the end of each episode, the agent can decide to stop collecting data, according to a *random* stopping time τ and outputs an empirical transition kernel $\widehat{p}$ built with the dataset $\mathcal{D}_\tau$.

Any RFE agent is therefore made of a triple $((\pi^t)_{t\in\mathbb{N}^\star}, \tau, \widehat{p})$. Our goal is to design an agent that is (ε, δ)-PAC, *probably approximately correct*, according to the following definition, for which the number of exploration episodes τ, i.e., the *sample complexity*, is as small as possible.

Definition 1 (PAC algorithm for RFE)**.** An algorithm is

[4] As explained by Fiechter (1994) and Kaufmann et al. (2021), if the first state is sampled randomly as $s_1 \sim p_0$, we can simply add an artificial first state s_0 such that for any action a, the transition probability is defined as the distribution $p_0(s_0, a) \triangleq p_0$.

(ε, δ)-PAC for reward-free exploration if

$$\mathbb{P}\Big(\text{for any reward function } r, V_1^\star(s_1;r) - V_1^{\widehat{\pi}_r^\star}(s_1;r) \le \varepsilon\Big) \ge 1-\delta,$$

where $\widehat{\pi}_r^\star$ is the optimal policy in the empirical MDP whose transitions are given by the transition kernel $\widehat{p}$ returned by the algorithm and whose reward function is r.

3.1. `RF-Express` algorithm

In this section, we present the `RF-Express` algorithm along with a high-probability bound on its sample complexity. `RF-Express` relies on upper bounds on the estimation error between the true value functions and their empirical counterparts. We start with the motivation for the design choices for `RF-Express`. We then introduce quantities to which we engrave our choices and which we subsequently use in the definition algorithm. In the algorithmic template we proceed as (Fiechter, 1994) and Kaufmann et al. (2021) by upper bounding the estimation-error for all the policies *with the striking difference that we only upper bound it at the initial state*. We finish this part by providing more intuition and discussion, in particular, we provide *technical insights into what* `RF-Express` *is optimizing* and then *explain our* $1/n$ *versus* $1/\sqrt{n}$ *exploration bonuses*, the reasons for choosing them and the challenge with analysing them.

Estimation error Given a policy π and an arbitrary reward function r, we define the estimation error as the absolute difference between the Q-value of π in the empirical MDP and its Q-value in the true MDP. Precisely, after episode t, for all (s, a, h), we define

$$\widehat{e}_h^{\,t,\pi}(s,a;r) \triangleq \big|\widehat{Q}_h^{t,\pi}(s,a;r) - Q_h^\pi(s,a;r)\big|.$$

To control the approximation error of the value of *any policy* for *any reward function* starting from the initial state s_1, we introduce the functions $W_h^t(s,a)$ defined inductively by $W_{H+1}^t(s,a) \triangleq 0$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and for all $h \in [H]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$W_h^t(s,a) \triangleq \min\Bigg(H, 15H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \left(1+\frac{1}{H}\right)\sum_{s'}\widehat{p}_h^{\,t}(s'|s,a)\max_{a'} W_{h+1}^t(s',a')\Bigg), \tag{1}$$

where $\beta(n_h^t(s,a),\delta)$ is a threshold that depends on how we build the confidence sets for the transitions probabilities. Notice that the W_h^t are all *independent* of the reward function r. As shown in the next lemma with proof in Appendix C, the function $W_1^t(s_1,a)$ can be used to upper bound the estimation error of any policy under any reward function in the initial state s_1.

Algorithm 1 `RF-Express`

```
**sampling rule:** the policy $\pi^{t+1}$ is the greedy policy with respect to $W_h^t$,
    $\forall s \in \mathcal{S}, \forall h \in [H],\ \ \pi_h^{t+1}(s) = \arg\max_{a\in\mathcal{A}} W_h^t(s,a)$
**stopping rule:**
    $\tau = \inf\Big\{t \in \mathbb{N} : 3e\sqrt{\pi_1^{t+1}W_1^t(s_1)} + \pi_1^{t+1}W_1^t(s_1) \le \varepsilon/2\Big\}$
**prediction rule:** output the empirical transition kernel $\widehat{p} = \widehat{p}^{\,\tau}$
```

Lemma 1. *With probability at least $1-\delta$, for any episode t, policy π, and reward function r,*

$$\widehat{e}_1^{\,t,\pi}(s_1,\pi_1(s_1);r) \le 3e\sqrt{\max_{a\in\mathcal{A}} W_1^t(s_1,a)} + \max_{a\in\mathcal{A}} W_1^t(s_1,a).$$

In particular, the above holds on the event $\mathcal{F}$ defined in Appendix A.

With all the above definitions, we are now ready to outline our `RF-Express` algorithm.

Next, we provide a bound on the sample complexity of `RF-Express` with a proof in Appendix C.

Theorem 1. *For $\delta \in (0,1)$, $\varepsilon \in (0,1]$,* `RF-Express` *with threshold $\beta(n,\delta) \triangleq \log(3SAH/\delta) + S\log(8e(n+1))$ is (ε,δ)-PAC for reward-free exploration. Moreover,* `RF-Express` *stops after τ episodes where, with probability at least $1-\delta$,*

$$\tau \le \frac{H^3SA}{\varepsilon^2}(\log(3SAH/\delta)+S)C_1 + 1$$

and where $C_1 \triangleq 5587e^6\log\big(e^{18}(\log(3SAH/\delta)+S)H^3SA/\varepsilon\big)^2$.

As a consequence, the sample complexity of `RF-Express` is of order $\widetilde{\mathcal{O}}\big(H^3SA(\log(1/\delta)+S)\big)$ and matches the lower bound of $\Omega(H^2S^2A/\varepsilon^2)$ of Jin et al. (2020) up to a factor of H and poly-log terms. This lower bound is informative in the regime where δ is considered as fixed and ε tends to zero. Moreover, our result also matches the lower bound of $\Omega\big(H^3SA\log(1/\delta)/\varepsilon^2\big)$ given by Domingues et al. (2021b) which is informative in the regime where ε is fixed and δ tends to 0. As explained in the introduction we believe that that the discrepancy with the dependence on the horizon is rather because of a sub-optimal lower bound. Indeed the lower-bound by Jin et al. (2020) is proved in the stationary setting and we conjecture it becomes $\Omega(H^3S^2A/\varepsilon^2)$ in our non-stationary setting (transition probabilities that could depend on h).

What is `RF-Express` optimizing? Contrary to `RF-UCRL` of Kaufmann et al. (2021), `RF-Express` does *not* build upper bounds on all estimation errors $\widehat{e}_h^{\,t,\pi}(s,a;r)$ for all $h \in [H]$ but *only for the one at the initial state* $\widehat{e}_1^{t,\pi}(s_1, \pi_1(s_1); r)$. Moreover, the upper bound is not $W_1^t(s_1, a)$ itself, but a function of this quantity, as can be seen in Lemma 1. Hence, if RF-Express actually follows the optimism-in-the-face-of-uncertainty principle, what quantities are W_h^t upper bounding? To answer this question and provide an intuition on the sampling rule of RF-Express, fix a policy π and let P^π be the probability distribution governing a random trajectory $(s_1, a_1, s_1, a_2, \ldots, s_H, a_H) \sim P^\pi$ in the MDP. Next, let $\widehat{P}^{t,\pi}$ be the probability distribution of a trajectory $(s_1^t, a_1^t, s_1^t, a_2^t, \ldots, s_H^t, a_H^t) \sim \widehat{P}^{t,\pi}$ in the empirical MDP built using the dataset $\mathcal{D}_t$ at episode t. Assuming that all the state action pairs have been visited at least once at time t, using the chain rule (see Garivier et al., 2019) we can compute the Kullback-Leibler divergence between these two probability distributions as

$$\mathrm{KL}(\widehat{P}^{t,\pi}, P^\pi) = \sum_{h=1}^{H} \sum_{s,a} \widehat{p}_h^{t,\pi}(s,a) \,\mathrm{KL}\big(\widehat{p}_h^{t,\pi}(s,a), p_h(s,a)\big),$$

where $\widehat{p}_h^{t,\pi}(s,a)$ is the probability to reach state-action (s,a) at step h under policy π in the empirical MDP in episode t. Notice now that the bonus of the form $\beta(n,\delta)/n$ used to define W^t is *by design chosen to be an upper-confidence bound* on the Kullback-Leibler divergence between the empirical transition probability and the transition probability. Indeed, in Appendix A we show that with high probability, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$,

$$\mathrm{KL}\big(\widehat{p}_h^t(s,a), p_h(s,a)\big) \le \frac{\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)}.$$

Therefore, omitting the clipping to H in (1), we have that

$$\max_\pi \mathrm{KL}(\widehat{P}^{t,\pi}, P^\pi) \lesssim \frac{\pi_1^{t+1} W_1^t(s_1)}{H^2}.$$

Therefore, RF-Express can be interpreted as an algorithm minimizing an upper-confidence bound on the loss of $\max_\pi \mathrm{KL}(\widehat{P}^{t,\pi}, P^\pi)$, which requires bonuses of the form $\beta(n,\delta)/n$ instead of $\sqrt{\beta(n,\delta)/n}$. Notice that this loss is of the same flavor as the one introduced by Hazan et al. (2019).

Bonuses of $1/n$ versus $1/\sqrt{n}$ Our approach differs from the bonuses typically used in regret minimization (e.g., Azar et al., 2017) and in prior work in reward-free exploration (Kaufmann et al., 2021; Zhang et al., 2020), which uses bonuses proportional to $\sqrt{1/n(s,a)}$. Intuitively, since $1/n$-bonuses decay faster with n, our algorithm is more exploratory: once a state-action pair (s,a) has been visited, the bonus associated to it will be more strongly reduced than if we used $\sqrt{1/n}$-bonuses and the algorithm tends to visit other state-action pairs before returning to (s,a) again. Empirically, we illustrate how $1/n$ bonuses can be beneficial for exploration in Appendix G. Technically, this might seem very surprising. Indeed, if we want to estimate the mean μ of a random variable X with an estimator $\widehat{\mu}_n$ computed with n i.i.d. samples from X, the error $|\mu - \widehat{\mu}_n|$ scales with $\sqrt{1/n}$ by Hoeffding's inequality, which explains the shape of the bonuses used in previous works. However, instead of bounding the error $|\mu - \widehat{\mu}_n|$, our concentration inequalities based on the KL divergence give us a bound on the quadratic term $(\mu - \widehat{\mu}_n)^2$, which scales with $1/n$. This allows us to use a Bellman-type equation for the variance of the value functions and reduce the sample complexity by a factor of H, similarly to previous work on regret minimization (Azar et al., 2017). The main challenge in our case is that, in reward free exploration, we need to upper bound the sum of variances for *any possible value function*, which makes this technique considerably more challenging to analyze than for the regret minimization.

3.2. Proof sketch

We first sketch the proof of Lemma 1. We begin as it is done in the analysis of Kaufmann et al. (2021). For a fixed policy π and an arbitrary reward function r, we decompose the estimation error of the Q-value function of π at the state-action pair (s,a) as, for all reward function r,

$$\begin{aligned}
\widehat{e}_h^{t,\pi}(s,a;r) &\le \big|\widehat{Q}_h^{t,\pi}(s,a;r) - Q_h^\pi(s,a;r)\big| \\
&\le \big|(\widehat{p}_h^t - p_h)V_{h+1}^\pi(s,a)\big| + \widehat{p}_h^t |\widehat{V}_{h+1}^{t,\pi} - V_{h+1}^\pi|(s,a) \\
&= \big|(\widehat{p}_h^t - p_h)V_{h+1}^\pi(s,a)\big| + \widehat{p}_h^t \pi_{h+1}^t \widehat{e}_{h+1}^{t,\pi}(s,a;r).
\end{aligned}$$

Similarly to Azar et al. (2017) and Zanette & Brunskill (2019), to obtain the optimal dependency with respect to the horizon H, we would like to apply the Bernstein inequality to control the first term. Since we need to do it for all value functions of all policies, we could use a covering of this function space and conclude with a union bound, see Domingues et al. (2020). Instead we show, via Lemma 10 in Appendix F.3, that from a control of the deviations of the empirical transition probabilities such that

$$\mathrm{KL}\big(\widehat{p}_h^t(s,a), p_h(s,a)\big) \le \frac{\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)}$$

with high probability, we deduce an empirical Bernstein inequality,

$$\begin{aligned}
\widehat{e}_h^{t,\pi}(s,a;r) \le 3\sqrt{\frac{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right)} \\
+ 15H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \left(1 + \frac{1}{H}\right)\widehat{p}_h^t \pi_{h+1}^t \widehat{e}_{h+1}^{t,\pi}(s,a;r),
\end{aligned}$$

where the variance of $\widehat{V}^{t,\pi}_{h+1}$, *in particular with respect to* $\widehat{p}^t_h(\cdot|s,a)$ is defined as

$$\mathrm{Var}_{\widehat{p}^t_h}(\widehat{V}^{t,\pi}_{h+1})(s,a;r) = \\ \sum_{s'} \widehat{p}^t_h(s'|s,a)\Big(\widehat{V}^{t,\pi}_{h+1}(s';r) - \mathbb{E}_{z\sim\widehat{p}^t_h(\cdot|s,a)}\Big[\widehat{V}^{t,\pi}_{h+1}(z;r)\Big]\Big)^2.$$

Therefore, defining $Z^{t,\pi}_{H+1}(s,a;r) \triangleq 0$ and recursively the functions

$$Z^{t,\pi}_h(s,a;r) \triangleq \min\Bigg(H, 3\sqrt{\frac{\mathrm{Var}_{\widehat{p}^t_h}(\widehat{V}^{t,\pi}_{h+1})(s,a;r)}{H^2}\left(\frac{H^2\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}\wedge 1\right)} \\ + 15H^2\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)} + \left(1+\frac{1}{H}\right)\widehat{p}^t_h\pi_{h+1}Z^{t,\pi}_{h+1}(s,a;r)\Bigg),$$

we prove by induction that for all (s,a,h),

$$\widehat{e}^{t,\pi}_h(s,a;r) \le Z^{t,\pi}_h(s,a;r). \tag{2}$$

We now *split* $Z^{t,\pi}$ in two terms. The first term is the one with the bonus in $\sqrt{1/n}$ and the second one with the bonus in $1/n$. Precisely, for all (s,a), we define recursively two other quantities $Y^{t,\pi}_{H+1}(s,a;r) \triangleq W^{t,\pi}_{H+1}(s,a) \triangleq 0$ and

$$Y^{t,\pi}_h(s,a;r) \triangleq 3\sqrt{\frac{\mathrm{Var}_{\widehat{p}^t_h}(\widehat{V}^{t,\pi}_{h+1})(s,a;r)}{H^2}\left(\frac{H^2\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}\wedge 1\right)} \\ + \left(1+\frac{1}{H}\right)\widehat{p}^t_h\pi_{h+1}Y^{t,\pi}_{h+1}(s,a;r)$$

$$W^{t,\pi}_h(s,a) \triangleq \min\left(H, 15H^2\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)} + \left(1+\frac{1}{H}\right)\widehat{p}^t_h\pi_{h+1}W^{t,\pi}_{h+1}(s,a)\right).$$

We then prove by induction (Appendix C, Step 2 of the proof of Lemma 1) that for all h,s,a,

$$Z^{t,\pi}_h(s,a;r) \le Y^{t,\pi}_h(s,a;r) + W^{t,\pi}_h(s,a). \tag{3}$$

Note that although $Z^{t,\pi}_h(\cdot;r)$ is a high-probability upper bound on $\widehat{e}^{t,\pi}_h(\cdot;r)$, we cannot use it to build a sampling rule reducing the errors as it still depends on the reward function r through the empirical variance term, and this knowledge is only available in the planning phase. To obtain an upper bound on $Z^{t,\pi}_h(\cdot;r)$ which does not depend on r, we now further upper-bound $Y^{t,\pi}(\cdot;r)$. The key tool for this purpose is to use the Bellman equation for the variances, see Appendix F.1. We denote by $\widehat{p}^{t,\pi}_h(s,a)$ the probability of reaching the state-action pair (s,a) at step h under the policy π in the empirical MDP at time t. Using Cauchy-Schwarz inequality, Lemma 7 in Appendix F.1, and the fact that that variance of the sum of reward is upper bounded by H^2, we get

$$\pi_1 Y^{t,\pi}_1(s_1;r) = \\ 2\sum_{s,a}\sum_{h=1}^{H}\widehat{p}^{t,\pi}_h(s,a)\left(1+\frac{1}{H}\right)^{h-1}\sqrt{\frac{\mathrm{Var}_{\widehat{p}^t_h}(\widehat{V}^{t,\pi}_{h+1})(s,a;r)}{H^2}B^t} \\ \le 3e\sqrt{\sum_{s,a}\sum_{h=1}^{H}\widehat{p}^{t,\pi}_h(s,a)\frac{\mathrm{Var}_{\widehat{p}^t_h}(\widehat{V}^{t,\pi}_{h+1})(s,a;r)}{H^2}}\sqrt{\sum_{s,a}\sum_{h=1}^{H}\widehat{p}^{t,\pi}_h(s,a)B^t} \\ \le 3e\sqrt{\sum_{s,a}\sum_{h=1}^{H}\widehat{p}^{t,\pi}_h(s,a)B^t} \le 3e\sqrt{W^{t,\pi}_1(s_1)},$$

where the last inequality is proved in Appendix C (Step 3 of the proof of Lemma 1) and we denoted

$$B^t = \frac{H^2\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)} \wedge 1\,.$$

Combining inequality $\pi_1 Y^{t,\pi}_1(s_1;r) \le 3e\sqrt{W^{t,\pi}_1(s_1)}$ with (2) and (3) yields that, for all π,

$$\widehat{e}^{t,\pi}_1(s_1,\pi_1(s_a);r) \le 3e\sqrt{\pi_1 W^{t,\pi}_1(s_1)} + \pi_1 W^{t,\pi}_1(s_1).$$

Finally, we note that by construction, $\pi_1 W^{t,\pi}_1(s_1) \le \max_{a\in\mathcal{A}} W^t_1(s_1,a)$, which allows us to conclude the proof of Lemma 1.

Next, we sketch the proof of Theorem 1. The fact that `RF-Express` is (ε,δ)-PAC is a simple consequence of Lemma 1. Indeed, on an event of probability at least $1-\delta$, if the algorithm stops at time τ we know that for all policy π and for all reward function r,

$$\frac{\varepsilon}{2} \ge 3e\sqrt{\max_{a\in\mathcal{A}} W^\tau_1(s_1,a)} + \max_{a\in\mathcal{A}} W^\tau_1(s_1,a) \\ \ge \widehat{e}^{\tau,\pi}_1(s_1,\pi_1(s_1);r) = |\widehat{V}^{\tau,\pi}_1(s_1;r) - V^\pi_1(s_1;r)|.$$

Therefore, still on the same event it holds that

$$V^\star_1(s_1;r) - V^{\widehat{\pi}^\star}_1(s_1;r) \le |V^{\pi^\star_r}_1(s_1;r) - \widehat{V}^{\tau,\pi^\star_r}_1(s_1;r)| \\ + |\widehat{V}^{\tau,\widehat{\pi}^\star_r}_1(s_1;r) - V^{\widehat{\pi}^\star_r}_1(s_1;r)| \le \varepsilon\,.$$

The proof of the bound on the sample complexity is close to the one of a regret bound. We fix a time $t<\tau$. We start by proving an upper-bound on $W^t_1(s_1,\pi^{t+1}(s_1))$. For that using again the empirical Bernstein inequality of Lemma 10, with high probability, it holds that

$$W^t_h(s,a) \le 21H^2\left(\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}\wedge 1\right) \\ + \left(1+\frac{3}{H}\right)p_h\pi^{t+1}_{h+1}W^t_{h+1}(s,a).$$

We denote by $p^t_h(s,a)$ the probability to reach the state-action pair (s,a) at step h under policy π^t in the true MDP. Unfolding the previous inequality and switching to the pseudo-counts, defined by $\bar{n}^t_h(s,a) \triangleq \sum_{\ell=1}^t p^\ell_h(s,a)$, by Lemma 8 proved in Appendix F.2 we get

$$\pi^{t+1}_1 W^t_1(s_1) \le 84e^3H^2\sum_{h=1}^{H}\sum_{s,a}p^{t+1}_h(s,a)\frac{\beta(\bar{n}^t_h(s,a),\delta)}{\bar{n}^t_h(s,a)\vee 1}. \tag{4}$$

Since $t<\tau$ we know that due to stopping rule

$$\varepsilon \le 3e\sqrt{\pi^{t+1}_1 W^t_1(s_1)} + \pi^{t+1}_1 W^t_1(s_1).$$

Summing the previous inequalities for $0 \leq t < \tau$ then using the Cauchy-Schwarz inequality we obtain

$$\tau\varepsilon \leq \sum_{t=0}^{\tau-1} 3e\sqrt{\pi_1^{t+1}W_1^t(s_1)} + \pi_1^{t+1}W_1^t(s_1)$$
$$\leq 3e\sqrt{\tau\sum_{t=0}^{\tau-1}\pi_1^{t+1}W_1^t(s_1)} + \sum_{t=0}^{\tau-1}\pi_1^{t+1}W_1^t(s_1).$$

We now upper-bound the sum that appears in the left-hand terms. Using successively (4), $\beta(\cdot,\delta)$ is increasing, Lemma 9 of Appendix F.2, we have

$$\sum_{t=0}^{\tau-1}\pi_1^{t+1}W_1^t(s_1) \leq 84e^3H^2\sum_{t=0}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}$$
$$\leq 84e^3H^2\beta(\tau-1,\delta)\sum_{h=1}^{H}\sum_{s,a}\sum_{t=0}^{\tau-1}\frac{\bar{n}_h^{t+1}(s,a)-\bar{n}_h^t(s,a)}{\bar{n}_h^t(s,a)\vee 1}$$
$$\leq 336e^3H^3SA\log(\tau+1)\beta(\tau-1,\delta).$$

Therefore, combining with the above inequality with the previous one, we get

$$\tau\varepsilon \leq 55e^3\sqrt{\tau H^3SA\log(\tau+1)\beta(\tau-1,\delta)}$$
$$+\, 336e^3H^3SA\log(\tau+1)\beta(\tau-1,\delta).$$

Using Lemma 13, we invert the inequality above and obtain an upper bound on τ, which allows us to conclude the proof of the theorem.

4. Best-policy identification

Unlike in the previous section, we now consider a more standard setup in which there is a single reward function r and in which the agent observes the reward at each step, during the exploration phase. To ease the presentation, we drop the dependence on the reward r in all data-dependent quantities introduced in this section.

Best-policy identification In BPI, the agent interacts with the MDP in a way described in Section 2. Notice that the difference from Section 3 is that the agent also observes the reward. In each episode t, the agent follows a policy π^t (the sampling rule) based only on the information collected up to and including episode $t-1$. At the end of each episode, the agent can decide to stop collecting data (we denote by τ its random stopping time) and outputs a guess $\widehat{\pi}$ for the optimal policy.

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

$$\mathbb{P}\Big(V_1^\star(s_1) - V_1^{\widehat{\pi}}(s_1) \leq \varepsilon\Big) \geq 1-\delta.$$

Algorithm 2 `BPI-UCBVI`

sampling rule: the policy π^{t+1} is the greedy policy with respect to $\widetilde{Q}_h^t$,

$$\forall s\in\mathcal{S}, \forall h\in[H],\;\; \pi_h^{t+1}(s) = \arg\max_{a\in\mathcal{A}}\widetilde{Q}_h^t(s,a)$$

stopping rule:

$$\tau = \inf\{t\in\mathbb{N} : \pi_1^{t+1}G_1^t(s_1) \leq \varepsilon\}$$

prediction rule: $\widehat{\pi} = \pi^{\tau+1}$

4.1. `BPI-UCBVI` algorithm

Similarly to Azar et al. (2017) and Zanette & Brunskill (2019), we define upper confidence bounds on the optimal Q-value and value functions as

$$\widetilde{Q}_h^t(s,a) \triangleq \min\Bigg(H, r_h(s,a) + 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}}$$
$$+14H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \utilde{V}_{h+1}^t)(s,a) + \widehat{p}_h^t\widetilde{V}_{h+1}^t(s,a)\Bigg),$$
$$\widetilde{V}_h^t(s) \triangleq \max_{a\in\mathcal{A}}\widetilde{Q}_h^t(s,a), \qquad \widetilde{V}_{H+1}^t(s) \triangleq 0,$$

where $\beta^\star$ is some exploration rate (that does *not* have a linear scaling in the number of states S, unlike β) and $\utilde{V}^t$ is a lower confidence bound on the optimal value function; see Appendix B for a complete definition.

As in RFE, we need to build an upper confidence bound on the gap $V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1)$, between the value of the optimal policy and the value of the current policy, to define the stopping rule. We recursively define the functions G^t as $G_{H+1}^t(s,a) \triangleq 0$ for all (s,a) and for all (s,a,h) with $h \leq H$ as

$$G_h^t(s,a) \triangleq \min\Bigg(H, 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}}$$
$$+36H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \left(1+\frac{3}{H}\right)\widehat{p}_h^t\pi_{h+1}^{t+1}G_{h+1}^t(s,a)\Bigg).$$

We prove the following result in Appendix D.

Lemma 2. *With probability at least* $1-\delta$*, for all t,*

$$V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1) \leq \pi_1^{t+1}G_1^t(s_1)\,.$$

In particular it holds on the event $\mathcal{G}$ *defined in Appendix A.*

We are now ready to define our `BPI-UCBVI` algorithm. We provide a sample complexity bound for `BPI-UCBVI` in the next theorem, which we prove in Appendix D.

Theorem 2. *For* $\delta\in(0,1)$, $\varepsilon\in(0,1/S^2]$, `BPI-UCBVI` *using thresholds* $\beta(n,\delta) \triangleq \log(3SAH/\delta) + S\log\big(8e(n+1)\big)$ *and* $\beta^\star(n,\delta) \triangleq \log(3SAH/\delta) + \log\big(8e(n+1)\big)$ *is* (ε,δ)*-PAC for best policy exploration. Moreover, with probability* $1-\delta$*,*

$$\tau \leq \frac{H^3SA}{\varepsilon^2}\big(\log(3SAH/\delta)+1\big)C_1 + 1,$$

where $C_1 \triangleq 5904e^{26}\log\big(e^{30}(\log(3SAH/\delta)+S)H^3SA/\varepsilon\big)^2$.

Therefore, the rate of `BPI-UCBVI` is of order $\widetilde{O}\big(H^3SA\log(1/\delta)/\varepsilon^2\big)$ when ε is small enough and matches the lower bound of $\Omega\big(H^3SA\log(1/\delta)/\varepsilon^2\big)$ by Domingues et al. (2021b) up to poly-log terms. To the best of our knowledge, `BPI-UCBVI` is the first algorithm for BPI whose sample complexity has an optimal dependence on S, A, ε, and δ.

Acknowledgements

The research presented was supported by European CHIST-ERA project DELTA, French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, French National Research Agency project BOLD (ANR19-CE23-0026-04). Anders Jonsson is partially supported by the Spanish grants PID2019-108141GB-I00 and PCIN-2017-082. Pierre Ménard is supported by the SFI Sachsen-Anhalt for the project RE-BCI ZS/2019/10/102024 by the Investitionsbank Sachsen-Anhalt.

References

Agarwal, Alekh, Kakade, Sham, and Yang, Lin F. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, 2020.

Azar, Mohammad Gheshlaghi, Munos, Rémi, and Kappen, Bert. On the sample complexity of reinforcement learning with a generative model. In *International Conference on Machine Learning*, 2012.

Azar, Mohammad Gheshlaghi, Munos, Rémi, and Kappen, Hilbert J. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Azar, Mohammad Gheshlaghi, Osband, Ian, and Munos, Rémi. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Bartlett, Peter L. and Tewari, Ambuj. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Uncertainty in Artificial Intelligence*, 2009.

Boucheron, Stéphane, Lugosi, Gábor, and Massart, Pascal. *Concentration inequalities*. Oxford University Press, 2013.

Cover, Thomas M. and Thomas, Joy A. *Elements of information theory*. John Wiley & Sons, 2006.

Dann, Christoph and Brunskill, Emma. Sample complexity of episodic fixed-horizon reinforcement learning. In *Neural Information Processing Systems*, 2015.

Dann, Christoph, Lattimore, Tor, and Brunskill, Emma. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Neural Information Processing Systems*, 2017.

Dann, Christoph, Li, Lihong, Wei, Wei, and Brunskill, Emma. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, 2019.

de la Peña, Victor H., Klass, Michael J., and Lai, Tze Leung. Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, 32:1902–1933, 2004.

Domingues, Omar Darwiche, Ménard, Pierre, Pirotta, Matteo, Kaufmann, Emilie, and Valko, Michal. Regret bounds for kernel-based reinforcement learning. *arXiv preprint arXiv:2004.05599*, 2020.

Domingues, Omar Darwiche, Flet-Berliac, Yannis, Leurent, Edouard, Ménard, Pierre, Shang, Xuedong, and Valko, Michal. rlberry - A Reinforcement Learning Library for Research and Education. `https://github.com/rlberry-py/rlberry`, 2021a.

Domingues, Omar Darwiche, Ménard, Pierre, Kaufmann, Emilie, and Valko, Michal. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, 2021b.

Durrett, Rick. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 4 edition, 2010.

Even-Dar, Eyal, Mannor, Shie, and Mansour, Yishay. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

Fiechter, Claude-Nicolas. Efficient reinforcement learning. In *Conference on Learning Theory*, 1994.

Garivier, Aurélien, Ménard, Pierre, and Stoltz, Gilles. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2): 377–399, 2019.

Hazan, Elad, Kakade, Sham, Singh, Karan, and Soest, Abby Van. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019.

Jaksch, Thomas, Ortner, Ronald, and Auer, Peter. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, 2010.

Jin, Chi, Allen-Zhu, Zeyuan, Bubeck, Sébastien, and Jordan, Michael I. Is Q-learning provably efficient? In *Neural Information Processing Systems*, 2018.

Jin, Chi, Krishnamurthy, Akshay, Simchowitz, Max, and Yu, Tiancheng. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 2020.

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

Kakade, Sham. *On the sample complexity of reinforcement learning*. PhD thesis, University College London, 2003.

Kaufmann, Emilie, Ménard, Pierre, Domingues, Omar Darwiche, Jonsson, Anders, Leurent, Edouard, and Valko, Michal. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, 2021.

Kearns, Michael J. and Singh, Satinder P. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Neural Information Processing Systems*, 1998.

Puterman, Martin L. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, New York, NY, 1994.

Sidford, Aaron, Wang, Mengdi, Wu, Xian, Yang, Lin F., and Ye, Yinyu. Near-optimal time and sample complexities for solving discounted Markov decision process with a generative model. In *Neural Information Processing Systems*, 2018.

Talebi, Mohammad Sadegh and Maillard, Odalric-Ambrym. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *Algorithmic Learning Theory*, 2018.

Wang, Ruosong, Du, Simon S, Yang, Lin F, and Salakhutdinov, Ruslan. On reward-free reinforcement learning with linear function approximation. In *Neural Information Processing Systems*, 2020.

Zanette, Andrea and Brunskill, Emma. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, 2019.

Zhang, Xuezhou, Ma, Yuzhe, and Singla, Adish. Task-agnostic exploration in reinforcement learning. In *Neural Information Processing Systems*, 2020.