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] $\widetilde{\mathcal{O}}\big(H^4 SA \log(1/\delta)/\varepsilon^2\big)$ 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.

[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.

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.

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 $\widetilde{\mathcal{O}}\big(H^5SA \log(N) \log(1/\delta)/\varepsilon^2\big)$.

Algorithm	Setting	Upper bound (*non-stationary case*)	Lower bound (*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	$\frac{H^3SA}{\varepsilon^2}\frac{\log(1/\delta)}{\delta^2}$	
`BPI-UCRL` (Kaufmann et al., 2021)	BPI	$\frac{H^4SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)$	$\frac{H^3SA}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$
`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	$\frac{H^7S^2A}{\varepsilon}\log^3\left(\frac{1}{\delta}\right)+\frac{H^5S^2A}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$	
`RF-UCRL` (Kaufmann et al., 2021)	RFE	$\frac{H^4SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)$	$\frac{SA}{\varepsilon^2}\left(H^3\log\left(\frac{1}{\delta}\right)+H^2S\right)$[3]
`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 $\widetilde{\mathcal{O}}\left(SAH^3(\log(1/\delta)+S)/\varepsilon^2\right)$, 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 $\Omega\left(H^3SA\log(1/\delta)/\varepsilon^2\right)$ 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.

[3] In Table 1, we combined the $\Omega\left(H^2S^2/\varepsilon^2\right)$ result of Jin et al. (2020) and the $\Omega\left(H^3SA\log(1/\delta)/\varepsilon^2\right)$ 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 $\Omega\left(H^3S^2/\varepsilon^2\right)$ in our non-stationary stetting (transition probabilities that could depend on h, see Section 2).

Contributions To sum up, we highlight our major contributions.

- BPI: we provide `BPI-UCBVI`, with a sample complexity of $\widetilde{\mathcal{O}}\left(H^3SA\log(1/\delta)\right)$ 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 $\widetilde{\mathcal{O}}\left(H^3SA(\log(1/\delta)+S)/\varepsilon^2\right)$. 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 $\left(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}\right)$, 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

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 $n_h^t(s,a) \triangleq \sum_{i=1}^t \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\}$ be the number of times the state action-pair (s,a) was visited in step h in the first t episodes and $n_h^t(s,a,s') \triangleq \sum_{i=1}^t \mathbb{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}$. 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 $(s,a) \in \mathcal{S}\times\mathcal{A}$,

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

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 π^{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 $(s_1, a_1, s_1, a_2, \dots, s_H, a_H) \sim P^\pi$ in the MDP. Next, let $\widehat{P}^{t,\pi}$ be the probability distribution of a trajectory $(s_1^t, a_1^t, s_1^t, a_2^t, \dots, s_H^t, a_H^t) \sim \widehat{P}^{t,\pi}$ 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 $\widehat{V}_{h+1}^{t,\pi}$, *in particular with respect to* $\widehat{p}_h^t(\cdot|s,a)$ is defined as

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

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

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

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

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

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 $Y_{H+1}^{t,\pi}(s,a;r) \triangleq W_{H+1}^{t,\pi}(s,a) \triangleq 0$ and

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

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

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

$$Z_h^{t,\pi}(s,a;r) \le Y_h^{t,\pi}(s,a;r) + W_h^{t,\pi}(s,a). \tag{3}$$

Note that although $Z_h^{t,\pi}(\cdot;r)$ is a high-probability upper bound on $\widehat{e}_h^{t,\pi}(\cdot;r)$, 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 $Z_h^{t,\pi}(\cdot;r)$ 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 $\widehat{p}_h^{t,\pi}(s,a)$ 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

$$\begin{aligned}
&\pi_1 Y_1^{t,\pi}(s_1;r) = \\
&2\sum_{s,a}\sum_{h=1}^H \widehat{p}_h^{t,\pi}(s,a)\left(1+\frac{1}{H}\right)^{h-1}\sqrt{\frac{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}B^t} \\
&\le 3e\sqrt{\sum_{s,a}\sum_{h=1}^H \widehat{p}_h^{t,\pi}(s,a)\frac{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}}\sqrt{\sum_{s,a}\sum_{h=1}^H \widehat{p}_h^{t,\pi}(s,a)B^t} \\
&\le 3e\sqrt{\sum_{s,a}\sum_{h=1}^H \widehat{p}_h^{t,\pi}(s,a)B^t} \le 3e\sqrt{W_1^{t,\pi}(s_1)},
\end{aligned}$$

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

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

Combining inequality $\pi_1 Y_1^{t,\pi}(s_1;r) \le 3e\sqrt{W_1^{t,\pi}(s_1)}$ with (2) and (3) yields that, for all π,

$$\widehat{e}_1^{t,\pi}(s_1,\pi_1(s_a);r) \le 3e\sqrt{\pi_1 W_1^{t,\pi}(s_1)} + \pi_1 W_1^{t,\pi}(s_1).$$

Finally, we note that by construction, $\pi_1 W_1^{t,\pi}(s_1) \le \max_{a\in\mathcal{A}} W_1^t(s_1,a)$, 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,

$$\begin{aligned}
\frac{\varepsilon}{2} &\ge 3e\sqrt{\max_{a\in\mathcal{A}} W_1^\tau(s_1,a)} + \max_{a\in\mathcal{A}} W_1^\tau(s_1,a) \\
&\ge \widehat{e}_1^{\tau,\pi}(s_1,\pi_1(s_1);r) = |\widehat{V}_1^{\tau,\pi}(s_1;r) - V_1^\pi(s_1;r)|.
\end{aligned}$$

Therefore, still on the same event it holds that

$$\begin{aligned}
V_1^\star(s_1;r) - V_1^{\widehat{\pi}^\star}(s_1;r) &\le |V_1^{\pi_r^\star}(s_1;r) - \widehat{V}_1^{\tau,\pi_r^\star}(s_1;r)| \\
&\quad + |\widehat{V}_1^{\tau,\widehat{\pi}_r^\star}(s_1;r) - V_1^{\widehat{\pi}_r^\star}(s_1;r)| \le \varepsilon\,.
\end{aligned}$$

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 $W_1^t(s_1,\pi^{t+1}(s_1))$. For that using again the empirical Bernstein inequality of Lemma 10, with high probability, it holds that

$$W_h^t(s,a) \le 21H^2\left(\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\wedge 1\right) + \left(1+\frac{3}{H}\right)p_h\pi_{h+1}^{t+1}W_{h+1}^t(s,a).$$

We denote by $p_h^t(s,a)$ 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 $\bar{n}_h^t(s,a) \triangleq \sum_{\ell=1}^t p_h^\ell(s,a)$, by Lemma 8 proved in Appendix F.2 we get

$$\pi_1^{t+1}W_1^t(s_1) \le 84e^3H^2\sum_{h=1}^H\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}. \tag{4}$$

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

$$\varepsilon \le 3e\sqrt{\pi_1^{t+1}W_1^t(s_1)} + \pi_1^{t+1}W_1^t(s_1).$$

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

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

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

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

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

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

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

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

Algorithm 2 `BPI-UCBVI`

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

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

stopping rule:

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

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

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

where $\beta^\star$ is some exploration rate (that does *not* have a linear scaling in the number of states S, unlike β) and $\underset{\sim}{V}^t$ 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 $h \le H$ as

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

We prove the following result in Appendix D.

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

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

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* $\beta(n,\delta) \triangleq \log\big(3SAH/\delta\big) + S\log\big(8e(n+1)\big)$ *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* $1-\delta$*,*

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

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 $\widetilde{\mathcal{O}}\big(H^3SA\log(1/\delta)/\varepsilon^2\big)$ 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.

Appendix

Table of Contents

A. Concentration events

We define the following favorable events: $\mathcal{E}$ the event where the empirical transition probabilities are close to the true ones, $\mathcal{E}^{\mathrm{cnt}}$ the event where the pseudo-counts are close to their expectation, and $\mathcal{E}^\star$ where the empirical means of the optimal value functions are close to the true ones,

$$\mathcal{E} \triangleq \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S}\times\mathcal{A}:\ \mathrm{KL}\big(\widehat{p}_h^t(\cdot|s,a), p_h(\cdot|s,a)\big) \le \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\right\},$$

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

$$\mathcal{E}^\star \triangleq \Bigg\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S}\times\mathcal{A}:$$

$$\left|(\widehat{p}_h^t - p_h)V_{h+1}^\star(s,a)\right| \le \min\left(H, \sqrt{2\mathrm{Var}_{p_h}(V_{h+1}^\star)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 3H\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}\right)\Bigg\}.$$

We also introduce the intersection of these events, $\mathcal{G} \triangleq \mathcal{E}\cap\mathcal{E}^{\mathrm{cnt}}\cap\mathcal{E}^\star$, and the intersection of only the first two events, $\mathcal{F} \triangleq \mathcal{E}\cap\mathcal{E}^{\mathrm{cnt}}$. Note that the event $\mathcal{F}$ is independent of the reward function r. We prove that for the right choice of the functions β the above events hold with high probability.

Lemma 3. *For the following choices of functions β,*

$$\beta(n,\delta) \triangleq \log(3SAH/\delta) + S\log(8e(n+1)),$$
$$\beta^{\mathrm{cnt}}(\delta) \triangleq \log(3SAH/\delta), \quad \textit{and}$$
$$\beta^\star(n,\delta) \triangleq \log(3SAH/\delta) + \log(8e(n+1)),$$

it holds that

$$\mathbb{P}(\mathcal{E}) \ge 1-\delta, \qquad \mathbb{P}(\mathcal{E}^{\mathrm{cnt}}) \ge 1-\delta, \qquad \textit{and} \qquad \mathbb{P}(\mathcal{E}^\star) \ge 1-\delta\,.$$

In particular, $\mathbb{P}(\mathcal{G}) \ge 1-\delta$ and $\mathbb{P}(\mathcal{F}) \ge 1-\delta$.

Proof. First, by Theorem 3, we have that

$$\mathbb{P}(\mathcal{E}) \ge 1 - \frac{\delta}{3}.$$

Second, by Theorem 4, we have that

$$\mathbb{P}(\mathcal{E}^{\mathrm{cnt}}) \ge 1 - \frac{\delta}{3}.$$

Finally, by Theorem 5, we have that

$$\mathbb{P}(\mathcal{E}^\star) \ge 1 - \frac{\delta}{3}.$$

Applying a union to the above three inequalities, we conclude that

$$\mathbb{P}(\mathcal{G}) \ge 1-\delta \qquad \mathbb{P}(\mathcal{E}) \ge 1-\delta\,.$$

Remark: Note that we can order $1 \le \beta^{\mathrm{cnt}}(\delta) \le \beta^\star(n,\delta) \le \beta(n,\delta)$. □

A.1. Deviation inequality for categorical distributions

Next, we reproduce the deviation inequality for categorical distributions by Jonsson et al. (2020, Proposition 1). Let $(X_t)_{t\in\mathbb{N}^\star}$ be i.i.d. samples from a distribution supported on $\{1,\dots,m\}$, 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 $k \in \{1,\dots,m\}$,

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

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

$$H(p) = \sum_{k=1}^{m} p_k \log\left(\frac{1}{p_k}\right).$$

Theorem 3. *For all $p \in \Sigma_m$ and for all $\delta \in [0,1]$,*

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

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

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

be the log-partition function, we have that the following M_n^λ is a martingale,

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

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

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

where in the second inequality we used Lemma 4. Next, we choose the uniform prior $\alpha = (1, \dots, 1)$ to obtain

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

Theorem 11.1.3 by Cover & Thomas (2006) shows us how to upper-bound the multinomial coefficient. In particular, for $M \in \mathbb{N}^\star$ and $x \in \{0, \dots, M\}^m$ such that $\sum_{k=1}^m x_k = M$,

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

Using this inequality we obtain

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

It remains to upper-bound entropic term which we do as follows,

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

Therefore, we lower bound the martingale as

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

Since for any supermartingale we have that

$$\mathbb{P}(\exists n \in \mathbb{N}^\star : M_n > 1/\delta) \leq \delta \cdot \mathbb{E}[M_1], \tag{5}$$

which is a well-known property of the method of mixtures (de la Peña et al., 2004), we conclude that

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

□

Lemma 4. *For $q, p \in \Sigma_m$ and $\lambda \in \mathbb{R}^{m-1}$,*

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

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

Proof. First note that

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

which implies that

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

Therefore, we obtain

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

□

A.2. Deviation inequality for sequence of Bernoulli random variables

Below, we reproduce the deviation inequality for Bernoulli distributions by Dann et al. (2017, Lemma F.4). Let $\mathcal{F}_t$ for $t \in \mathbb{N}$ be a filtration and $(X_t)_{t\in\mathbb{N}^\star}$ be a sequence of Bernoulli random variables with $\mathbb{P}(X_t = 1|\mathcal{F}_{t-1}) = P_t$ with P_t being $\mathcal{F}_{t-1}$-measurable and X_t being $\mathcal{F}_t$-measurable.

Theorem 4. *For all $\delta > 0$,*

$$\mathbb{P}\left(\exists n :\ \sum_{t=1}^{n} X_t < \sum_{t=1}^{n} P_t/2 - \log\frac{1}{\delta}\right) \leq \delta. \tag{6}$$

Proof. $P_t - X_t$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_t$. Since X_t is nonnegative and has finite second moment, we have for any $\lambda > 0$ that $\mathbb{E}\big[e^{-\lambda(X_t - P_t)}|\mathcal{F}_{t-1}\big] \leq e^{\lambda^2 P_t/2}$ (Exercise 2.9, Boucheron et al., 2013). Hence, we have

$$\mathbb{E}\Big[e^{\lambda(P_t - X_t) - \lambda^2 P_t/2}|\mathcal{F}_{t-1}\Big] \leq 1 \tag{7}$$

and by setting $\lambda \triangleq 1$, we see that

$$M_n = e^{\sum_{t=1}^n (-X_t + P_t/2)} \tag{8}$$

is a supermartingale. Therefore by Markov's inequality,

$$\mathbb{P}\left(\sum_{t=1}^n (-X_t + P_t/2) \geq \log\frac{1}{\delta}\right) = \mathbb{P}\left(M_n \geq \frac{1}{\delta}\right) \leq \delta\mathbb{E}[M_n] \leq \delta \tag{9}$$

which gives us

$$\mathbb{P}\left(\sum_{t=1}^n X_t \leq \sum_{t=1}^n P_t/2 - \log\frac{1}{\delta}\right) \leq \delta \tag{10}$$

for a fixed n. We define now the stopping time $\tau \triangleq \min\{t \in \mathbb{N} : M_t > \frac{1}{\delta}\}$ and the sequence $\tau_n = \min\{t \in \mathbb{N} : M_t > \frac{1}{\delta} \vee t \geq n\}$. Applying the convergence theorem for nonnegative supermartingales (Theorem 5.2.9 by Durrett, 2010), we get that $\lim_{t\to\infty} M_t$ is well-defined almost surely. Therefore, M_τ is well-defined even when $\tau = \infty$. By the optional stopping theorem for nonnegative supermartingales (Theorem 5.7.6 by Durrett, 2010), we have $\mathbb{E}[M_{\tau_n}] \leq \mathbb{E}[M_0] \leq 1$ for all n and applying Fatou's lemma, we obtain $\mathbb{E}[M_\tau] = \mathbb{E}[\lim_{n\to\infty} M_{\tau_n}] \leq \liminf_{n\to\infty} \mathbb{E}[M_{\tau_n}] \leq 1$. Using Markov's inequality, we can finally bound

$$\mathbb{P}\left(\exists n : \sum_{t=1}^n X_t < \frac{1}{2}\sum_{t=1}^n P_t - \log\frac{1}{\delta}\right) \leq \mathbb{P}(\tau < \infty) \leq \mathbb{P}\left(M_\tau > \frac{1}{\delta}\right) \leq \delta\mathbb{E}[M_\tau] \leq \delta. \tag{11}$$

□

A.3. Deviation inequality for bounded distributions

Below, we reproduce the self-normalized Bernstein-type inequality by Domingues et al. (2020). Let $(Y_t)_{t\in\mathbb{N}^\star}$, $(w_t)_{t\in\mathbb{N}^\star}$ be two sequences of random variables adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. We assume that the weights are in the unit interval $w_t \in [0,1]$ and predictable, i.e. $\mathcal{F}_{t-1}$ measurable. We also assume that the random variables Y_t are bounded $|Y_t| \leq b$ and centered $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$. Consider the following quantities

$$S_t \triangleq \sum_{s=1}^t w_s Y_s, \quad V_t \triangleq \sum_{s=1}^t w_s^2 \cdot \mathbb{E}\big[Y_s^2|\mathcal{F}_{s-1}\big], \quad \text{and} \quad W_t \triangleq \sum_{s=1}^t w_s$$

and let $h(x) \triangleq (x+1)\log(x+1) - x$ be the Cramér transform of a Poisson distribution of parameter 1.

Theorem 5 (Bernstein-type concentration inequality)**.** *For all $\delta > 0$,*

$$\mathbb{P}\left(\exists t \geq 1, (V_t/b^2 + 1)h\left(\frac{b|S_t|}{V_t + b^2}\right) \geq \log(1/\delta) + \log(4e(2t+1))\right) \leq \delta.$$

The previous inequality can be weakened to obtain a more explicit bound: with probability at least $1-\delta$, for all $t \geq 1$,

$$|S_t| \leq \sqrt{2V_t \log(4e(2t+1)/\delta)} + 3b\log(4e(2t+1)/\delta)\,.$$

Proof. By homogeneity we can assume that $b = 1$ to prove the first part. First note that for all $\lambda > 0$,

$$e^{\lambda w_t Y_t} - \lambda w_t Y_t - 1 \leq (w_t Y_t)^2(e^\lambda - \lambda - 1)\,,$$

because the function $y \to (e^y - y - 1)/y^2$ (extended by continuity at zero) is non-decreasing. Taking the expectation yields

$$\mathbb{E}\big[e^{\lambda w_t Y_t}|\mathcal{F}_{t-1}\big] - 1 \leq w_t^2 \mathbb{E}\big[Y_t^2|\mathcal{F}_{t-1}\big](e^\lambda - \lambda - 1)\,,$$

therefore using $y + 1 \leq e^y$ we get

$$\mathbb{E}\Big[e^{\lambda(w_t Y_t)}|\mathcal{F}_{t-1}\Big] \leq e^{w_t^2 \mathbb{E}\left[Y_t^2|\mathcal{F}_{t-1}\right](e^\lambda - \lambda - 1)}\,.$$

We just proved that the following quantity is a supermartingale with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$,

$$M_t^{\lambda,+} = e^{\lambda(S_t + V_t) - V_t(e^\lambda - 1)}\,.$$

Similarly, using that the same inequality holds for $-X_t$, we have

$$\mathbb{E}\big[e^{-\lambda w_t Y_t}|\mathcal{F}_{n-1}\big] \leq e^{w_t^2 \mathbb{E}\left[Y_t^2|\mathcal{F}_{t-1}\right](e^\lambda - \lambda - 1)}\,,$$

therefore, we can also define the supermartingale

$$M_t^{\lambda,-} = e^{\lambda(-S_t + V_t) - V_t(e^\lambda - 1)}\,.$$

We now choose the prior over $\lambda_x = \log(x+1)$ with $x \sim \mathcal{E}(1)$, and consider the (mixture) supermartingale

$$M_t = \frac{1}{2}\int_0^{+\infty} e^{\lambda_x(S_t+V_t) - V_t(e_x^\lambda - 1)} e^{-x}\mathrm{d}x + \frac{1}{2}\int_0^{+\infty} e^{\lambda_x(-S_t+V_t) - V_n(e_x^\lambda - 1)} e^{-x}\mathrm{d}x\,.$$

Note that by construction it holds $\mathbb{E}[M_t] \leq 1$. We now apply the method of mixtures to that super martingale therefore we need to lower bound it with the quantity of interest. To this aim we lower bound the integral by the one only around the maximum of the integrand. Using the change of variable $\lambda = \log(1+x)$, we obtain

$$\begin{aligned}
M_t &\geq \frac{1}{2}\int_0^{+\infty} e^{\lambda_x(|S_t|+V_t) - V_t(e^{\lambda_x} - 1)} e^{-x}\mathrm{d}x \geq \frac{1}{2}\int_0^{+\infty} e^{\lambda(|S_t|+V_t+1) - (V_t+1)(e^\lambda - 1)}\mathrm{d}\lambda \\
&\geq \frac{1}{2}\int_{\log(|S_t|/(V_t+1)+1)}^{\log(|S_t|/(V_t+1)+1+1/(V_t+1))} e^{\lambda(|S_t|+V_t+1) - (V_t+1)(e^\lambda - 1)}\mathrm{d}\lambda \\
&\geq \frac{1}{2}\int_{\log(|S_t|/(V_t+1)+1)}^{\log(|S_t|/(V_t+1)+1+1/(V_t+1))} e^{\log(|S_t|/(V_t+1)+1)(|S_t|+V_t+1) - |S_t| - 1}\mathrm{d}\lambda \\
&= \frac{1}{2e} e^{(V_t+1)h(|S_t|/(V_t+1))} \log\left(1 + \frac{1}{|S_t| + V_t + 1}\right) \geq \frac{1}{4e(2t+1)} e^{(V_t+1)h(|S_t|/(V_t+1))}\,,
\end{aligned}$$

where in the last line we used $\log(1 + 1/x) \geq 1/(2x)$ for $x \geq 1$ and the trivial bounds $|S_t| \leq 1$, $V_t \leq t$. The method of mixtures, see (de la Peña et al., 2004), allows us to conclude for the first inequality of the lemma. The second inequality is a straightforward consequence of the previous one. Indeed, using that (see Exercise 2.8 of Boucheron et al., 2013) for $x \geq 0$

$$h(x) \geq \frac{x^2}{2(1 + x/3)},$$

we get

$$\frac{|S_t|/b}{V_t/b^2 + 1} \leq \sqrt{\frac{2\log(4e(2t+1)/\delta)}{V_t/b^2 + 1}} + \frac{2}{3}\frac{\log(4e(2t+1)/\delta)}{V_t/b^2 + 1}.$$

Multiplying by $b(V_t/b^2 + 1)$ the previous inequality allows us to conclude,

$$\begin{aligned}
|S_t| &\leq \sqrt{2(V_t + b)\log(4e(2t+1)/\delta)} + \frac{2b}{3}\log(4e(2t+1)/\delta) \\
&\leq \sqrt{2V_t \log(4e(2t+1)/\delta)} + 3b\log(4e(2t+1)/\delta).
\end{aligned}$$

□

B. Confidence intervals on values

In this section, we define confidence regions for the Q-value and value functions. To define these confidence regions, we first introduce a confidence region for the transition probabilities,

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

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

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

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 policy operator. We also define upper and lower confidence bounds on the optimal value and Q-value functions as

$$\begin{aligned}
\overline{Q}_h^t(s,a) &\triangleq (r_h + \max_{\overline{p}_h \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{V}_{h+1}^t)(s,a) & \underline{Q}_h^t(s,a) &\triangleq (r_h + \min_{\underline{p}_h \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{V}_{h+1}^t)(s,a) \\
\overline{V}_h^t(s) &\triangleq \max_a \overline{Q}_h^t(s,a) & \underline{V}_h^t(s) &\triangleq \max_a \underline{Q}_h^t(s,a) \\
\overline{V}_{H+1}^t(s) &\triangleq 0 & \underline{V}_{H+1}^t(s) &\triangleq 0 \\
\overline{p}_h^t(s,a) &\in \operatorname*{arg\,max}_{\overline{p} \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{V}_{h+1}^t(s,a) & \underline{p}_h^t(s,a) &\in \operatorname*{arg\,min}_{\underline{p} \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{V}_{h+1}^t(s,a) \\
\overline{\pi}_h^t(s,a) &\in \operatorname*{arg\,max}_{a \in \mathcal{A}} \overline{Q}_h^t(s,a) & \underline{\pi}_h^t(s,a) &\in \operatorname*{arg\,max}_{a \in \mathcal{A}} \underline{Q}_h^t(s,a).
\end{aligned}$$

We now build upper confidence bounds that *are not obtained by controlling the deviation of the Kullback-Leibler divergence between the empirical transition probabilities and the true transition probabilities* but only when these empirical transition probabilities are applied to the optimal value function. Precisely, similarly to Azar et al. (2017) and Zanette & Brunskill (2019) we define upper-bounds that take into account the variance as

$$\begin{aligned}
\widetilde{Q}_h^t(s,a) &\triangleq \min \Bigg(H, r_h(s,a) + 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a) \frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 14H^2 \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\qquad + \frac{1}{H} \widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underline{V}_{h+1}^t)(s,a) + \widehat{p}_h^t \widetilde{V}_{h+1}^t(s,a) \Bigg) \\
\widetilde{V}_h^t(s) &\triangleq \max_{a \in \mathcal{A}} \widetilde{Q}_h^t(s,a) \\
\widetilde{V}_{H+1}^t(s) &\triangleq 0, \quad \text{and} \\
\underset{\sim}{Q}_h^t(s,a) &\triangleq \max \Bigg(0, r_h(s,a) - 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a) \frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} - 14H^2 \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\qquad - \frac{1}{H} \widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underline{V}_{h+1}^t)(s,a) + \widehat{p}_h^t \underline{V}_{h+1}^t(s,a) \Bigg) \\
\underline{V}_h^t(s) &\triangleq \max_{a \in \mathcal{A}} \underset{\sim}{Q}_h^t(s,a) \\
\underline{V}_{H+1}^t(s) &\triangleq 0.
\end{aligned}$$

Note that, since the Bernstein inequality is not very sharp, we rather use a loose confidence bound in order to obtain simpler proof. The following confidence bounds are valid with high probability.

Lemma 5. *If $\beta^\star(n,\delta) \le \beta(n,\delta)$ then on event $\mathcal{G}$, we have that for all t, all $h \in [H]$, and all (s,a),*

$$\underset{\sim}{Q}{}_h^t(s,a) \le Q_h^\star(s,a) \le \widetilde{Q}_h^t(s,a) \quad \textit{and}$$
$$\underset{\sim}{V}{}_h^t(s) \le V_h^\star(s) \le \widetilde{V}_h^t(s).$$

Proof. We proceed by induction. For $h = H+1$ the result is trivially true. Assume the inequalities hold for $h' > h$. We prove only the upper bounds, the proof is very similar for the lower bounds. Fix (s,a) and assume that $\widetilde{Q}_h^t(s,a) < H$, otherwise the we trivially have that $\widetilde{Q}_h^t(s,a) \ge Q_h^\star(s,a)$. Note that it implies that $n_h^t(s,a) > 0$. In such case, by construction we have that

$$\begin{aligned}\widetilde{Q}_h(s,a) - Q_h^\star(s,a) \ge 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 10H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ + \frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underset{\sim}{V}{}_{h+1}^t)(s,a) + (\widehat{p}_h^t - p_h)V_{h+1}^\star(s,a) \\ + \widehat{p}_h^t(\widetilde{V}_{h+1}^t - V_{h+1}^\star)(s,a). \end{aligned} \tag{12}$$

Then, by definition of event $\mathcal{G}$ and in particular $\mathcal{E}^\star$, we know that

$$\left|(\widehat{p}_h^t - p_h)V_{h+1}^\star(s,a)\right| \le \sqrt{2\mathrm{Var}_{p_h}(V_{h+1}^\star)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 3H\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}.$$

Using Lemma 11 and Lemma 12, the definition of $\mathcal{G}$ and the induction hypothesis, we get

$$\begin{aligned}\mathrm{Var}_{p_h}(V_{h+1}^\star)(s,a) &\le 2\mathrm{Var}_{\widehat{p}_h^t}(V_{h+1}^\star)(s,a) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\le 4\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a) + 4H\widehat{p}_h^t(\widetilde{V}_{h+1}^t - V_{h+1}^\star)(s,a) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\le 4\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a) + 4H\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underset{\sim}{V}{}_{h+1}^t)(s,a) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}.\end{aligned}$$

Therefore using $\beta^\star(n,\delta) \le \beta(n,\delta)$, $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and $\sqrt{xy} \le x + y$ for $x, y \ge 0$, we obtain

$$\begin{aligned}\left|(\widehat{p}_h^t - p_h)V_{h+1}^\star(s,a)\right| &\le 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \\ &\quad + \sqrt{\frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underset{\sim}{V}{}_{h+1}^t)(s,a)8H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + (\sqrt{8}+3)H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\le 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 14H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\quad + \frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underset{\sim}{V}{}_{h+1}^t)(s,a).\end{aligned}$$

Now going back to (12), due to the inequality above, we arrive to

$$\widetilde{Q}_h(s,a) - Q_h^\star(s,a) \ge \widehat{p}_h^t(\widetilde{V}_{h+1}^t - V_{h+1}^\star)(s,a) \ge 0,$$

where we used the induction hypothesis. Therefore, we have that for all (s,a), $\widetilde{Q}_h^t(s,a) \ge Q_h^\star(s,a)$ and then by definition $\widetilde{V}_h^t(s) = \max_a \widetilde{Q}_h^t(s,a) \ge V_h^\star(s)$. □

C. Proofs of reward-free exploration results

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.

Proof. Assume that we are on event $\mathcal{F}$ and we fix a policy π.

Step 1: Empirical Bernstein inequality From Lemma 10 we know that for any state-action pair (s,a), if $n_h^t(s,a)>0$, then

$$\begin{aligned}\widehat{e}_h^{t,\pi}(s,a;r) &= \left|\widehat{Q}_h^{t,\pi}(s,a;r) - Q_h^\pi(s,a;r)\right| = \left|(\widehat{p}_h^t - p_h)V_{h+1}^\pi(s,a;r)\right| + \widehat{p}_h^t|\widehat{V}_{h+1}^{t,\pi} - V_{h+1}^\pi|(s,a)\\ &\le \sqrt{2\mathrm{Var}_{p_h}(V_{h+1}^\pi)(s,a;r)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \frac{2}{3}H\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \widehat{p}_h^t(\widehat{V}_{h+1}^{t,\pi} - V_{h+1}^\pi)(s,a). \qquad (13)\end{aligned}$$

Now notice that by Lemma 11 with the fact that $0 \le V_{h+1}^\pi \le H$

$$\begin{aligned}\mathrm{Var}_{p_h}(V_{h+1}^\pi)(s,a;r) &\le 2\mathrm{Var}_{\widehat{p}_h^t}(V_{h+1}^\pi)(s,a;r) + 4H^2\,\mathrm{KL}(\widehat{p}_h^t(s,a),p_h(s,a))\\ &\le 2\mathrm{Var}_{\widehat{p}_h^t}(V_{h+1}^\pi)(s,a;r) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}.\end{aligned}$$

Using Lemma 12, we replace the value function by its estimate to get

$$\mathrm{Var}_{\widehat{p}_h^t}(V_{h+1}^\pi)(s,a;r) \le 2\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r) + 2H\widehat{p}_h^t|V_{h+1}^\pi - \widehat{V}_{h+1}^{t,\pi}|(s,a).$$

Combining these two inequalities and using $\sqrt{xy}\le x+y$ for $x,y\ge 0$, we arrive to

$$\begin{aligned}\sqrt{2\mathrm{Var}_{p_h}(V_{h+1}^\pi)(s,a;r)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} &\le 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \sqrt{8}H\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\\ &\quad + \sqrt{\frac{1}{H}\widehat{p}_h^t|V_{h+1}^\pi - \widehat{V}_{h+1}^{t,\pi}|(s,a)8H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}\\ &\le 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}\\ &\quad + (\sqrt{8}H + 8H^2)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \frac{1}{H}\widehat{p}_h^t|V_{h+1}^\pi - \widehat{V}_{h+1}^{t,\pi}|(s,a).\end{aligned}$$

Injecting the above inequality in the initial bound (13) of the estimation error yields

$$\begin{aligned}\widehat{e}_h^{t,\pi}(s,a;r) &\le 3\sqrt{\frac{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 12H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\\ &\quad + \left(1+\frac{1}{H}\right)\widehat{p}_h^t|V_{h+1}^\pi - \widehat{V}_{h+1}^{t,\pi}|(s,a)\\ &\le 3\sqrt{\frac{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\wedge 1\right)} + 15H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\\ &\quad + \left(1+\frac{1}{H}\right)\widehat{p}_h^t|V_{h+1}^\pi - \widehat{V}_{h+1}^{t,\pi}|(s,a),\end{aligned}$$

where in the last inequality we used that if $H^2\beta(n_h^t(s,a),\delta)/n_h^t(s,a)\ge 1$ then

$$3\sqrt{\frac{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \le 3\sqrt{\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \le 3\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}.$$

We therefore obtain the following bound on the error of estimation of the Q-value function at a state-action pair (s,a) in the case when $n_h^t(s,a) > 0$,

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

where we recall the operator notation $\pi_h g(s') \triangleq g(s',\pi_h(s'))$. Defining recursively the functions Z as $Z_{H+1}^{t,\pi}(s,a;r) \triangleq 0$ and for $h \le h$ as

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

and noting that $\widehat{e}_h^{t,\pi}(s,a;r) \le H$ we can show by induction that for all (s,a,h),

$$\widehat{e}_h^{t,\pi}(s,a;r) \le Z_h^{t,\pi}(s,a;r). \tag{14}$$

Step 2: Law of total variance For all (s,a), we now recursively define Y and W. In particular, we set $Y_{H+1}^{t,\pi}(s,a;r) \triangleq W_{H+1}^{t,\pi}(s,a) \triangleq 0$ and

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

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

Again by induction we show that for all h,s,a

$$Z_h^{t,\pi}(s,a;r) \le Y_h^{t,\pi}(s,a;r) + W_h^{t,\pi}(s,a).$$

Indeed, the case $h = H+1$ is trivially true and if we assume the inequality is true at step $h+1$ then using that $\min(x,y+z) \le \min(x,y) + \min(x,z)$ for $x,y,z \ge 0$,

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

Now using (14) we have

$$\pi_1\widehat{e}_1^{t,\pi}(s_1;r) \le \pi_1 Y_1^{t,\pi}(s_1;r) + \pi_1 W_1^{t,\pi}(s_1). \tag{15}$$

Next, we upper-bound the $Y_h^{t,\pi}$ term to *remove the dependency on the empirical variance* of the the value function the policy π. We let $\widehat{p}_h^{t,\pi}(s,a)$ be the probability of reaching the sate-action pair (s,a) at step h under policy π with the empirical transitions at round t. By successive use of the definition of $Y_h^{t,\pi}$, the Cauchy-Schwarz inequality, and Lemma 7

in the empirical MDP, we get

$$\begin{aligned}
\pi Y_1^{t,\pi}(s_1;r) &= 3\sum_{s,a}\sum_{h=1}^{H}\widehat{p}_h^{t,\pi}(s,a)\left(1+\frac{1}{H}\right)^{h-1}\sqrt{\frac{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\wedge 1\right)} \\
&\le 3e\sqrt{\sum_{s,a}\sum_{h=1}^{H}\widehat{p}_h^{t,\pi}(s,a)\frac{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}}\sqrt{\sum_{s,a}\sum_{h=1}^{H}\widehat{p}_h^{t,\pi}(s,a)\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\wedge 1\right)} \\
&\le 3e\sqrt{\frac{1}{H^2}\mathbb{E}_{\pi,\widehat{p}_h^t}\left[\left(\sum_{h=1}^{H}r_h(s_h,a_h)-\widehat{V}_1^{\pi}(s_1;r)\right)^2\right]}\sqrt{\sum_{s,a}\sum_{h=1}^{H}\widehat{p}_h^{t,\pi}(s,a)\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\wedge 1\right)} \\
&\le 3e\sqrt{\sum_{s,a}\sum_{h=1}^{H}\widehat{p}_h^{t,\pi}(s,a)\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\wedge 1\right)}.
\end{aligned}$$

Step 3: Clipping For this step we recursively define $\widetilde{W}^{t,\pi}$ by $\widetilde{W}_{H+1}^{t,\pi}(s,a)\triangleq 0$ and

$$\widetilde{W}_h^{t,\pi}(s,a)\triangleq\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\wedge 1\right)+\widehat{p}_h^{t,\pi}\pi_{h+1}\widetilde{W}_{h+1}^{t,\pi}(s,a),$$

such that by construction

$$\sum_{s,a}\sum_{h=1}^{H}\widehat{p}_h^{t,\pi}(s,a)\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\wedge 1\right)=\pi_1\widetilde{W}_1^{t,\pi}(s_1).$$

By induction we can prove that $\widetilde{W}_h^{t,\pi}\le W_h^{t,\pi}$. Indeed, the inequality is true for $h=H+1$ and if we assume it is true for step $h+1$ then using that by construction $\widetilde{W}_h^{t,\pi}(s,a)\le H$, for all (s,a),

$$\begin{aligned}
\widetilde{W}_h^{t,\pi}(s,a) &= \min\left(H,\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\wedge 1\right)+\widehat{p}_h^{t,\pi}\pi_{h+1}\widetilde{W}_{h+1}^{t,\pi}(s,a)\right) \\
&\le \min\left(H,\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}+\widehat{p}_h^{t,\pi}\pi_{h+1}W_{h+1}^{t,\pi}(s,a)\right) \\
&\le W_h^{t,\pi}(s,a).
\end{aligned}$$

Therefore, we just proved that $\pi_1 Y_1^{t,\pi}(s_1;r)\le 3e\sqrt{\pi_1 W_1^{t,\pi}(s_1)}$ and going back to (15) we get

$$\pi_1\widehat{e}_1^{t,\pi}(s_1;r)\le 3e\sqrt{\pi_1 W_1^{t,\pi}(s_1)}+\pi_1 W_1^{t,\pi}(s_1). \tag{16}$$

Conclusion To finish the proof it remains to note that for all π, for all (s,a,h) we have that

$$W_h^{t,\pi}(s,a)\le W_h^t(s,a)\qquad\text{and}\qquad \pi_h W_h^{t,\pi}(s)\le\max_{a\in\mathcal{A}}W_h^t(s)=\pi_h^{t+1}W_h^t(s,a).$$

□

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

Proof. We first prove that `RF-Express` is (ε, δ)-PAC. This is a simple consequence of Lemma 1. Indeed if the algorithm stops at round τ we know that on event $\mathcal{F}$ for any policy π,

$$\frac{\varepsilon}{2} \geq 2e\sqrt{\max_{a\in\mathcal{A}} W_1^\tau(s_1,a)} + \max_{a\in\mathcal{A}} W_1^\tau(s_1,a) \geq \pi_1 \widehat{e}_1^{\tau,\pi}(s_1;r) = |\widehat{V}_1^{\pi,\tau}(s_1;r) - V_1^\pi(s_1;r)|.$$

Recalling with are still on event $\mathcal{F}$, we have

$$\begin{aligned} V_1^\star(s_1;r) - V_1^{\widehat{\pi}^\star,\tau}(s_1;r) &= V_1^\star(s_1;r) - \widehat{V}_1^{\tau,\pi^\star}(s_1;r) + \widehat{V}_1^{\tau,\pi^\star}(s_1;r) - \widehat{V}_1^{\tau,\widehat{\pi}^\star,\tau}(s_1;r) \\ &\quad + \widehat{V}_1^{\tau,\widehat{\pi}^\star,\tau}(s_1;r) - V_1^{\widehat{\pi}^\star,\tau}(s_1;r) \\ &\leq |V_1^\star(s_1;r) - \widehat{V}_1^{\tau,\pi^\star}(s_1;r)| + |\widehat{V}_1^{\tau,\widehat{\pi}^\star,\tau}(s_1;r) - V_1^{\widehat{\pi}^\star,\tau}(s_1;r)| \leq \varepsilon. \end{aligned}$$

We can conclude the first part of the theorem by noting that by Lemma 3, $\mathbb{P}(\mathcal{F}) \geq 1-\delta$.

It remains to prove the *bound on the sample complexity*. In the rest of the proof we again assume that event $\mathcal{F}$ holds. We start by fixing a round $T < \tau$.

Step 1: Upper bound on W_1^t We first provide an upper bound on $W_h^t(s,a)$ for all (s,a,h) and $t \leq T$. By definition, if $n_h^t(s,a) > 0$, then

$$\begin{aligned} W_h^t(s,a) &\leq 15H^2 \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \left(1+\frac{1}{H}\right)\sum_{s'} \widehat{p}_h^t(s'|s,a) \max_{a'} W_{h+1}^t(s',a') \\ &= 15H^2 \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \left(1+\frac{1}{H}\right)(\widehat{p}_h^t - p_h)\pi_{h+1}^{t+1} W_{h+1}^t(s,a) \\ &\quad + \left(1+\frac{1}{H}\right) p_h \pi_{h+1}^{t+1} W_{h+1}^t(s,a). \end{aligned}$$

Using Lemma 10 we apply Bernstein inequality to get

$$(\widehat{p}_h^t - p_h)\pi_{h+1}^{t+1} W_{h+1}^t(s,a) \leq \sqrt{2\mathrm{Var}_{p_h}(\pi_{h+1}^{t+1} W_{h+1}^t)(s,a)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \frac{2}{3}H\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}.$$

Now since the variance is bounded by $\mathrm{Var}_{p_h}(\pi_{h+1}^{t+1} W_{h+1}^t)(s,a) \leq H p_h \pi_{h+1}^{t+1} W_{h+1}^t(s,a)$, we use the fact that $\sqrt{xy} \leq x+y$ for $x, y \geq 0$ and split the square-root term into two other terms

$$\begin{aligned} \sqrt{2\mathrm{Var}_{p_h}(\pi_{h+1}^{t+1} W_{h+1}^t)(s,a)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} &\leq \sqrt{\frac{1}{H} p_h \pi_{h+1}^{t+1} W_{h+1}^t(s,a) 2H^2 \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \\ &\leq \frac{1}{H} p_h \pi_{h+1}^{t+1} W_{h+1}^t(s,a) + 2H^2 \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}. \end{aligned}$$

We plug this upper bound in the Bernstein inequality above to obtain

$$\left(1+\frac{1}{H}\right)(\widehat{p}_h^t - p_h)\pi_{h+1}^{t+1} W_{h+1}^t(s,a) \leq \frac{16}{3}H^2 \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \frac{2}{H} p_h \pi_{h+1}^{t+1} W_{h+1}^t(s,a).$$

Next, we go back to the initial upper bound and note that by our construction, $W_h^t(s,a) \leq H \leq H^2$ we get for all $n_h^t(s,a) \geq 0$,

$$W_h^t(s,a) \leq 21H^2\left(\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right) + \left(1+\frac{3}{H}\right) p_h \pi_{h+1}^{t+1} W_{h+1}^t(s,a). \tag{17}$$

Unfolding (17) and using that $(1+3/H)^H \leq e^3$ yields

$$\pi_1^{t+1} W_1^t(s_1) \leq 21e^3 H^2 \sum_{h=1}^H \sum_{s,a} p_h^{t+1}(s,a)\left(\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right).$$

Using Lemma 8 we replace the counts by the pseudo-counts in the above inequality

$$\pi_1^{t+1} W_1^t(s_1) \le 84 e^3 H^2 \sum_{h=1}^{H} \sum_{s,a} p_h^{t+1}(s,a) \frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a) \vee 1}, \tag{18}$$

where we recall that $n_h^t(s,a) \triangleq \sum_{\ell=1}^{t} p_h^\ell(s,a)$.

Step 2: Summing over $t \le T$ Since $T < \tau$, we know that for $t \le T$ due to the stopping rule,

$$\varepsilon \le 3e\sqrt{\pi_1^{t+1} W_1^t(s_1)} + \pi_1^{t+1} W_1^t(s_1).$$

Summing the over the above inequalities for $0 \le t \le T$, followed by Cauchy-Schwarz inequality,

$$\begin{aligned}(T+1)\varepsilon &\le \sum_{t=0}^{T} 3e\sqrt{\pi_1^{t+1} W_1^t(s_1)} + \pi_1^{t+1} W_1^t(s_1) \\ &\le 3e\sqrt{(T+1)\sum_{t=0}^{T} \pi_1^{t+1} W_1^t(s_1)} + \sum_{t=0}^{T} \pi_1^{t+1} W_1^t(s_1).\end{aligned}$$

Next, we upper-bound the sum that appears in the left-hand terms. Using successively (18), the property that $\beta(\cdot,\delta)$ is increasing and Lemma 9, we have

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

Therefore, combining just obtained inequality with the previous one, we get

$$(T+1)\varepsilon \le 55 e^3 \sqrt{(T+1) H^3 SA \log(T+2) \beta(T,\delta)} + 336 e^3 H^3 SA \log(T+2) \beta(T,\delta).$$

We now assume that $\tau > 0$ otherwise the result is trivially true. Since the above inequality is true for all $T < \tau$, we get the functional inequality on τ

$$\varepsilon\tau \le 55 e^3 \sqrt{\tau H^3 SA \log(\tau+1) \beta(\tau-1,\delta)} + 336 e^3 H^3 SA \log(\tau+1) \beta(\tau-1,\delta). \tag{19}$$

Step 3: Upper bound on τ It remains to invert of (19). Due the the specific choice of β we are able to further upper-bound τ by

$$\begin{aligned}\varepsilon\tau \le{}& 55 e^3 \sqrt{\tau H^3 SA (\log(3SAH/\delta) \log(8e\tau) + S \log(8e\tau)^2)} \\ &+ 336 e^3 H^3 SA \big(\log(3SAH/\delta) \log(8e\tau) + S \log(8e\tau)^2\big).\end{aligned}$$

We are now ready to use Lemma 13 with

$$C = 55 e^3 \sqrt{H^3 SA}/\varepsilon,\ A = \log(3SAH/\delta),\ B = E = S,\ D = 336 e^3 H^3 SA/\varepsilon, \text{ and } \alpha = 8e$$

to get

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

where we chose $C_1 \triangleq 5587 e^6 \log\big(e^{18} (\log(3SAH/\delta) + S) H^3 SA/\varepsilon\big)^2$ and used that $\varepsilon \le 1$.

□

D. Proofs of best-policy identification results

For this section, we recursively define upper bound on the gap $V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1)$ in order to build the stopping rule. Precisely consider $G_{H+1}^t(s,a) \triangleq 0$ for all (s,a) and for all (s,a,h),

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

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

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

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

Proof. First, we define the following quantities

$$\begin{aligned}
\mathring{Q}_h^t(s,a) \triangleq \min\Bigg(r_h(s,a) + p_h\mathring{V}_h^t(s,a), \max\Bigg(0, r_h(s,a) - 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \\
- 14H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} - \frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t)(s,a) + \widehat{p}_h^t\mathring{V}_{h+1}^t(s,a)\Bigg)\Bigg) \\
\mathring{V}_h^t(s) \triangleq \pi_h^{t+1}\mathring{Q}_h^t(s,a) \\
\mathring{V}_{H+1}^t(s) \triangleq 0.
\end{aligned}$$

On the event $\mathcal{G}$, using Lemma 6 below, we upper-bound the gap at time t as

$$V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1) \le \widetilde{V}_1^t(s_1) - V_1^{\pi^{t+1}}(s_1) \le \widetilde{V}_1^t(s_1) - \mathring{V}_1^t(s_1).$$

Next, we upper-bound the last term in the bound above. We do it by induction on h, showing that for all state-action pairs (s,a),

$$\widetilde{Q}_h^t(s,a) - \mathring{Q}_h^t(s,a) \le G_h^t(s,a). \tag{20}$$

The inequality is trivially true for $h = H+1$. For the induction step, assume it is true for $h+1$ and fix a state-action pair (s,a). Since by construction $\widetilde{Q}_h^t \le H$ and $\mathring{Q}_h^t \ge 0$, then if $G_h^t(s,a) = H$ the inequality is trivially true. We therefore assume that $G_h^t(s,a) < H$ which implies in particular that $n_h^t(s,a) > 0$. We now distinguish the two possible values of $\mathring{Q}_h^t(s,a)$.

Step 1: First case If $\mathring{Q}_h^t(s,a) = r_h(s,a) + p_h\mathring{V}_h^t(s,a)$, then we have

$$\begin{aligned}
\widetilde{Q}_h^t(s,a) - \mathring{Q}_h^t(s,a) \le 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 14H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
+ \frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t)(s,a) + \widehat{p}_h^t\widetilde{V}_{h+1}^t(s,a) - p_h\mathring{V}_{h+1}^t(s,a)\,.
\end{aligned}$$

We upper-bound the last term separately as

$$\begin{aligned}
\widehat{p}_h^t\widetilde{V}_{h+1}^t(s,a) - p_h\mathring{V}_{h+1}^t(s,a) = \widehat{p}_h^t(\widetilde{V}_{h+1}^t - \mathring{V}_{h+1}^t)(s,a) + (\widehat{p}_h^t - p_h)V_{h+1}^\star(s,a) \\
+ (p_h - \widehat{p}_h^t)(V_{h+1}^\star - \mathring{V}_{h+1}^t)(s,a).
\end{aligned} \tag{21}$$

First, proceeding as in the proof of Lemma 5 we know that

$$\begin{aligned}
(\widehat{p}_h^t - p_h)V_{h+1}^\star(s,a) \le 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 14H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
+ \frac{1}{H}\widehat{p}_h^t\left(\widetilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t\right)(s,a).
\end{aligned}$$

Second, using Lemma 10 yields

$$(p_h - \widehat{p}_h^t)(V_{h+1}^\star - \mathring{V}_{h+1}^t)(s,a) \le \sqrt{2\mathrm{Var}_{p_h}(V_{h+1}^\star - \mathring{V}_{h+1}^t)(s,a)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \frac{2}{3}H\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}.$$

Using Lemma 11, we simplify the variance term that appears in the inequality above as

$$\begin{aligned}\mathrm{Var}_{p_h}(V_{h+1}^\star - \mathring{V}_{h+1}^t)(s,a) &\le 2\mathrm{Var}_{\widehat{p}_h^t}(V_{h+1}^\star - \mathring{V}_{h+1}^t)(s,a) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\le 2H\widehat{p}_h^t(V_{h+1}^\star - \mathring{V}_{h+1}^t)(s,a) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}.\end{aligned}$$

Therefore, using the above inequality in the previous one in combination with $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and $\sqrt{xy} \le x + y$ leads to

$$\begin{aligned}(p_h - \widehat{p}_h^t)\Big(V_{h+1}^\star - \mathring{V}_{h+1}^t\Big)(s,a) &\le \frac{1}{H}\widehat{p}_h^t(V_{h+1}^\star - \mathring{V}_{h+1}^t)(s,a) + \left(4H^2 + \sqrt{8}H + \frac{2}{3}H\right)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\le \frac{1}{H}\widehat{p}_h^t\Big(V_{h+1}^\star - \mathring{V}_{h+1}^t\Big)(s,a) + 8H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}.\end{aligned}$$

Going back to (21) and applying just derived bounds and combining them with Lemma 6, we get

$$\begin{aligned}\widehat{p}_h^t\widetilde{V}_{h+1}^t(s,a) - p_h\mathring{V}_{h+1}^t(s,a) &\le \widehat{p}_h^t\Big(\widetilde{V}_{h+1}^t - \mathring{V}_{h+1}^t\Big)(s,a) \\ &\quad + 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 14H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\quad + \frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - V_{h+1}^t)(s,a) \\ &\quad + \frac{1}{H}\widehat{p}_h^t(V_{h+1}^\star - \mathring{V}_{h+1}^t)(s,a) + 8H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\le 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 22H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\quad + \left(1 + \frac{2}{H}\right)\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \mathring{V}_{h+1}^t)(s,a).\end{aligned}$$

Now, using the induction hypothesis and Lemma 6 again, we obtain

$$\begin{aligned}\widetilde{Q}_h^t(s,a) - \mathring{Q}_h^t(s,a) &\le 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 36H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\quad + \left(1 + \frac{3}{H}\right)\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \mathring{V}_{h+1}^t)(s,a) \\ &\le 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 36H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ &\quad + \left(1 + \frac{3}{H}\right)\widehat{p}_h^t\pi_{h+1}^{t+1}G_{h+1}^t(s,a).\end{aligned}$$

Step 2: Second case In the alternative case, we have that

$$\begin{aligned}\mathring{Q}_h^t(s,a) = \max\Bigg(0, r_h(s,a) - 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} - 14H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\ - \frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - V_{h+1}^t)(s,a) + \widehat{p}_h^t\mathring{V}_{h+1}^t(s,a)\Bigg),\end{aligned}$$

and we use Lemma 6 and the induction hypothesis to get

$$\begin{aligned}
\widetilde{Q}_h^t(s,a) - \mathring{Q}_h^t(s,a) &\le 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 24H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\quad + \frac{2}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t)(s,a) + \widehat{p}_h^t(\widetilde{V}_{h+1}^t - \mathring{V}_{h+1}^t)(s,a) \\
&\le 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 24H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\quad + \left(1+\frac{2}{H}\right)\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \mathring{V}_{h+1}^t)(s,a) \\
&\le 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 24H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\quad + \left(1+\frac{2}{H}\right)\widehat{p}_h^t\pi_{h+1}^{t+1}G_{h+1}^t(s,a)\,.
\end{aligned}$$

Conclusion In both cases, since we assumed that $G_h^t(s,a) < H$ and using the definition of G_h^t, we proved that

$$\widetilde{Q}_h^t(s,a) - \mathring{Q}_h^t(s,a) \le G_h^t(s,a)\,.$$

In particular, for $h = 1$ we conclude that

$$\widetilde{V}_1^t(s_1) - \mathring{V}_1^t(s_1) = \pi_1^{t+1}(\widetilde{Q}_1^t - \mathring{Q}_1^t)(s_1) \le \pi_1^{t+1}G_1^t(s_1).$$

□

Lemma 6. *On event $\mathcal{G}$, we have that for all (s,a,h),*

$$\begin{aligned}
\mathring{Q}_h^t(s,a) &\le \min\left(\underset{\sim}{Q}_h^t(s,a), Q_h^{\pi^{t+1}}(s,a)\right) \\
\mathring{V}_h^t(s) &\le \min\left(\underset{\sim}{V}_h^t(s), V_h^{\pi^{t+1}}(s)\right),
\end{aligned}$$

where $\mathring{Q}_h^t$ and $\mathring{V}_h^t$ are defined in the proof of Lemma 2.

Proof. We proceed by induction on h. For $h = H+1$ the inequalities are trivially true. For that induction step, we assume them true for $h+1$ and therefore for all (s,a), we have

$$\begin{aligned}
\mathring{Q}_h^t(s,a) &\le r_h(s,a) + p_h\mathring{V}_{h+1}^t(s,a) \\
&\le r_h(s,a) + p_h V_{h+1}^{\pi^{t+1}}(s,a) = Q_h^{\pi^{t+1}}(s,a)\,.
\end{aligned}$$

Similarly to the above, we obtain

$$\begin{aligned}
\mathring{Q}_h^t(s,a) &\le \max\Bigg(0, r_h(s,a) - 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \\
&\qquad - 14H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} - \frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t)(s,a) + \widehat{p}_h^t\mathring{V}_{h+1}^t(s,a)\Bigg) \\
&\le \max\Bigg(0, r_h(s,a) - 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \\
&\qquad - 14H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} - \frac{1}{H}\widehat{p}_h^t(\widetilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t)(s,a) + \widehat{p}_h^t\underset{\sim}{V}_{h+1}^t(s,a)\Bigg) = \underset{\sim}{Q}_h^t(s,a).
\end{aligned}$$

It remains to prove the inequalities for the value functions. Using the inequalities with the Q-value functions we get

$$\mathring{V}_h^t(s) \le \pi_h^{t+1} Q_h^{\pi^{t+1}}(s) = V_h^{\pi^{t+1}}(s), \quad \text{and}$$
$$\mathring{V}_h^t(s) \le \pi_h^{t+1} \underset{\sim}{Q}_h^t(s) \le \max_{a \in \mathcal{A}} \underset{\sim}{Q}_h^t(s,a) = \underset{\sim}{V}_h^t(s)\,.$$

□

Theorem 2. *For $\delta \in (0,1)$, $\varepsilon \in (0, 1/S^2]$,* `BPI-UCBVI` *using thresholds $\beta(n,\delta) \triangleq \log\big(3SAH/\delta\big) + S\log\big(8e(n+1)\big)$ 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 $1-\delta$,*

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

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

Proof. We first prove that `BPI-UCBVI` is (ε,δ)-PAC. This is a simple consequence of Lemma 2. Indeed, if our algorithm stops at time τ we know that on event $\mathcal{G}$

$$V_1^{\widehat{\pi}}(s_1) = V_1^{\pi^{\tau+1}}(s_1) \ge V_1^\star(s_1) - \pi_1^{\tau+1} G_1^\tau(s_1) \ge V_1^\star(s_1) - \varepsilon\,.$$

We can then conclude the first part of the theorem by noting that by Lemma 3, $\mathbb{P}(\mathcal{G}) \ge 1-\delta$.

It remains to prove the bound on the sample complexity. In the rest of the proof we assume that the event $\mathcal{G}$ holds and fix a round $T < \tau$.

Step 1: Upper bound on G_1^t We first provide an upper bound on $G_h^t(s,a)$ for all (s,a,h) and $t \le T$. If $n_h^t(s,a) > 0$ by definition of G_h^t we have

$$G_h^t(s,a) \le 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 36H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$
$$+ \left(1 + \frac{3}{H}\right)\widehat{p}_h^t \pi_{h+1}^{t+1} G_{h+1}^t(s,a). \tag{22}$$

Now, we replace the empirical transition probability by the true one. Using Lemma 10 and that $0 \le G_h^t \le H$, we get

$$(\widehat{p}_h^t - p_h)\pi_{h+1}^{t+1} G_{h+1}^t(s,a) \le \sqrt{2\mathrm{Var}_{p_h}(\pi_{h+1}^{t+1}G_{h+1}^t)(s,a)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \frac{2}{3}H\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$
$$\le \frac{1}{H}p_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a) + 3H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)},$$

where in the last line we used $\mathrm{Var}_{p_h}(\pi_{h+1}^{t+1}G_{h+1}^t)(s,a) \le H\pi_{h+1}^{t+1}G_{h+1}^t(s,a)$ and $\sqrt{xy} \le x+y$ for all $x,y \ge 0$. We also need to replace the variance of the upper confidence bound under the empirical transition by the variance of the optimal value function under the true transition probability. Using Lemma 11 then Lemma 12, we obtain

$$\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a) \le 2\mathrm{Var}_{p_h}(\widetilde{V}_{h+1}^t)(s,a) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$
$$\le 4\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a) + 4Hp_h(\widetilde{V}_{h+1}^t - V_{h+1}^{\pi^{t+1}})(s,a) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$
$$\le 4\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a) + 4Hp_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$

where we used (20) from the proof of Lemma 2 in the last inequality. Next, using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, $\sqrt{xy} \leq x + y$, and $\beta^\star(n,\delta) \leq \beta(n,\delta)$ leads to

$$\begin{aligned}
\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} &\leq 2\sqrt{\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + (2H+4H^2)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\quad + \frac{1}{H}p_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a) \\
&\leq 2\sqrt{\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 6H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\quad + \frac{1}{H}p_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a).
\end{aligned}$$

Combining these two inequalities with (22) yields

$$\begin{aligned}
G_h^t(s,a) &\leq 12\sqrt{\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 36H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\quad + \frac{6}{H}p_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a) + 36H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\quad + \left(1+\frac{3}{H}\right)\frac{1}{H}p_h\pi_{h+1}^{t+1}G_{h+1}^t + \left(1+\frac{3}{H}\right)3H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\quad + \left(1+\frac{3}{H}\right)p_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a) \\
&\leq 12\sqrt{\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 84H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\quad + \left(1+\frac{13}{H}\right)p_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a).
\end{aligned}$$

Since by construction, $G_h^t(s,a) \leq H$, we have that for all $n_h^t(s,a) \geq 0$,

$$\begin{aligned}
G_h^t(s,a) &\leq 12\sqrt{\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)\left(\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right)} + 84H^2\left(\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right) \\
&\quad + \left(1+\frac{13}{H}\right)p_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a).
\end{aligned}$$

Unfolding the previous inequality and using $(1+13/H)^H \leq e^{13}$ we get

$$\begin{aligned}
\pi_1 G_1^t(s_1) &\leq 12e^{13}\sum_{h=1}^H\sum_{s,a}p_h^{t+1}(s,a)\sqrt{\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)\left(\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right)} \\
&\quad + 84e^{13}H^2\sum_{h=1}^H\sum_{s,a}p_h^{t+1}(s,a)\left(\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right),
\end{aligned}$$

where we recall that $p_h^{t+1}(s,a)$ is the probability of reaching a state-action pair (s,a) at step h under policy π^{t+1}. Using Lemma 8 we replace the counts by the pseudo-counts in the previous inequality as

$$\begin{aligned}
\pi_1 G_1^t(s_1) &\leq 24e^{13}\sum_{h=1}^H\sum_{s,a}p_h^{t+1}(s,a)\sqrt{\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)\frac{\beta^\star(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}} \\
&\quad + 336e^{13}H^2\sum_{h=1}^H\sum_{s,a}p_h^{t+1}(s,a)\frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}. \qquad (23)
\end{aligned}$$

Step 2: Law of total variance. Using Lemma 7 we further upper-bound the first sum in (23). In particular, by Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\sqrt{\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)\frac{\beta^\star(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}} \\
&\qquad\le \sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)}\sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta^\star(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}} \\
&\qquad\le \sqrt{\mathbb{E}_{\pi^{t+1}}\left[\left(\sum_{h=1}^{H} r_h(s_h,a_h) - V_1^{\pi^{t+1}}(s_1)\right)^2\right]}\sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta^\star(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}} \\
&\qquad\le H\sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta^\star(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}}.
\end{aligned}$$

Therefore, going back to (23), we obtain

$$\begin{aligned}
\pi_1 G_1^t(s_1) \le\ & 24e^{13}H\sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta^\star(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}} \\
&+ 336e^{13}H^2\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}. \qquad (24)
\end{aligned}$$

Step 3: Summing over $t \le T$ Since $T < \tau$, we know that for $t \le T$, using due to the design of the stopping rule,

$$\varepsilon \le \pi_1^{t+1} G_1^t(s_1)\,.$$

Therefore, summing the previous inequalities in combination with (24), yields

$$\begin{aligned}
(T+1)\varepsilon &\le 24e^{13}H\sum_{t=1}^{T}\sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta^\star(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}} + 336e^{13}H^2\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1} \\
&\le 24e^{13}H\sqrt{T}\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta^\star(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}} + 336e^{13}H^2\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1}.
\end{aligned}$$

Using successively that $\beta(.,\delta)$ and $\beta^\star(.,\delta)$ are increasing, and then Lemma 9, we have

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta^\star(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1} &\le \beta^\star(T,\delta)\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{1}{\bar{n}_h^t(s,a)\vee 1} \\
&\le \beta^\star(T,\delta)\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s,a}\frac{\bar{n}_h^{t+1}(s,a)-\bar{n}_h^t(s,a)}{\bar{n}_h^t(s,a)\vee 1} \\
&\le 4HSA\beta^\star(T,\delta)\log(T+2).
\end{aligned}$$

Similarly, for the second sum we get

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a)\vee 1} \le 4HSA\beta(T,\delta)\log(T+2).$$

Therefore, we obtain that

$$(T+1)\varepsilon \le 48e^{13}\sqrt{T}\sqrt{H^3SA\beta^\star(T,\delta)\log(T+2)} + 1344e^{13}H^3SA\beta(T,\delta)\log(T+2).$$

We assume that $\tau > 0$ otherwise the result is trivially true. Since the above inequality is true for all $T < \tau$, we get the functional inequality for τ

$$\varepsilon\tau \le 48e^{13}\sqrt{\tau H^3 SA\beta^\star(\tau-1,\delta)\log(\tau+1)} + 1344e^{13}H^3SA\beta(\tau-1,\delta)\log(\tau+1). \tag{25}$$

Step 3: Upper bound on τ It remains to invert (25). Using the definition of β and $\beta^\star$ yields

$$\begin{aligned}\varepsilon\tau \le{}& 48e^{13}\sqrt{\tau H^3 SA\big(\log(3SAH/\delta)\log(8e\tau)+\log(8e\tau)^2\big)}\\ &+ 1344e^{13}H^3SA\big(\log(3SAH/\delta)\log(8e\tau)+S\log(8e\tau)^2\big).\end{aligned}$$

We can now use Lemma 13 with

$$C = 48e^{13}\sqrt{H^3SA}/\varepsilon,\ A = \log(3SAH/\delta),\ B = 1,\ E = S,\ D = 1344e^{13}H^3SA/\varepsilon,\ \text{ and } \alpha = 8e$$

to get

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

with $C_1 \triangleq 5904e^{26}\log\big(e^{30}(\log(3SAH/\delta)+S)H^3SA/\varepsilon\big)^2$ and used that $\varepsilon \le 1/S^2$. □

E. From regret-minimization to BPI

The main difficulty for converting a regret-minimizing algorithm to BPI lies in high-probability prediction of an ε-optimal policy. Indeed, assume that a regret-minimizing algorithm generates a sequence of policy $(\pi^t)_{t\in[T]}$ for a number of episodes $T \in \mathbb{N}$ large enough, with a controlled regret such that with probability at least $1-\delta'$,

$$\sum_{t=1}^{T} V^\star(s_1) - V_1^{\pi^t}(s_1) \le C\sqrt{H^3SA\log(1/\delta')T},$$

for some constant C. Then a straightforward choice for the prediction is to choose $\widehat{\pi}$ at random among the sequence $(\pi^t)_{t\in[T]}$, as suggested by (Jin et al., 2018). Markov's inequality implies that

$$\begin{aligned}\mathbb{P}\Big(V_1^\star(s_1) - V_1^{\pi^t}(s_1) \ge \varepsilon\Big) &\le \frac{1}{\varepsilon}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} V^\star(s_1) - V_1^{\pi^t}(s_1)\right] \\ &\le \frac{1}{\varepsilon}\left(C\sqrt{\frac{H^3SA}{T}\log(1/\delta')} + \delta' H\right).\end{aligned}$$

Therefore, if we choose $\delta' \triangleq \frac{\varepsilon}{H}\cdot\frac{\delta}{2}$ and

$$T \triangleq \frac{2H^3SA}{\varepsilon^2\delta^2}\log\left(\frac{2H}{\varepsilon\delta}\right),$$

we get $\mathbb{P}\big(V^\star(s_1) - V_1^{\pi^t}(s_1) \ge \varepsilon\big) \le \delta$ which implies that the described algorithm is (ε,δ)-PAC for BPI. Unfortunately, the choice of T above gives a sample complexity that scales with $1/\delta^2$ whereas we prefer and expect $\log(1/\delta)$.

Another natural choice for the prediction is to return the average policy. Precisely, let us define $\bar{p}_h^t(s,a) \triangleq \sum_{\ell=1}^t p_h^\ell(s,a)$ be the average state-action distribution, we define the average distribution as the one for which its state-action distribution is precisely $\bar{p}^t$,

$$\bar{\pi}_h^t(a|s) \triangleq \begin{cases} \frac{\bar{p}_h^t(s,a)}{\sum_{b\in\mathcal{A}}\bar{p}_h^t(s,b)} & \text{if } \sum_{b\in\mathcal{A}}\bar{p}_h^t(s,b) > 0, \text{ and} \\ 1/A & \text{otherwise.}\end{cases}$$

Note that the average policy $\bar{\pi}_h^t(a|s)$ is not the average of the policies $\sum_{\ell=1}^t \pi_h^\ell(a|s)/t$. The interest of $\bar{\pi}_h^t$ is that the average regret at time T is *equal* to the simple regret of the policy $\bar{\pi}^T$,

$$\begin{aligned}V_1^\star(s_1) - V_1^{\bar{\pi}^T}(s_1) &\le \frac{1}{T}\sum_{t=1}^{T} V^\star(s_1) - V_1^{\pi^t}(s_1) \\ &\le C\sqrt{\frac{H^3SA}{T}\log(1/\delta')},\end{aligned}$$

with probability $1-\delta'$. Choosing $\delta' \triangleq \delta$ and $T \triangleq CH^3SA\log(1/\delta)/\varepsilon^2$ and predicting policy $\bar{\pi}^T$ would lead to an (ε,δ)-PAC algorithm for BPI with a minimax optimal sample complexity. However, since the agent does not know the transition kernel, it cannot compute the average policy $\bar{\pi}^T$! Therefore, `BPI-UCBVI` rather relies on an upper bound on the simple regret (see Lemma 2) computable by the agent to detect if the policy currently used by the sampling rule is ε-optimal.

F. Technical lemmas

F.1. A Bellman-type equation for the variance

For a deterministic policy π we define Bellman-type equations for the variances as follows

$$\begin{aligned}\sigma Q_h^\pi(s,a) &\triangleq \mathrm{Var}_{p_h} V_{h+1}^\pi(s,a) + p_h \sigma V_{h+1}^\pi(s,a)\\ \sigma V_h^\pi(s) &\triangleq \sigma Q_h^\pi(s,\pi(s))\\ \sigma V_{H+1}^\pi(s) &\triangleq 0,\end{aligned}$$

where $\mathrm{Var}_{p_h}(f)(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}\big[(f(s') - p_h f(s,a))^2\big]$ denotes the *variance operator*. In particular, the function $s \mapsto \sigma V_1^\pi(s)$ represents the average sum of the local variances $\mathrm{Var}_{p_h} V_{h+1}^\pi(s,a)$ over a trajectory following the policy π, starting from (s,a). Indeed, the definition above implies that

$$\sigma V_1^\pi(s_1) = \sum_{h=1}^H \sum_{s,a} p_h^\pi(s,a) \mathrm{Var}_{p_h}(V_{h+1}^\pi)(s,a).$$

The lemma below shows that we can relate the global variance of the cumulative reward over a trajectory to the average sum of local variances.

Lemma 7 (Law of total variance)**.** *For any deterministic policy π and for all $h \in [H]$,*

$$\mathbb{E}_\pi\left[\left(\sum_{h'=h}^H r_{h'}(s_{h'},a_{h'}) - Q_h^\pi(s_h,a_h)\right)^2 \middle| (s_h,a_h) = (s,a)\right] = \sigma Q_h^\pi(s,a).$$

In particular,

$$\mathbb{E}_\pi\left[\left(\sum_{h=1}^H r_h(s_h,a_h) - V_1^\pi(s_1)\right)^2\right] = \sigma V_1^\pi(s_1) = \sum_{h=1}^H \sum_{s,a} p_h^\pi(s,a)\mathrm{Var}_{p_h}(V_{h+1}^\pi)(s,a).$$

Proof. We proceed by induction. The statement in Lemma 7 is trivial for $h = H+1$. We now assume that it holds for $h+1$ and prove that it also holds for h. For this purpose, we compute

$$\begin{aligned}&\mathbb{E}_\pi\left[\left(\sum_{h'=h}^H r_{h'}(s_{h'},a_{h'}) - Q_h^\pi(s_h,a_h)\right)^2 \middle| (s_h,a_h)\right]\\ &= \mathbb{E}_\pi\left[\left(Q_{h+1}^\pi(s_{h+1},a_{h+1}) - p_h V_{h+1}^\pi(s_h,a_h) + \sum_{h'=h+1}^H r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)^2 \middle| (s_h,a_h)\right]\\ &= \mathbb{E}_\pi\left[\big(Q_{h+1}^\pi(s_{h+1},a_{h+1}) - p_h V_{h+1}^\pi(s_h,a_h)\big)^2 \middle| (s_h,a_h)\right]\\ &+ \mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^H r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)^2 \middle| (s_h,a_h)\right]\\ &+ 2\mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^H r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)\big(Q_{h+1}^\pi(s_{h+1},a_{h+1}) - p_h V_{h+1}^\pi(s_h,a_h)\big) \middle| (s_h,a_h)\right].\end{aligned}$$

The definition of $Q_{h+1}^\pi(s_{h+1},a_{h+1})$ implies that

$$\mathbb{E}_\pi\left[\sum_{h'=h+1}^H r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1}) \middle| (s_{h+1},a_{h+1})\right] = 0.$$

Therefore, the law of total expectation gives us

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'}) - Q_h^\pi(s_h,a_h)\right)^2 \middle| (s_h,a_h)\right] \\
&= \mathbb{E}_\pi\Big[\big(V_{h+1}^\pi(s_{h+1}) - p_h V_{h+1}^\pi(s_h,a_h)\big)^2 \Big| (s_h,a_h)\Big] + \mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^{H} r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)^2 \middle| (s_h,a_h)\right] \\
&= \mathrm{Var}_{p_h} V_{h+1}^\pi(s_h,a_h) \\
&\quad + \sum_{(s_{h+1},a_{h+1})} p_h(s_{h+1}|s_h,a_h)\mathbb{1}_{(a_{h+1}=\pi(s_{h+1}))}\mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^{H} r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)^2 \middle| (s_{h+1},a_{h+1})\right] \\
&= \mathrm{Var}_{p_h} V_{h+1}^\pi(s_h,a_h) + p_h \sigma V_{h+1}^\pi(s_h,a_h) \\
&= \sigma Q_h^\pi(s_h,a_h)
\end{aligned}$$

where in the third equality we used the inductive hypothesis and the definition of σV_{h+1}^π. □

F.2. Counts to pseudo-counts

Lemma 8. *On event $\mathcal{E}^{\mathrm{cnt}}$, $\forall h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$,*

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

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

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

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

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

where we used that $x \mapsto \beta(x,\delta)/x$ is non-increasing for $x \ge 1$, $x \mapsto \beta(x,\delta)$ is non-decreasing, and $\beta^{\mathrm{cnt}}(\delta) \ge 1$. Second, if $\beta^{\mathrm{cnt}}(\delta) > \frac{1}{4}\bar{n}_\ell^t(s,a)$, a simple derivation gives that

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

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

Lemma 9. *For $T \in \mathbb{N}^\star$ and $(u_t)_{t\in\mathbb{N}^\star}$, for a sequence where $u_t \in [0,1]$ and $U_t \triangleq \sum_{l=1}^t u_\ell$, we get*

$$\sum_{t=0}^{T} \frac{u_{t+1}}{U_t \vee 1} \le 4\log(U_{T+1}+1).$$

Proof. Notice that

$$\begin{aligned}
\sum_{t=0}^{T} \frac{u_{t+1}}{U_t\vee 1} &\le 4\sum_{t=0}^{T}\frac{u_{t+1}}{2U_t+2} \\
&\le 4\sum_{t=0}^{T}\frac{U_{t+1}-U_t}{U_{t+1}+1} \\
&\le 4\sum_{t=0}^{T}\int_{U_t}^{U_{t+1}}\frac{1}{x+1}\mathrm{d}x \\
&= 4\log(U_{T+1}+1).
\end{aligned}$$

□

F.3. On Bernstein inequality

We reproduce a Bernstein-type inequality similar to the one by Talebi & Maillard (2018).

Lemma 10. *Let $p, q \in \Sigma_S$, where Σ_S denotes the probability simplex of dimension $S-1$. For all $\alpha > 0$, for all functions f defined on $\mathcal{S}$ with $0 \le f(s) \le b$, for all $s \in \mathcal{S}$, if $\mathrm{KL}(p,q) \le \alpha$ then*

$$|pf - qf| \le \sqrt{2\mathrm{Var}_q(f)\alpha} + \frac{2}{3}b\alpha,$$

where use the expectation operator defined as $pf \triangleq \mathbb{E}_{s\sim p} f(s)$ and the variance operator defined as $\mathrm{Var}_p(f) \triangleq \mathbb{E}_{s\sim p}\big(f(s) - \mathbb{E}_{s'\sim p} f(s')\big)^2 = p(f-pf)^2$.

Proof. We only prove that

$$qf - pf \le \sqrt{2\mathrm{Var}_q(f)\alpha} + \frac{2}{3}b\alpha$$

as an upper bound on $pf - qf$ can be obtained analogically. We assume that $qf > pf$, otherwise the inequality is trivially true. Furthermore, without loss of generality, we consider that $0 \le f(s) \le 1$ for all $s \in \mathcal{S}$: when f is bounded by b, we can apply the inequality to f/b and by homogeneity, we get the desired general version. Using the variational formula for the Kullback-Leibler divergence, we have

$$\begin{aligned}\mathrm{KL}(p,q) &= \sup_{g\in\mathbb{R}^S} pg - \log(qe^g) \\ &\ge \sup_{\lambda\ge 0} \lambda(qf - pf) - \log\Big(qe^{\lambda(qf-f)}\Big),\end{aligned}$$

where we chose $g \triangleq \lambda(qf - f)$ and used that $pqf = qf$ since qf is a scalar. Given that the mapping $u \to (e^u - u - 1)/u^2$ is non-decreasing in u and that $qf - f(s) \le 1$, we obtain

$$e^{\lambda\left(qf - f(s)\right)} - \lambda\big(qf - f(s)\big) - 1 \le \big(qf - f(s)\big)^2(e^\lambda - \lambda - 1).$$

Furthermore, by taking the expectation of the previous inequality with respect to q, using the property $\log(1+x) \le x$, and define function $\phi(\lambda) \triangleq e^\lambda - \lambda - 1$, we get

$$\log(qe^{\lambda(qf-f)}) \le \log\big(1 + \mathrm{Var}_q(f)\phi(\lambda)\big) \le \mathrm{Var}_q(f)\phi(\lambda).$$

Now using the previous inequality in the variational formula above with the fact that the convex conjugate of ϕ is for $u \ge 0$, $\sup_{\lambda\ge 0} \lambda u - \phi(\lambda) \triangleq h(u) \ge u^2/(2(1 + u/3))$, yields

$$\begin{aligned}\mathrm{KL}(p,q) &\ge \sup_{\lambda\ge 0} \lambda(qf - pf) - \mathrm{Var}_q(f)\phi(\lambda) \\ &= \mathrm{Var}_q(f) h\left(\frac{qf - pf}{\mathrm{Var}_q(f)}\right) \\ &\ge \frac{(qf - pf)^2}{2(\mathrm{Var}_q(f) + (qf - pf)/3)}.\end{aligned}$$

Since by assumption $\mathrm{KL}(p,q) \le \alpha$, we get

$$2\alpha(\mathrm{Var}_q(f) + (qf - pf)/3) - (qf - pf)^2 \ge 0.$$

Hence $qf - pf$ is upper bounded by the positive root of the polynomial in the left hand side,

$$\frac{\alpha}{3} + \sqrt{2\alpha\mathrm{Var}_q(f) + \frac{\alpha^2}{9}} \le \sqrt{2\alpha\mathrm{Var}_q(f)} + \frac{2}{3}\alpha,$$

using that $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$. □

Lemma 11. *Let $p, q \in \Sigma_S$ and f is a function defined on $\mathcal{S}$ such that $0 \leq f(s) \leq b$ for all $s \in \mathcal{S}$. If $\mathrm{KL}(p, q) \leq \alpha$ then*

$$\mathrm{Var}_q(f) \leq 2\mathrm{Var}_p(f) + 4b^2\alpha \quad \textit{and}$$
$$\mathrm{Var}_p(f) \leq 2\mathrm{Var}_q(f) + 4b^2\alpha.$$

Proof. Let $\widetilde{p}$ be the distribution of the pair of random variables (X, Y) where X, Y are i.i.d. according to the distribution p. Similarly, let $\widetilde{q}$ be the distribution of the pair of random variables (X, Y) where X, Y are i.i.d. according to distribution q. Since Kullback–Leibler divergence is additive for independent distributions, we know that

$$\mathrm{KL}(\widetilde{p}, \widetilde{q}) = 2\,\mathrm{KL}(p, q) \leq 2\alpha.$$

Using Lemma 10 for the function $g(x, y) = (f(x) - f(y))^2$ defined on $\mathcal{S}^2$, such that $0 \leq g \leq b^2$, we get

$$\begin{aligned}
|\widetilde{p}g - \widetilde{q}g| &\leq \sqrt{4\mathrm{Var}_{\widetilde{q}}(g)\alpha} + \frac{4}{3}b^2\alpha \\
&\leq \sqrt{4b^2\alpha\widetilde{q}g} + \frac{4}{3}b^2\alpha \\
&\leq \frac{1}{2}\widetilde{q}g + \frac{10}{3}b^2\alpha\,,
\end{aligned}$$

where in the last line we used $2\sqrt{xy} \leq x + y$ for $x, y \geq 0$. In particular we obtain

$$\begin{aligned}
\widetilde{p}g &\leq \frac{3}{2}\widetilde{q}g + \frac{10}{3}b^2\alpha \\
\widetilde{q}g &\leq 2\widetilde{p}g + \frac{20}{3}b^2\alpha\,.
\end{aligned}$$

To conclude, it remains to note that

$$\widetilde{p}g = 2\mathrm{Var}_p(f) \text{ and } \widetilde{q}g = 2\mathrm{Var}_q(f).$$

□

Lemma 12. *For $p, q \in \Sigma_S$, for f, g two functions defined on $\mathcal{S}$ such that $0 \leq g(s), f(s) \leq b$ for all $s \in \mathcal{S}$, we have that*

$$\mathrm{Var}_p(f) \leq 2\mathrm{Var}_p(g) + 2bp|f - g| \quad \textit{and}$$
$$\mathrm{Var}_q(f) \leq \mathrm{Var}_p(f) + 3b^2\|p - q\|_1,$$

where we denote the absolute operator by $|f|(s) = |f(s)|$ for all $s \in \mathcal{S}$.

Proof. First note that

$$\mathrm{Var}_p(f) = p(f - g + g - pg + pg - pf)^2 \leq 2p(f - g - pf + pg)^2 + 2p(g - pg)^2 = 2\mathrm{Var}_p(f - g) + 2\mathrm{Var}_p(g).$$

From the above we can immediately conclude the proof of the first inequality with

$$\mathrm{Var}_p(f - g) \leq p(f - g)^2 \leq bp|f - g|,$$

where we used that for all $s \in \mathcal{S}$, $0 \leq |f(s) - g(s)| \leq b$. For the second inequality, using the Hölder inequality,

$$\begin{aligned}
\mathrm{Var}_q(f) &= pf^2 - (pf)^2 + (q - p)f^2 + (pf)^2 - (qf)^2 \\
&\leq \mathrm{Var}_p(f) + b^2\|p - q\|_1 + 2b^2\|p - q\|_1 \\
&\leq \mathrm{Var}_p(f) + 3b^2\|p - q\|_1.
\end{aligned}$$

□

F.4. An auxiliary inequality

Lemma 13. *Let A, B, C, D, E, and α be positive scalars such that $1 \leq B \leq E$ and $\alpha \geq e$. If $\tau \geq 0$ satisfies*

$$\tau \leq C\sqrt{\tau\big(A\log(\alpha\tau) + B\log(\alpha\tau)^2\big)} + D\big(A\log(\alpha\tau) + E\log(\alpha\tau)^2\big) \tag{26}$$

then

$$\tau \leq C^2(A+B)C_1^2 + \Big(D + 2\sqrt{D}C\Big)(A+E)C_1^2 + 1,$$

where

$$C_1 = \frac{8}{5}\log\big(11\alpha^2(A+E)(C+D)\big).$$

Proof. We can assume that $\tau \geq 1$ otherwise the result is trivially true. Using $\log(x) \leq x^\beta/\beta$ for $x \geq 0, \beta > 0$, and $\alpha \geq 1$ we get

$$\begin{aligned}
\tau\big(A\log(\alpha\tau) + B\log(\alpha\tau)^2\big) &\leq \tau\Big(4A\alpha^{1/4}\tau^{1/4} + B(8\alpha^{1/8}\tau^{1/8})^2\Big) \\
&\leq (4A\alpha^{1/4} + 64B\alpha^{1/4})\tau^{5/4} \\
&\leq 64\alpha^{1/4}(A+B)\tau^{5/4},
\end{aligned}$$

and consequently,

$$C\sqrt{\tau(A\log(\alpha\tau) + B\log(\alpha\tau)^2)} \leq 8C\alpha\sqrt{A+B}\tau^{5/8}.$$

Similarly,

$$\begin{aligned}
D\big(A\log(\alpha\tau) + E\log(\alpha\tau)^2\big) &\leq D\left(8A\alpha\tau^{5/8} + E\left(\frac{16}{5}\alpha^{5/16}\tau^{5/16}\right)^2\right) \\
&\leq 11D\alpha(A+E)\tau^{5/8}.
\end{aligned}$$

Using the above two inequalities in combination with (26) we obtain a first crude upper bound

$$\tau \leq (11\alpha(A+E)(C+D))^{8/5}.$$

if we let C_1 be the constant defined as

$$C_1 \triangleq \frac{8}{5}\log\big(11\alpha^2(A+E)(C+D)\big),$$

then upper-bounding the log in (26) with C_1 yields a simpler inequality verified by τ,

$$\tau \leq C\sqrt{\tau(AC_1 + BC_1^2)} + D\big((AC_1 + EC_1^2\big).$$

Next, define the constants $a \triangleq C\sqrt{AC_1 + BC_1^2}$ and $b \triangleq D(AC_1 + BC_1^2)$. Hence, $x = \sqrt{\tau}$ satisfies

$$x^2 - ax - b \leq 0,$$

which implies that x is upper-bounded by the larger roots of the polynomial above

$$x \ \leq \frac{a + \sqrt{a^2 + 4b}}{2} \leq a + \sqrt{b}.$$

We there deduce the following bound on τ

$$\begin{aligned}
\tau &\leq (a + \sqrt{b})^2 \\
&\leq a^2 + 2\sqrt{b}a + b \\
&\leq C^2(A+B)C_1^2 + \Big(D + 2\sqrt{D}C\Big)(A+E)C_1^2,
\end{aligned}$$

where in the last inequality we used that $C_1 \leq C_1^2$ since $\alpha \geq e$ and $B \leq E$. □

G. Experiments

In this section, we illustrate experimentally how $1/n$ bonuses can improve exploration when compared to $1/\sqrt{n}$ in a reward-free setting. Using a grid-world environment consisting of 15 rooms with 25 states per room, we compare `RF-Express` ($1/n$ bonuses) to `RF-UCRL` ($1/\sqrt{n}$). Let $\widehat{p}_{\mathrm{v}}^t(s) = \frac{1}{t}\sum_{h,a} n_h^t(s,a)$ be the empirical distribution of state visits after t episodes. Figure 1 shows the entropy of $\widehat{p}_{\mathrm{v}}^t$ and the number of new visited states versus the number of episodes t for each algorithm. We observe that $\widehat{p}_{\mathrm{v}}^t$ obtained with $1/n$ bonuses is closer to the uniform distribution than the one obtained with $1/\sqrt{n}$, and that `RF-Express` also discovers new states more quickly.

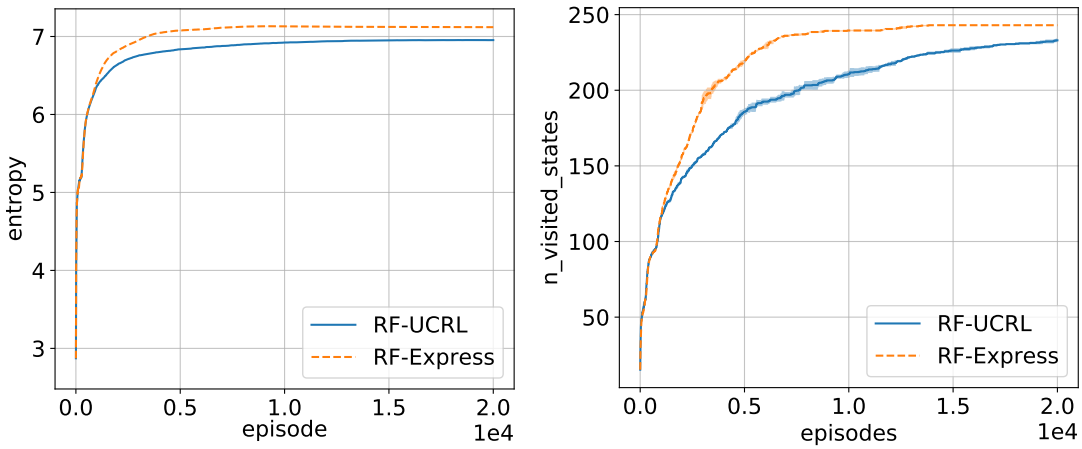

Figure 1. Entropy of the empirical distribution over states (left) and the number of visited states (right) versus number of episodes. Horizon $H = 30$, average over 4 runs.

The code to reproduce the experiments is available on GitHub, and uses the `rlberry` library (Domingues et al., 2021a).