Optimal Posterior Sampling for Policy Identification in Tabular Markov Decision Processes

Cyrille Kone
Univ. Lille, CNRS, Inria, Centrale Lille,
UMR 9189-CRIStAL, F-59000 Lille, France

Kevin Jamieson
University of Washington,
Seattle, USA

Abstract

We study the (ε, δ)-PAC policy identification problem in finite-horizon episodic Markov Decision Processes. Existing approaches provide finite-time guarantees for approximate settings ($\varepsilon > 0$) but suffer from high computational cost, rendering them hard to implement, and also suffer from suboptimal dependence on $\log(1/\delta)$. We propose a randomized and computationally efficient algorithm for best policy identification that combines posterior sampling with an online learning algorithm to guide exploration in the MDP. Our method achieves asymptotic optimality in sample complexity, also in terms of posterior contraction rate, and runs in $O(S^2AH)$ per episode, matching standard model-based approaches. Unlike prior algorithms such as MOCA and PEDEL, our guarantees remain meaningful in the asymptotic regime and avoid sub-optimal polynomial dependence on $\log(1/\delta)$. Our results provide both theoretical insights and practical tools for efficient policy identification in tabular MDPs.

1 INTRODUCTION

In online reinforcement learning (RL), an agent sequentially interacts with an unknown Markov decision process (MDP) by observing trajectories composed of transitions and rewards generated by the environment. We are interested in the setting where interactions occur in episodes of finite length H: an episodic finite-horizon setting (Puterman 1994). Put formally, the MDP is described by $M \triangleq (\mathcal{S}, \mathcal{A}, H, p, \mu, s_{\text{init}})$ where $\mathcal{S}$ is the state space of size S, $\mathcal{A}$ is the action space of size A, p and μ denote the transition and reward kernels, and s_{init} is the initial state.

Each episode starts in the initial state s_{init}. At stage $h \in [H]$, when in state $s_h \in \mathcal{S}$, the agent selects an action $a_h \in \mathcal{A}$ according to its strategy, collects a (random) reward r_h, and transitions to the (random) next state s_{h+1} according to the environment dynamics. After H steps, the environment resets to s_{init}. The agent must learn to efficiently explore the environment in order either to maximize cumulative rewards across episodes (Jaksch et al. 2010) or, after an exploration phase, to identify a "good" policy that maximizes the expected sum of rewards per episode (Fiechter 1994; Kearns & Singh 1998). In this work, we focus on the latter objective, studied in the celebrated (ε, δ)-PAC (Probably Approximately Correct) policy identification setting, where the agent explores adaptively and eventually stops to recommend a policy it believes to be ε-optimal. The goal is to keep the exploration phase as short as possible while guaranteeing the correctness of the recommendation.

In this framework, the agent follows adaptive exploratory policies and stops at a (random) time τ, when it recommends a candidate policy $\hat{\pi}_\tau$ believed to be optimal. The objective is to minimize the exploration time while ensuring PAC correctness. This setting is a pure exploration problem, since performance is evaluated only through the final recommended policy, not during the exploration phase.

An algorithm for (ε, δ)-PAC policy identification typically has three components : (i) an adaptive exploration policy ρ_t to gather new observations at episode t, (ii) a recommendation rule $\hat{\pi}_t$ representing the guess of the learning agent for the best policy and (iii) a stopping rule τ specifying when to terminate interaction with the environment.

Early work on (ε, δ)-PAC policy identification adapted algorithms from the regret-minimization setting by incorporating adaptive stopping rules, and established

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

worst-case guarantees on the sample complexity (Azar, Munos, et al. 2012; Kaufmann, Ménard, et al. 2021a; Menard et al. 2021). However, subsequent results (A. J. Wagenmaker, Simchowitz, et al. 2022; Tirinzoni, Al-Marjani, et al. 2023) proved such rates to be inherently suboptimal from a problem-dependent perspective. Recent research has therefore shifted toward identifying the exact complexity term governing policy identification. In this regime, racing-style elimination algorithms have been proposed (Tirinzoni, Al Marjani, et al. 2022; A. Wagenmaker & Jamieson 2022; A. J. Wagenmaker, Simchowitz, et al. 2022; Narang et al. 2024). These approaches often rely on sophisticated algorithmic designs or exhaustive enumeration of deterministic policies, which leads to significant computational challenges. Another line of work has introduced tracking-based algorithms (Al Marjani et al. 2021; Marjani & Proutiere 2021); while promising, these methods are also computationally costly. Importantly, none of these approaches achieves instance-dependent optimality when δ is small.

In this work, we revisit the problem of (ε, δ)-PAC policy identification in the tabular episodic setting with a focus on the instance-dependent complexity when δ is small. We propose an algorithm that relies on posterior sampling and avoids costly policy enumeration and optimization steps required by many prior methods.

1.1 Related Work

The earliest results on (ε, δ)-PAC policy identification focused on establishing worst-case sample complexity bounds, i.e., the dependence on S, A, H, ε, and δ (Fiechter 1994; Kearns & Singh 1998; Azar, Munos, et al. 2012; Azar, Munos, et al. 2013; Dann & Brunskill 2015; Sidford, Wang, Wu, L. Yang, et al. 2018; Sidford, Wang, Wu & Ye 2018; Agarwal et al. 2020; Menard et al. 2021). Notably, the most recent of these results matches the lower bound $O\left(\frac{SAH^3}{\varepsilon^2}\log(1/\delta)\right)$ proven by Domingues et al. 2021. The corresponding algorithms are often derived from optimistic methods originally designed for regret minimization (Azar, Osband, et al. 2017) or from online-to-batch and regret-to-PAC reductions (Jin, Allen-Zhu, et al. 2018; Tirinzoni, Al-Marjani, et al. 2023). While such worst-case guarantees provide a complete characterization of performance in the hardest environments, they do not capture the true, instance-dependent complexity of PAC policy identification. In particular, many MDP instances can be solved with far fewer samples than the worst-case bounds would suggest.

A series of works has recently attempted to characterize the true instance-dependent complexity of (ε, δ)-PAC policy identification. The seminal work of A. J. Wagenmaker, Simchowitz, et al. 2022 notably observed that algorithms relying on the optimism principle, and, more generally, any algorithm satisfying good regret properties, cannot be instance-optimal in the (ε, δ)-PAC setting for all MDPs. The authors then introduced MOCA, a racing-type algorithm with action elimination, and showed that, with high probability, its sample complexity is upper-bounded by $C_{\mathrm{MOCA}}(M, \varepsilon)\log(1/\delta) + \mathrm{poly}(\log(1/\delta), \log(1/\varepsilon), SAH)/\varepsilon$, where $C_{\mathrm{MOCA}}(M, \varepsilon)$ scales with $H^2 \sum_{h=1}^{H} \min_\rho \max_{s,a} \frac{1}{(\varepsilon \vee \Delta_h(s,a))^2 w_h^\rho(s,a)}$, and $w_h^\rho(s,a)$ is the probability that (s,a) is visited at step h when ρ is played during an episode. $C_{\mathrm{MOCA}}(M, \varepsilon)$ is therefore identified as the leading complexity term when ε is small or when the suboptimality gaps are small. However, due to second-order terms, it may cease to be the dominating term when δ is small (A. J. Wagenmaker, Simchowitz, et al. 2022).

A. Wagenmaker & Jamieson 2022 proposed PEDEL, a policy-elimination algorithm that estimates policy values using a sophisticated G-optimal design. For tabular MDPs the sample complexity of PEDEL scales with $C_{\mathrm{PEDEL}}(M, \varepsilon) \triangleq H^4 \sum_{h=1}^{H} \min_\rho \max_{\pi \in \Pi_{\mathrm{det}}} \sum_{s,a} \frac{w_h^\pi(s,a)^2}{w_h^\rho(s,a)(\varepsilon \vee \Delta(\pi))^2}$ where Π_{det} is the set of deterministic (Markovian) policies and $\Delta(\pi)$ is the policy gap. See Tirinzoni, Al-Marjani, et al. 2023 for further discussion of the different gaps in (ε, δ)-PAC policy identification. By design, PEDEL requires enumerating all deterministic policies, which for finite-horizon MDPs is of size A^{SH}. The identified complexities for PEDEL and MOCA are generally not comparable.

Al-Marjani et al. 2023a introduced an algorithm whose identified complexity ranges between MOCA and PEDEL but suffers from a highly suboptimal polynomial dependence on $\log(1/\delta)$.

Narang et al. 2024 improved upon PEDEL by estimating not the individual policy values but the differences in policy values with respect to a reference policy. However, like PEDEL, their algorithm still requires enumerating the set of deterministic policies.

Tirinzoni, Al Marjani, et al. 2022 introduced EPRL, a fully adaptive action-elimination algorithm for MDPs with deterministic transitions. The sample complexity of EPRL also scales with the sum of the inverse squared suboptimality gaps of suboptimal reachable actions. The authors proved a lower bound on the sample complexity of a (ε, δ)-PAC algorithm for deterministic MDPs, showing that EPRL is optimal up to $O(H^2)$ terms. Importantly, this optimality guarantee holds only in the deterministic setting and does not extend to stochastic MDPs.

Marjani & Proutiere 2021 studied the instance-dependent complexity of (ε, δ)-PAC policy identification with a *generative model* for $\varepsilon = 0$ and stated the optimal complexity as a solution to an intensive min–max program, similar to the techniques popularized in Garivier & Kaufmann 2016. The authors proposed an algorithm that solves a proxy of this problem. Applying the same technique with a *forward model*, Al Marjani et al. 2021 proposed MDP-NaS and proved a sample complexity asymptotically (as $\delta \to 0$) equivalent to $C(M)\log(1/\delta)$, where $C(M)$ is worse than $C_{\text{MOCA}}(M, 0)$.

To date, all existing algorithms for policy identification either suffer from computational intractability, exhibit a suboptimal dependence on the leading term involving $\log(1/\delta)$ as $\delta \to 0$, or use regret minimization algorithms for their sampling rules.

Posterior Sampling and Top-Two Methods. Posterior sampling has emerged as a powerful tool for exploration in bandits and MDPs. Algorithms such as Top-Two Thompson Sampling (Russo 2016; Jourdan et al. 2022; Z. Li et al. 2024) achieve optimal exploration in bandits by maintaining uncertainty over the true model. They appear to be statistically optimal, computationally efficient, and empirically competitive with other approaches. In RL, particularly for regret minimization, posterior sampling algorithms have become popular for designing efficient and tractable methods (Osband et al. 2013; Agrawal & Jia 2017; Russo 2019; Tiapkin et al. 2022). Yet, to the best of our knowledge, no posterior sampling algorithm has been analyzed for best policy identification in RL. We aim to close this gap by proposing an optimal and tractable algorithm for the episodic tabular setting. However, we emphasize that our goal is not merely to establish a sample complexity bound for an existing regret-minimization algorithm, as such algorithms are known to be suboptimal from an instance-dependent perspective (see above). Instead, we aim to design an instance-optimal algorithm that retains the computational efficiency of posterior sampling.

1.2 Main Contributions

In this paper, we focus on the theoretical analysis of the high-confidence regime (small δ) and the exact identification case ($\varepsilon = 0$):

- *Algorithmic framework.* We introduce PIPS, a top-two, posterior-sampling-based algorithm for PAC policy identification, which breaks the policy-enumeration bottleneck by using posterior-driven exploration coupled with an online learner.
- *Theoretical guarantees.* We establish optimal instance-dependent sample complexity bounds in the high-confidence and exact identification regimes. In addition, we prove sharp posterior contraction results, providing a refined characterization of how uncertainty about the optimal policy decreases with the number of episodes. We further discuss the extension of PIPS to the $\varepsilon > 0$ case for which a simple modification of PIPS is proposed.
- *Empirical validation.* We complement our theory with experiments showing that our algorithm outperforms existing baselines, both in terms of sample efficiency and computational scalability.

2 PROBLEM FORMULATION

We consider an episodic, finite-horizon Markov decision process (MDP) with finite state space $\mathcal{S}$ of size S and action space $\mathcal{A}$ of size A. The MDP is defined by the tuple $M \triangleq (\mathcal{S}, \mathcal{A}, \{p_h\}_{h=1}^H, \{R_h\}_{h=1}^H, s_{\text{init}})$ where p_h and R_h denote the transition and reward kernels at step h, and s_{init} is the initial state. A (Markovian) policy is a sequence $\pi = \{\pi_h\}_{h=1}^H$ with $\pi_h(\cdot|s)$ a distribution over actions given state $s \in \mathcal{S}$. Under policy π, at each stage $h \in [H]$ the agent draws $a_h \sim \pi_h(s_h)$, receives reward $r_h \sim R_h(s_h, a_h)$ with mean $\mu_h(s_h, a_h) = \mu(s, a, h)$, and transitions to $s_{h+1} \sim p_h(\cdot|s_h, a_h) \equiv p(\cdot \mid s, a, h)$. A policy is called *deterministic* if each $\pi_h(\cdot|s)$ assigns all probability mass to a single action denoted $\pi_h(s)$ or $\pi(s, h)$.

Learning Problem. Given the MDP M, for any policy $\pi = \{\pi_h\}_{h=1}^H$, the Q-function is defined at each state-action as $Q_h^\pi(s, a) \triangleq \mathbb{E}_M^\pi\left[\sum_{k=h}^H r_k | s_h = s, a_h = a\right]$, i.e., the expected accumulated reward when starting from (s, a) at step h and following π thereafter. The corresponding state-value function is $V_h^\pi(s) \triangleq \mathbb{E}_M^\pi\left[\sum_{k=h}^H r_k | s_h = s\right]$ and we write $V_0^\pi \triangleq V_1^\pi(s_{\text{init}})$ for the expected return when starting from the initial state. Let Π denote the set of all policies (deterministic and stochastic). The learner's goal is to identify an optimal policy $\pi^\star \in \operatorname{argmax}_{\pi \in \Pi} V_0^\pi$ using as few episodes as possible. For an MDP $\mathbf{M}$, we let $\Pi^\star(M)$ (resp. $\Pi^\star_{\text{det}}(M)$) denote the set of optimal (resp. deterministic optimal) policies, and drop the dependence on M when clear from context. Finally, we denote by $\{\mathcal{H}_t\}_{t \geqslant 1}$ the filtration of the learning process, i.e., the information collected up to the end of the t-th episode.

Assumption 1. *The MDP M admits a unique optimal policy $\pi^\star$.*

We assume normally distributed rewards, i.e. $R_h(s, a) \sim \mathcal{N}(\mu(s, a, h), 1)$, and we further assume

without loss of generality that the expected reward $\mu(s,a,h) \in (0,1)$. Any tabular MDP in this parametric set is uniquely identified by its transition probabilities $p \triangleq \{p_h\}_{h=1}^H$ and expected reward functions $\mu \triangleq \{\mu_h\}_{h=1}^H$. Maintaining a distribution over these parameters is therefore equivalent to maintaining a distribution over MDP models. In the sequel, when the context is clear, we identify an MDP by its transition probabilities p and expected rewards μ. We denote by $\mathbf{M}$ the set of such parametric MDPs.

Empirical MDP. For $t \geqslant 1$, let $n_t(s' \mid s,a,h) \triangleq \sum_{i=1}^t \mathbf{1}\big(s_h^i = s,\, a_h^i = a,\, s_{h+1}^i = s'\big)$ be the number of transitions to state s' after taking action a in state s at step h during the first t episodes. We also define $n_t(s,a,h) \triangleq \sum_{s'} n_t(s' \mid s,a,h) = \sum_{k=1}^{t-1} \mathbf{1}\big(s_h^k = s,\, a_h^k = a\big)$, the total number of visits to (s,a,h) up to the end of the t-th episode. The empirical transition probability at (s,a,h) is then given by

$$\hat{p}_t(s'|s,a,h) \triangleq \begin{cases} \dfrac{n_t(s' \mid s,a,h)}{n_t(s,a,h)}, & \text{if } n_t(s,a,h) > 0, \\ \dfrac{1}{S}, & \text{otherwise.} \end{cases}$$

We denote by $\hat{\mu}_t(s,a,h)$ the empirical mean reward at (s,a,h). The empirical reward and transition kernels $(\hat{p}_t, \hat{\mu}_t)$ define a unique MDP in $\mathbf{M}$, which we denote by $\hat{M}_t$. For any policy $\pi \in \Pi$, we write $(\hat{V}_{t,h}^\pi)$ and $(\hat{Q}_{t,h}^\pi)$ for the value and Q-functions of π in $\hat{M}_t$.

Prior and Posterior over MDPs. Our distribution over MDPs is defined over the parameters that uniquely determine the model. For transitions, independent Dirichlet priors are considered, which, by conjugacy, yield Dirichlet posteriors. For rewards, we assign independent uniform priors on $(0,1)$, leading to truncated normal posteriors. After observing an episode $\{(s_h^t, a_h^t, r_h^t, s_{h+1}^t)\}_{h=1}^H$, the posterior is updated via Bayes' rule. Direct computation gives the closed form

$$\Pr(p,\mu|\mathcal{H}_{t-1}) = \prod_{s,a,h} f_{s,a,h}(\mu_h(s,a)) g_{s,a,h}(p_h(\cdot|s,a)), \tag{1}$$

where $f_{s,a,h}$ is the density of density of the truncated normal $\mathcal{N}_{[0,1]}\big(\hat{\mu}_t(s,a,h), \frac{1}{n_t(s,a,h)}\big)$, and $g_{s,a,h}$ is the density of $\mathrm{Dir}\big(\boldsymbol{u} + n_t(\cdot \mid s,a,h)\big)$, the Dirichlet with parameter $+n_t(\cdot \mid s,a,h)$, for $\in \mathbb{R}_+^d$, that we will specify latter.

Our goal is not to develop a full Bayesian model; the posterior primarily serves as an algorithmic tool to guide exploration.

Information-theoretic Lower Bounds. For two probability distributions P, Q supported on a discrete space $\mathcal{S}$, the Kullback–Leibler (KL) divergence is $\mathrm{KL}(P \,\|\, Q) \triangleq \sum_{s \in \mathcal{S}} P(s) \log \frac{P(s)}{Q(s)}$. For two MDPs M, M', we say that M is absolutely continuous with respect to M' (denoted $M \ll M'$) if for every (s,a,h), $p_h(\cdot|s,a) \ll p'_h(\cdot|s,a)$ and $R_h(s,a) \ll R'_h(s,a)$. Given an MDP M, define $\mathrm{Alt}(M) \triangleq \big\{M' \in \mathbf{M} : M \ll M' \text{ and } \Pi^\star(M) \cap \Pi^\star(M') = \emptyset\big\}$, the set of alternative MDPs M' such that M is absolutely continuous with respect to M' but where no policy optimal in M remains optimal in M'. We denote by $(\mathrm{KL}(M \,\|\, M'))_{s,a,h}$ the KL divergence between the transition and reward kernels of M and M'. For each (s,a,h), introducing $M_{s,a,h} \triangleq R_h(s,a) \otimes p_h(\cdot|s,a)$,

$$\begin{aligned} \mathrm{KL}(M\|M')_{s,a,h} &= \mathrm{KL}\big(M_{s,a,h}\|M'_{s,a,h}\big) \\ &= \mathrm{KL}\big(R_{s,a,h}\|R'_{s,a,h}\big) + \mathrm{KL}\big(p_{s,a,h}\|p'_{s,a,h}\big), \end{aligned}$$

where $p_{s,a,h} \triangleq p_h(\cdot|s,a)$ and $R_{s,a,h} \triangleq R_h(s,a)$.

An algorithm is said to be δ-PAC for best policy identification (BPI) if its stopping time τ and recommendation rule $\hat{\pi}_\tau$ satisfy

$$\Pr_M\big(\tau < \infty,\ \hat{\pi}_\tau \neq \pi^\star\big) \leqslant \delta, \quad \forall M \in \mathbf{M}\,.$$

Let Ω_M denote the set of visitation probabilities induced by Markovian policies on M, $\Omega_M \triangleq \{\{w_h^\pi(s,a)\}_{s,a,h} : \pi \in \Pi\}$.

From the information-contraction principle and change-of-distribution techniques (as popularized in Garivier & Kaufmann 2016), the following holds.

Theorem 1. *Any δ-PAC algorithm with stopping time τ satisfies,*

$$\mathbb{E}_M[\tau] \geqslant \Gamma_M^{-1} \log \frac{1}{2.4\delta}, \quad \textit{where}$$

$$\Gamma_M \triangleq \sup_{w \in \Omega_M} \inf_{\widetilde{M} \in \mathrm{Alt}(M)} \left[\sum_{s,a,h} w_h(s,a)\, \mathrm{KL}(M\|\widetilde{M})_{(s,a,h)}\right].$$

We can express Γ_M as

$$\Gamma_M = \max_{\pi^{\mathrm{exp}}} \inf_{\widetilde{M} \in \mathrm{Alt}(M)} \mathbb{E}_M^{\pi^{\mathrm{exp}}} \left[\sum_{h=1}^H \mathrm{KL}(M\|\widetilde{M})_{s_h,a_h,h}\right], \tag{2}$$

where the expectation is taken over trajectories generated by policy ρ with dynamics governed by M and the KL terms act as deterministic rewards. The problem in Eq.(2) can be interpreted as the value of a two-player zero-sum game: the learner chooses an exploration policy π^{exp} to maximize the accumulated information (measured by KL terms), while an adversary selects an alternative MDP $\widetilde{M}$ (with different optimal policies than in M) to minimize it.

The quantity in Eq.(2) not only characterizes the sample complexity of BPI but also governs the contraction

rate of the posterior distribution of any adaptive algorithm for BPI.

Theorem 2. *Fix an MDP $M \in \mathbf{M}$ with optimal policy $\pi^\star$ and let ν_t denote the posterior distribution defined in Eq. (1). For any adaptive algorithm for best policy identification, with probability one,*

$$\limsup_{t\to\infty} -\tfrac{1}{t}\log \Pr_{\widetilde{M}\sim\nu_t \,|\, \mathcal{H}_{t-1}}\big(\pi^\star \notin \Pi^\star(\widetilde{M})\big) \;\leqslant\; \Gamma_M \,.$$

In words, for any adaptive algorithm, the posterior probability of misidentifying the optimal policy decays at most as $e^{-t\Gamma_M}$ asymptotically in t. While analogous results are known in the multi-armed bandit setting (Russo 2016), our result is novel in the context of tabular MDPs (see Appendix H for a full proof).

3 POLICY IDENTIFICATION VIA POSTERIOR SAMPLING

Our algorithm uses the posterior distribution to guide exploration in the environment. At episode t, the learner computes the optimal policy of the empirical MDP $\hat{M}_t$, denoted by $\hat{\pi}_t$, based on all data observed up to episode t. The core idea is then to construct a *challenger* MDP $\widetilde{M}_t$, sampled from an inflated posterior distribution, such that $\hat{\pi}_t$ is not optimal in $\widetilde{M}_t$. This challenger highlights uncertainty about the optimal policy and directs further exploration.

Formally, let $1/\eta_t$ be the inflation parameter. We define the density of the inflated posterior distribution over MDPs as

$$\nu_t^{\eta_t}(p,\mu) \;\triangleq\; \prod_{s,a,h} f^t_{s,a,h}(\mu(s,a,h))\, g^t_{s,a,h}(p(\cdot|s,a,h)) \,,$$

where $f^t_{s,a,h}$ is the density of inflated $N_{[0,1]}(\hat{\mu}_t(s,a,h), \frac{1}{\eta_t n_t(s,a,h)})$ and $g^t_{s,a,h}$ is the density of the inflated Dirichlet $\mathrm{Dir}(\boldsymbol{u} + \eta_t n_t(\cdot \mid s,a,h))$, and we set the prior hyperparameter to $\boldsymbol{u} = (1,\dots,1)$. The challenger MDP $\widetilde{M}_t$ is generated via conditional posterior sampling: we draw candidate MDPs $\widetilde{M}_{t,1}, \widetilde{M}_{t,2}, \dots$ *i.i.d* from $\nu_t^{\eta_t}$ and select the first one for which $\hat{\pi}_t$ is not optimal. Concretely, for each sampled MDP $\widetilde{M}_{t,k}$, we compute its optimal Q-function $\widetilde{Q}^{k,\star}_{t,h}$ by backward induction (Puterman 1994) and then check whether $\hat{\pi}_t$ is optimal for this Q-function. As soon as a sample is found where $\hat{\pi}_t$ is suboptimal, this instance is designated as the challenger MDP $\widetilde{M}_t$.

This sampling procedure naturally favors randomness in insufficiently explored state-action pairs, where deviations from the empirical model are most likely to produce a different optimal policy. Consequently, the challenger MDP $\widetilde{M}_t$ serves as an informative alternative to the current estimate, concentrating exploration on the state-action pairs that are most relevant for distinguishing optimal from suboptimal policies. Once the challenger is obtained, an online RL learner directs exploration toward the regions where the empirical model and the challenger differ the most, as measured by their KL divergence.

Exploration Policy. The learner maintains an online RL reward-maximization strategy over the space of exploration policies to ensure that sufficient evidence is collected to distinguish between the empirical MDP $\hat{M}$ and the challenger MDP $\widetilde{M}$. An initial exploration policy π_1^{exp} is given. At the end of each episode, the exploration policy is updated via a policy improvement step performed by a mirror descent learner. By letting $\widetilde{M}_t \triangleq (\theta_t, \Phi_t)$, we define the exploration reward kernel as

$$r_t^{\text{exp}}(s,a,h) \;\triangleq\; \frac{\big(\hat{\mu}_t(s,a,h) - \theta_t(s,a,h)\big)^2}{2} \;+\; \sum_{s'} \hat{p}_t(s' \mid s,a,h) \log \frac{\hat{p}_t(s' \mid s,a,h)}{\max\big(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\big)} \,. \tag{3}$$

which corresponds to the KL divergence between the marginals $\widetilde{M}_t(s,a,h)$ and $\hat{M}_t(s,a,h)$ with transition in $\widetilde{M}_t$ clipped to e^{-t^α} to avoid singularities and control the magnitude of the reward. An optimistic policy evaluation step estimates the Q-function of π_t^{exp} in the MDP with reward kernel r_t^{exp} and (unknown) transition kernel p (hence the use of optimism). Concretely, using Bernstein type confidence bonuses, and initializing $\bar{V}_{t,H+1}(\cdot) = 0$, the optimisitic Q-function is recursively evaluated as

$$\bar{Q}_{t,h}(s,a) \triangleq r_t^{\text{exp}}(s,a,h) \;+\; \hat{p}_{t,h}\bar{V}_{t,h+1}(s,a) \;+\; \frac{2 g_{t,h}\beta^p(n_t(s,a,h), 1/t^3)}{3\max(1, n_t(s,a,h))} \;+\; \sqrt{\frac{2\,\mathrm{Var}_{\hat{p}_{t,h}}\big[\bar{V}_{t,h+1}\big](s,a)\cdot\beta^p(n_t(s,a,h),\,1/t^3)}{\max(1,n_t(s,a,h))}} \,, \tag{4}$$

with the associated value function $\bar{V}_{t,h+1}(s) \triangleq \sum_a \bar{Q}_{t,h}(s,a)\rho_h(a|s)$ and $g_{t,h} \triangleq \max_s \bar{V}_{t,h+1}(s)$. We introduce $\hat{p}_{t,h}\bar{V}_{t,h+1}(s,a) \triangleq \mathbb{E}_{s'\sim\hat{p}_{t,h}(\cdot|s,a)}\big[\bar{V}_{t,h+1}(s')\big]$, and $\mathrm{Var}_{\hat{p}_{t,h}}\big[\bar{V}_{t,h+1}\big](s,a) = \mathrm{Var}_{s'\sim\hat{p}_{t,h}(\cdot|s,a)}(\bar{V}_{t,h+1}(s'))$. As proven in Appendix D, for some properly tuned threshold β^p, the Q-function computed in Eq.(4) is, with probability at least $1 - 1/t^3$, an optimistic estimate of $Q_{t,h}^{\pi_t^{\text{exp}}}(s,a)$, the Q-function of π_t^{exp} in (r_t^{exp}, p). Next, one step of policy improvement via an online learner on the

$\mathcal{A}$-simplex. We maintain SH independent online learners $(\mathcal{L}_{s,h})_{s,h}$ on the simplex $\triangle_A$ (distributions over actions).

The policy improvement step updates each online learner to improve the exploration policy for each (s,h) via the estimated exploration Q-function. Notable choices include AdaHedge and other instances of mirror descent learners. This is analyzed in Appendix D.

To correct for potential empirical bias in initial estimates of some state-action pairs, the exploration policy π_t^{exp} is mixed with a forced exploration policy c_t, which ensures that any reachable state will be explored sufficiently. Given some parameter $\gamma \in (0,1)$, at episode t, with probability $t^{-\gamma}$, the behaviour policy π_t^{exp} is replaced a forced exploration policy c_t (explained below). This ensures broad coverage while keeping the forced exploration negligible: the total number of episodes allocated is sublinear in t, and for γ close to 1, only $O(\log t)$ episodes are allocated to forced exploration, letting the exploration mainly be led by ComputeExplorationPolicy.

The forced exploration policy performs a single iteration update of UCBVI (Azar, Osband, et al. 2017) with a reward function designed as an indicator of a randomly picked state-action pair. A triplet $(\tilde{s}^t, \tilde{a}^t, \tilde{h}^t)$ is selected randomly and a reward kernel is defined as $r_t^{\text{f}} : (s,a,h) \mapsto \mathbf{1}\left(s = \tilde{s}^t, a = \tilde{a}^t, h = \tilde{h}^t\right)$ and c_{t+1} is defined as the greedy policy wrt an optimistic estimate of the Q-function for the reward kernel r_t^{f}. The procedure is described and analyzed in Appendix E.

In Appendix L, we describe a simple modification to construct the challenger MDP when the goal is to identify an ε-optimal policy (with $\varepsilon > 0$).

Computational Cost. The computational cost of Algorithm 1 is dominated by the sampling and planning steps performed at each episode. Concretely, the algorithm generates SAH Dirichlet samples of dimension S and SAH samples from a normal distribution. It then verifies whether the sampled MDP and the empirical MDP share the same optimal policy, which can be done via backward induction (Puterman 1994) in time $O(S^2AH)$. Altogether, the per-episode computational complexity is $O(S^2AH)$, with a memory requirement of $O(S^2AH)$, consistent with standard model-based approaches.

4 MAIN THEORETICAL RESULTS

In this section, we present the guarantees attained by PIPS both in terms of posterior contraction rate and sample complexity upper bound.

Algorithm 1: PIPS: Policy Identification via Posterior Sampling in Tabular MDPs

Require: Initial exploration $\pi_1^{\text{exp}} = \{\pi_1^{\text{exp}}(\cdot|s,h)\}_{s,h}$; initial forced exploration c_1; mixing parameter $\gamma \in (0,1)$; clipping parameter $\alpha \in (0, \frac{1}{2})$; posterior inflation $(\eta_t)_{t \geqslant 1}$; initial MDP $\hat{M}_0 \in \mathbf{M}$ (arbitrary)

for *episode* $t = 1, 2, \ldots$ **do**

// Empirical best policy

Compute $\hat{\pi}_t \in \arg\max_\pi \hat{V}_{t,0}^{\pi}$

// Computing a challenger MDP

for $k = 1, 2, \ldots$ **do**

Sample candidate challenger $\widetilde{M}_{t,k} \sim \nu_t^{\eta_t}$

Compute the optimal Q-function $\widetilde{Q}_t^{k,\star}$ of $\widetilde{M}_{t,k}$ by backward induction

if $\hat{\pi}_t$ *is suboptimal for* $\widetilde{Q}_t^{k,\star}$ **then**

Set challenger $\widetilde{M}_t \leftarrow \widetilde{M}_{t,k}$ and **break**

end

end

Compute c_t as in Algorithm 4

Draw $Z_t \sim \text{Bernoulli}(t^{-\gamma})$ and define $\forall (s,h) \in \mathcal{S} \times [H]$

$$\tilde{\pi}_t^{\text{exp}}(\cdot|s,h) \leftarrow \begin{cases} c_t(\cdot|s,h), & \text{if } Z_t = 1, \\ \pi_t^{\text{exp}}(\cdot|s,h), & \text{if } Z_t = 0, \end{cases}$$

// Rollout and data collection

Set $s_1^t \leftarrow s_{\text{init}}$

for $h = 1$ **to** H **do**

Sample action $a_h^t \sim \tilde{\pi}_t^{\text{exp}}(\cdot|s_h^t, h)$

Observe reward r_h^t and transition to s_{h+1}^t

end

// KL-based bonuses for the online learner

Construct bonus kernel $r_t^{\text{exp}}(s,a,h)$ from $\hat{M}_t$ and $\widetilde{M}_t$ as in Eq.(3)

// Behaviour policy improvement

Update exploration policy $\pi_{t+1}^{\text{exp}} \leftarrow$ ComputeExplorationPolicy$\left(t, \pi_t^{\text{exp}}, r_t^{\text{exp}}, \hat{p}_t, n_t\right)$

end

Our first result characterizes the posterior contraction rate. If we were to sample an MDP from the posterior the "Bayesian" error would be $\Pr_{\widetilde{M} \sim \nu_t | \mathcal{H}_{t-1}}(\pi^\star \notin \Pi^\star(\widetilde{M})) = \int_{\text{Alt}(M)} \mathrm{d}\nu_t((p,\mu))$ which by Theorem 2 is asymptotically larger than $e^{-t\Gamma_M}$ for any adaptive algorithm. The result below shows that PIPS reaches this optimal rate.

Algorithm 2: EXPLORATION POLICY HELPER FUNCTION

Function
ComputeExplorationPolicy$(t, \pi_t^{\exp}, r_t^{\exp}, \hat{p}_t, n_t)$
Require: Independent online learning instances $(\mathcal{L}_{s,h})_{s,h}$
Set $\bar{V}_{t,H+1}(s) \leftarrow 0 \quad \forall\, s \in \mathcal{S}$
for *episode stage* $h = H, H-1, \dots, 1$ **do**
  foreach $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    $\bar{Q}_{t,h}(s,a) \leftarrow$ Eq.(4)
  // Optimistic policy evaluation
  $\bar{V}_{t,h}(s) \leftarrow \sum_{a\in\mathcal{A}} \pi_t^{\exp}(a|s,h)\, \bar{Q}_{t,h}(s,a) \quad \forall\, s \in \mathcal{S}$
// Policy improvement via online Mirror descent learner
foreach $(s,h) \in \mathcal{S} \times [H]$ **do**
  Feed learner $\mathcal{L}_{s,h}$ with gain $\bar{Q}_{t,h}(s,\cdot)$
  Update $\pi_{t+1}^{\exp}(\cdot \mid s,h)$ from $\mathcal{L}_{s,h}$
return policy $\pi_{t+1}^{\exp}$

Theorem 3. *Consider an MDP $M \in \mathbf{M}$ with a unique optimal policy $\pi^\star$. Let $\gamma \in (0,1)$, $\alpha \in (0,1/2)$, $\alpha < \gamma$, $\varsigma \in (0,1)$, with $\varsigma > 2\alpha$. Run with parameterss γ, α and learning rate $\eta_t = O(t^{-\varsigma})$, letting ν_t denote the non-inflated posterior distribution, Algorithm 1 satisfies with probability one*

$$\limsup_{t\to\infty} -\frac{1}{t} \log \Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\big(\pi^\star \notin \Pi^\star(\widetilde{M})\big) \;=\; \Gamma_M.$$

Combined with Theorem 2, this result ensures that when running Algorithm 1, the posterior probability of mis-identifying the optimal policy decays exponentially fast with an optimal rate.

The next result we prove is in terms of sample complexity. We show that when equipped with a valid stopping procedure, Algorithm 1 attains the optimal sample complexity prescribed by Theorem 1.

Sample Complexity. PIPS can be coupled with a posterior-sampling-based stopping rule similar to the stopping rule used in Kone et al. 2025 for bandits. This stopping rule is based on the number of *i.i.d* samples collected from $\nu_t^{\eta_t}$ before finding a challenger MDP, i.e., by clipping the number of iterations in line ?? of Algorithm 1 by a threshold $B(t,\delta)$ to be defined.

Recalling that $\widetilde{M}_{t,1}, \widetilde{M}_{t,2}, \dots$ are sampled *i.i.d* from the posterior distribution $\nu_t^{\eta_t}$, the stopping time is formally defined as

$$\tau_\delta \triangleq \inf\Big\{ t \geqslant 1 : \forall\, 1 \leqslant k \leqslant B(t,\delta), \hat{\pi}_t \in \Pi^\star(\widetilde{M}_{t,k}) \Big\}. \quad (5)$$

The number of resampling is calibrated to ensure the δ-correctness of this stopping rule.

Theorem 4. *Let $\gamma \in (0,1)$, $\alpha \in (0,1/2)$, $\alpha < \gamma$, $\varsigma \in (0,1)$, with $\varsigma > 2\alpha$. If $B(t,\delta)$ satisfies $\limsup_{\delta\to 0} \frac{\log B(t,\delta)}{\eta_t \log\frac{1}{\delta}} \leqslant 1$, then, PIPS coupled with the stopping time τ_δ described in (5) and run with parameterss γ, α and learning rate $\eta_t = O(t^{-\varsigma})$, satisfies*

$$\mathbb{E}_M[\tau_\delta] \leqslant \Gamma_M^{-1} \log\frac{1}{\delta} + o\big(\log\tfrac{1}{\delta}\big)\,, \text{ and}$$
$$\Pr_M\Big(\limsup_{\delta\to 0} \frac{\tau_\delta}{\log\frac{1}{\delta}} \leqslant \Gamma_M^{-1} \Big) = 1\,.$$

Theorem 4 provides a sufficient condition on $B(t,\delta)$ to guarantee optimal sample complexity, but it does not ensure δ-correctness with this threshold. The lemma below provides the correctness guarantee when δ is small.

Lemma 5. *Let β be a threshold such that for any $\delta \in (0,1)$, the event*

$$\mathcal{E}_\delta \triangleq \left\{ \forall t \geqslant 1, \sum_{x\in\mathcal{X}} n_t(x)\, \mathrm{KL}(\hat{M}_t \,\|\, M)_x \leqslant \beta(t,\delta) \right\}$$

holds with probability at least $1-\delta$, where $R_h(s,a) \triangleq \mathcal{N}(\mu(s,a,1),1)$, $\hat{R}_h^t(s,a) \triangleq \mathcal{N}(\hat{\mu}_t(s,a,h), 1/n_t(s,a,h))$ and $\mathrm{KL}(\hat{M}_t \,\|\, M)_{s,a,h} \triangleq \mathrm{KL}(\hat{R}_h^t(s,a) \otimes \hat{p}_t(\cdot|s,a,h) \,\|\, R_h(s,a) \otimes p(\cdot|s,a,h))$. x denotes a generic s,a,h and $\mathcal{X} = \mathcal{S}\times\mathcal{A}\times[H]$. For $B(t,\delta) = O(\exp(\eta_t\beta(t,\delta))\log(t/\delta))$ the posterior stopping time defined in Eq.(5) satisfies $\limsup_{\delta\to 0} \Pr_M(\tau_\delta < \infty, \hat{\pi}_{\tau_\delta} \neq \pi^\star)/\delta \leqslant 1$.

Kaufmann & Koolen 2021 prescribes correct thresholds $\beta(t,\delta) = \log(1/\delta) + O(\log(t))$. Thus $B(t,\delta)$ as defined in Lemma 5 satisfies $\limsup_{\delta\to 0} \frac{\log B(t,\delta)}{\eta_t \log\frac{1}{\delta}} \leqslant 1$, so that Theorem 4 applies and PIPS is asymptically optimal and correct with the value of $B(t,\delta)$ prescribed by Lemma 5.

5 EXPERIMENTS

In this section, we evaluate our algorithm against a range of competitors. We examine two performance metrics : (i) the correct identification rate, defined as 1 if the recommended policy is optimal and 0 otherwise, and (ii) the performance factor $1 - \frac{V_0^\star - V^{\hat{\pi}^t}}{V_0^\star}$. The average identification rate reflects the fraction of times the recommended policy is optimal, while the performance factor measures the value of the recommended policy relative to the optimal value.

To ensure fairness, all algorithms are run without a stopping rule. Each experiment is repeated 50 times, and we report the average value of both metrics.

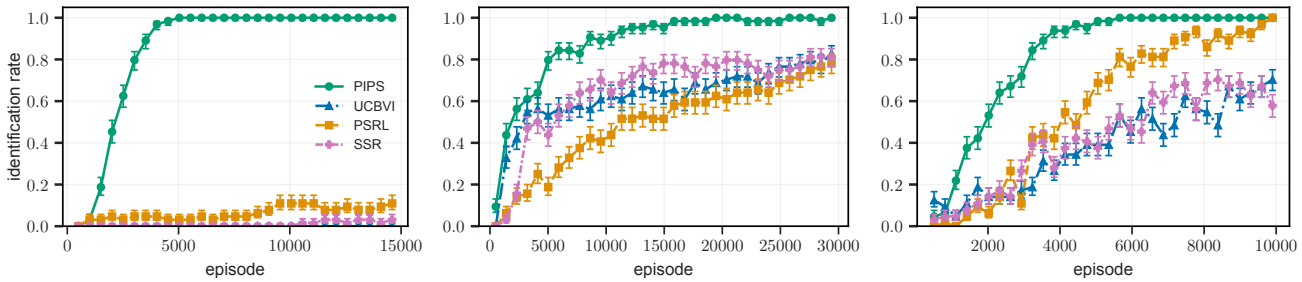

Figure 1: Empirical identification rate averaged on 64 runs with 95% Wilson's confidence interval. From left to right: MOCA instance, RiverSwim, CombinationLock.

Baselines. Algorithm 1, is compared to several established baselines. We evaluate BPI-UCBVI (Menard et al. 2021). Its sampling rule corresponds to the greedy policy with respect to an optimistic estimate of the optimal Q-function. Another baseline is PSRL (Osband et al. 2013), a randomized exploration algorithm that maintains a posterior distribution over MDPs and acts greedily wrt a sampled MDP, and SSR (Xiong et al. 2022), a variant of PSRL with distributions over Q-functions.

5.1 Environments

We evaluate the algorithms on challenging environments that require efficient exploration.

MOCA. The first environment, adapted from A. J. Wagenmaker, Simchowitz, et al. 2022 and shown in Fig. 2, contains $(L+1)$ states and $(L+1)$ actions. Among them, $a^\star$ is the optimal action at each stage but transitions to the next states with small probability. In contrast, the other actions $a_1, \dots, a_L$ are suboptimal, but taking a_l in s_{init} deterministically leads to state s_l. This environment poses a subtle challenge: identifying the optimal action in states $s_1, \dots, s_L$ may require first taking suboptimal actions in s_{init}. As a result, it is particularly difficult for regret-minimizing algorithms applied to BPI, since they tend to avoid actions with poor short-term performance.

We also evaluate on the classic RiverSwim MDP (see e.g. Strehl & Littman 2008), and the CombinationLock MDP (see e.g. Zhang et al. 2022).

RiverSwim. In the RiverSwim environment, the optimal policy always selects the `RIGHT` action to reach the high-reward state at the end of the chain, despite the lower immediate rewards along the way. We use a chain of length $L = 8$, set the horizon to $H = 10$, and take $s_{\text{init}} = s_1$. The rewards are Gaussian with variance $1/1000$. The environment is described in Appendix M.2, Fig.8.

Combination Lock. We consider a unichain, episodic combination-lock MDP with $S = H$ states $s_1, \dots, s_H$ and an action set of size $A \geqslant 2$. Each non-terminal state s_l $(l = 1, \dots, H-1)$ has a unique optimal action denoted $a_l^\star$. If the agent selects the optimal action $a_l^\star$ in state s_l, it transitions to s_{l+1} with probability $1-\varepsilon$ and returns to the initial state s_1 with probability ε. Any suboptimal action $a \neq a_l^\star$ at s_l sends the agent back to s_1 with probability 1. The terminal state s_H is absorbing, and the agent receives a reward of 1 only upon reaching s_H (i.e., after taking the optimal action at every step); all other transitions yield a reward of 0. In our experiments, we set $H = 10$, $A = 3$, and $\varepsilon = 0.01$.

5.2 Summary

Fig. 1 plots the identification rate vs the number of episodes for the three MDPs. The most significant improvement of PIPS over its competitors is achieved in the MOCA MDP. This is expected as in this MDP, optimistic algorithms essentially explore by following the policy with the highest value, which, as described in Fig. 2, only reach some states with probability exponentially small in S.

Fig. 4 shows the performance factor results. Additional experiments, presented in Appendix M.2, Fig. 5,6, 7, compare the posterior contraction rates of PIPS and PSRL on the MOCA instance, highlighting the significant advantage of PIPS over its competitors. Further implementation details are provided in Appendix M.1.

6 DISCUSSION AND OPEN PROBLEMS

We have studied the problem of best policy identification in finite-horizon episodic tabular MDPs. In this work, we introduced a computationally efficient algorithm that leverages posterior sampling combined with an online learning algorithm to explore the MDP and iteratively solve the optimization problem characterizing the optimal lower bound for best policy identification. Our analysis establishes optimal asymptotic

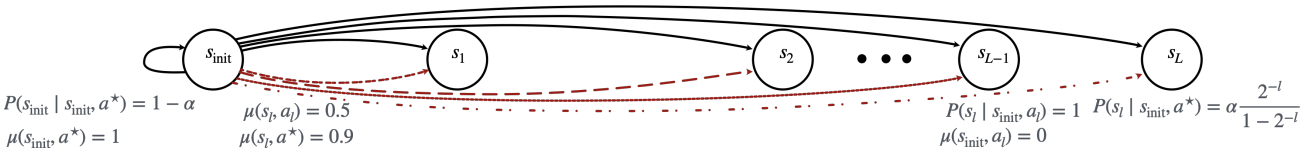

Figure 2: Custom MOCA environment where reaching the optimal policy requires exploratory steps involving sub-optimal actions $a_1, \ldots, a_L$ (red dotted lines).

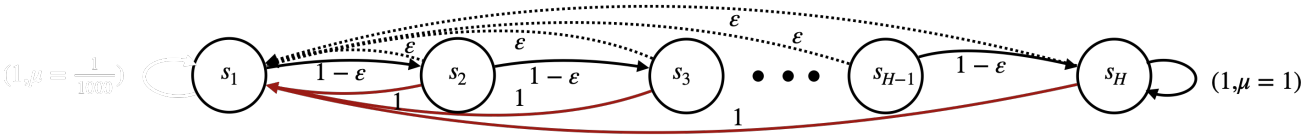

Figure 3: The H-state CombinationLock environment. Each nonterminal state offers A actions. The unique optimal action at state s_l advances the agent to s_{l+1} with probability $1-\varepsilon$ (solid black arrow) and returns it to the initial state s_{init} with probability ε (dashed arrow); any suboptimal action (red arrow) sends the agent back to s_{init} with probability 1. The terminal state s_H is absorbing and yields a reward 1 (all other transitions give a reward 0). Transition probabilities are annotated on the corresponding arrows.

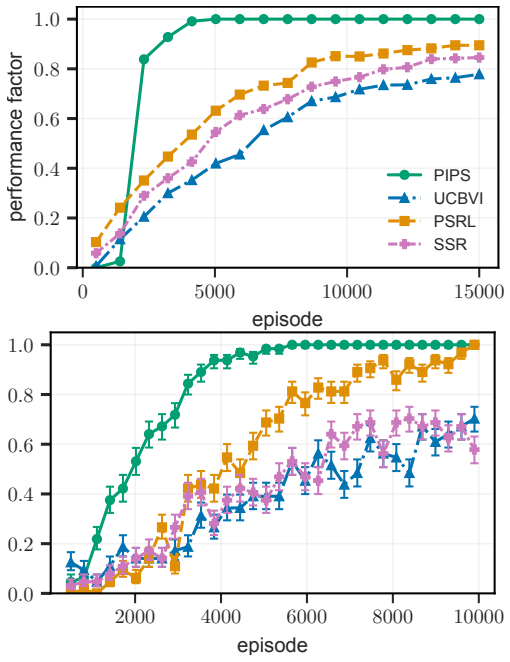

Figure 4: Performance factor averaged on 64 runs with 95% Wilson's confidence interval. MOCA instance (top), and CombinationLock (bottom).

guarantees on posterior contraction and sample complexity when the algorithm is equipped with a δ-PAC-calibrated threshold for the posterior stopping rule.

Despite its computational efficiency, our algorithm still requires sampling $O(S^2AH)$ parameters per round to generate a challenger MDP, which can be a bottleneck in tabular MDPs with large state spaces. Furthermore, while its theoretical guarantees are tight in the asymptotic regime, they do not provide explicit dependence on the state, action, or horizon parameters, which may not be negligible in the non-asymptotic setting, particularly when $\delta = \Omega(1)$. Obtaining tight finite-sample guarantees in the moderate confidence regime remains an open challenge as current algorithms scale with diverse and often incomparable complexity terms, including substantial "second-order" terms, such as those polynomial in $\log(1/\delta)$ (A. Wagenmaker & Jamieson 2022; A. J. Wagenmaker, Simchowitz, et al. 2022; Al-Marjani et al. 2023a; Narang et al. 2024).

This work can be extended in several promising directions. For instance, one could consider extensions to the linear function approximation setting (Jin, Z. Yang, et al. 2020), or to the (ε, δ)-PAC framework, where the objective is to identify an ε-optimal policy for some $\varepsilon > 0$. We briefly discuss an extension of PIPS to the (ε, δ)-PAC setting in Appendix L, although the resulting guarantees are sub-optimal. Extending our main results to this framework remains an open question.

Acknowledgements

KJ was funded in part by NSF Awards 2141511, 2023239. This work was done while CK was visiting KJ at the University of Washington.

References

Fiechter, C.-N. (1994). "Efficient reinforcement learning". *Proceedings of the Seventh Annual Conference on Computational Learning Theory*. COLT '94. Association for Computing Machinery.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1st. John Wiley & Sons, Inc.

Kearns, M. & S. Singh (1998). "Finite-Sample Convergence Rates for Q-Learning and Indirect Algorithms". *Advances in Neural Information Processing Systems*. Vol. 11. MIT Press.

Strehl, A. L. & M. L. Littman (2008). "An analysis of model-based Interval Estimation for Markov Decision Processes". *Journal of Computer and System Sciences* 74.

Lu, D. & W. V. Li (2009). "A note on multivariate Gaussian estimates". *Journal of Mathematical Analysis and Applications*.

Jaksch, T., R. Ortner & P. Auer (2010). "Near-optimal Regret Bounds for Reinforcement Learning". *Journal of Machine Learning Research* 11.

Azar, M. G., R. Munos & H. J. Kappen (2012). "On the sample complexity of reinforcement learning with a generative model". *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Omnipress.

– (2013). "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model". *Mach. Learn.* 91.

Osband, I., D. Russo & B. V. Roy (2013). *(More) Efficient Reinforcement Learning via Posterior Sampling*.

De Rooij, S., T. Van Erven, P. D. Grünwald & W. M. Koolen (2014). "Follow the leader if you can, hedge if you must". *J. Mach. Learn. Res.*

Dann, C. & E. Brunskill (2015). "Sample complexity of episodic fixed-horizon reinforcement learning". *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. MIT Press.

Krichene, W., M. Balandat, C. Tomlin & A. Bayen (2015). "The Hedge Algorithm on a Continuum". *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. PMLR.

Garivier, A. & E. Kaufmann (2016). "Optimal Best Arm Identification with Fixed Confidence". *29th Annual Conference on Learning Theory*. Vol. 49. Proceedings of Machine Learning Research. PMLR.

Russo, D. (2016). "Simple Bayesian Algorithms for Best Arm Identification". *29th Annual Conference on Learning Theory*. PMLR.

Agrawal, S. & R. Jia (2017). "Optimistic posterior sampling for reinforcement learning: worst-case regret bounds". *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Azar, M. G., I. Osband & R. Munos (2017). "Minimax Regret Bounds for Reinforcement Learning". *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR.

Dann, C., T. Lattimore & E. Brunskill (2017). "Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning". *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Yang, Z.-H., W. Qian, Y. Chu & W. Zhang (2017). "On Rational Bounds for the Gamma Function". *Journal of Inequalities and Applications* 2017.

Jin, C., Z. Allen-Zhu, S. Bubeck & M. I. Jordan (2018). "Is Q-Learning Provably Efficient?" *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc.

Sidford, A., M. Wang, X. Wu, L. Yang & Y. Ye (2018). "Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model". *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc.

Sidford, A., M. Wang, X. Wu & Y. Ye (2018). "Variance reduced value iteration and faster algorithms for solving markov decision processes". *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '18. Society for Industrial and Applied Mathematics.

Russo, D. (2019). "Worst-Case Regret Bounds for Exploration via Randomized Value Functions". *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Agarwal, A., S. Kakade & L. F. Yang (2020). "Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal". *Proceedings of Thirty Third Conference on Learning Theory*. Vol. 125. Proceedings of Machine Learning Research. PMLR.

Jin, C., Z. Yang, Z. Wang & M. I. Jordan (2020). "Provably efficient reinforcement learning with linear function approximation". *Proceedings of Thirty Third Conference on Learning Theory*. Vol. 125. Proceedings of Machine Learning Research. PMLR.

Al Marjani, A., A. Garivier & A. Proutiere (2021). "Navigating to the Best Policy in Markov Decision Processes". *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc.

Domingues, O. D., P. Ménard, E. Kaufmann & M. Valko (2021). "Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited". *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*. Vol. 132. Proceedings of Machine Learning Research. PMLR.

Kaufmann, E. & W. M. Koolen (2021). "Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals". *Journal of Machine Learning Research* 22.

Kaufmann, E., P. Ménard, O. Darwiche Domingues, A. Jonsson, E. Leurent & M. Valko (2021a). "Adaptive

Reward-Free Exploration". *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*. Proceedings of Machine Learning Research. PMLR.

– (2021b). "Adaptive Reward-Free Exploration". *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*. Vol. 132. Proceedings of Machine Learning Research. PMLR.

Marjani, A. A. & A. Proutiere (2021). "Adaptive Sampling for Best Policy Identification in Markov Decision Processes". *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR.

Menard, P., O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent & M. Valko (2021). "Fast active learning for pure exploration in reinforcement learning". *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR.

Jourdan, M., R. Degenne, D. Baudry, R. de Heide & E. Kaufmann (2022). "Top Two Algorithms Revisited". *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc.

Tiapkin, D., D. Belomestny, D. Calandriello, E. Moulines, R. Munos, A. Naumov, M. Rowland, M. Valko & P. Ménard (2022). "Optimistic Posterior Sampling for Reinforcement Learning with Few Samples and Tight Guarantees". *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc.

Tirinzoni, A., A. Al Marjani & E. Kaufmann (2022). "Near Instance-Optimal PAC Reinforcement Learning for Deterministic MDPs". *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc.

Wagenmaker, A. & K. G. Jamieson (2022). "Instance-Dependent Near-Optimal Policy Identification in Linear MDPs via Online Experiment Design". *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc.

Wagenmaker, A. J., Y. Chen, M. Simchowitz, S. Du & K. Jamieson (2022). "Reward-Free RL is No Harder Than Reward-Aware RL in Linear Markov Decision Processes". *Proceedings of the 39th International Conference on Machine Learning*. Proceedings of Machine Learning Research.

Wagenmaker, A. J., M. Simchowitz & K. Jamieson (2022). "Beyond No Regret: Instance-Dependent PAC Reinforcement Learning". *Proceedings of Thirty Fifth Conference on Learning Theory*. Vol. 178. Proceedings of Machine Learning Research. PMLR.

Xiong, Z., R. Shen, Q. Cui, M. Fazel & S. S. Du (2022). "Near-Optimal Randomized Exploration for Tabular Markov Decision Processes". *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc.

Zhang, X., Y. Song, M. Uehara, M. Wang, A. Agarwal & W. Sun (2022). "Efficient Reinforcement Learning in Block MDPs: A Model-free Representation Learning approach". *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR.

Al-Marjani, A., A. Tirinzoni & E. Kaufmann (2023a). "Active Coverage for PAC Reinforcement Learning". *Proceedings of Thirty Sixth Conference on Learning Theory*. Proceedings of Machine Learning Research. PMLR.

– (2023b). *Towards Instance-Optimality in Online PAC Reinforcement Learning*.

Tirinzoni, A., A. Al-Marjani & E. Kaufmann (2023). "Optimistic PAC Reinforcement Learning: the Instance-Dependent View". *Proceedings of The 34th International Conference on Algorithmic Learning Theory*. Vol. 201. Proceedings of Machine Learning Research. PMLR.

Li, Z., K. Jamieson & L. Jain (2024). "Optimal Exploration is no harder than Thompson Sampling". *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. PMLR.

Narang, A., A. Wagenmaker, L. J. Ratliff & K. Jamieson (2024). "Sample Complexity Reduction via Policy Difference Estimation in Tabular Reinforcement Learning". *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc.

Kone, C., M. Jourdan & E. Kaufmann (2025). "Pareto Set Identification With Posterior Sampling". *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*. Vol. 258. Proceedings of Machine Learning Research. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**] All claims are proven in the Appendix.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**] The complexity is analyzed in a paragraph in the main, and all sample complexity claims are proven in the Appendix.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Not Applicable**]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [**Yes**] Statements made in the main are proven in the Appendix
(b) Complete proofs of all theoretical results. [**Yes**] The results are proven in the Appendix
(c) Clear explanations of any assumptions. [**Yes**]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**] Section M describes the experimental setup and other implementation details for reproducibility.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**] Described in Section M
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**] Described in Section 5
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**] Described in Section M

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Not Applicable**]
   (b) The license information of the assets, if applicable. [**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]
   (d) Information about consent from data providers/curators. [**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

Contents

A EXTENDED NOTATION TABLE

We recall the notation used in the main and introduce some notation used in the appendix. In what follows, M is always fixed, and its optimal policy is also fixed. In general, $\hat{\cdot}$ denotes empirical quantities, whereas $\widetilde{\cdot}$ denotes quantities and variables related to the posterior distribution.

Table 1: Additional notations used in the main and in the appendix

Symbol	Definition
M	$(\mathcal{S}, \mathcal{A}, H, P, \mu, s_{\text{init}})$
$\hat{M}_t$	$(\mathcal{S}, \mathcal{A}, H, \hat{p}_t, \hat{\mu}_t, s_{\text{init}})$
$\pi^\star$	Unique (deterministic) optimal policy of M
$\triangle$	$\mathcal{S}$-Probability simplex
$\mathbf{M}$	Set of parameters of MDPs : $\{(0,1) \times \triangle\}^{SAH}$
$\mathbf{M}^\varepsilon$	$\{(r,q) \in \mathbf{M} : \forall (s,a,h,s')\ , q(s' \mid s,a,h) \geqslant \varepsilon\}$
$\mathcal{R}$	Expected reward kernel space: $(0,1)^{SAH}$
$\mathcal{P}$	Transition kernel space: $\triangle^{SAH}$
A^π	$\left\{(r,q) \in \mathbf{M} : V_0^\pi(r,q) \geqslant V_0^{\pi^\star}(r,q)\right\}$
$A^\pi(q)$	$\{q : \exists r \in \mathcal{R} \text{ such that } (r,q) \in A^\pi\}$
Π_{det}	Deterministic Markovian policies
$(\eta_t)_t$	Learning rate or posterior inflation rate
π_t^{exp}	Behavioral or exploration policy at episode t
c_t	Forced exploration or coverage policy at episode t
$n_t(s,a,t)$	Number of pulls of (s,a) up to the end of episode t
$n_t(s' \mid s,a,h)$	Number of transitions (s,a,s') at stage h
$\hat{\mu}_t(s,a,h)$	Empirical reward at (s,a,h) based upon the first t episodes
$\Pi^\star(M)$	Optimal start-state policies of M
$\Pi^\star_{\text{det}}(M)$	Optimal deterministic policies of M
$\text{Alt}(M)$	Alternative models : $\left\{M' \in \mathbf{M} : M \ll M' \text{ and } \Pi^\star(M) \cap \Pi^\star(M') = \emptyset\right\}$
$\nu_{\eta_t}^t(\cdot \mid \text{Alt}(\hat{M}_t))$	Truncated and inflated posterior at the start episode t
$\left\{(s_h^t, a_h^t, r_h^t, s_{h+1}^t)\right\}_{h=1}^H$	Trajectory collected during episode t
$\hat{p}_t(\cdot \mid s,a,h) \triangleq \hat{p}_t(\cdot \mid s,a,h)$	Empirical transition probabilities at (s,a,h) during episode t
$W_h(s,a)$	Maximum visitation probability of (s,a,h): $\max_\rho w_h^\rho(s,a)$

Value gaps. We recall that a policy π is *globally optimal* if its support at state and stage is included in the argmax set of the optimal Q-function. For a policy π, we define the *gap* at (s,a,h) as

$$\Delta_h^\pi(s,a) \ = \ V_h^\pi(s) - Q_h^\pi(s,a).$$

Note that, if π is globally optimal then the above gap is called the *value gap* and quantifies the suboptimality of action a in state s at stage h relative to the optimal action. We say that an MDP admits a *strongly globally policy* if, for every (s,h), the argmax set of the optimal Q-function is a singleton : $|\operatorname{argmax}_{a \in \mathcal{A}} Q_h^\star(s,a)| = 1$ i.e., there is a unique optimal action at each state and stage. If in addition, the start-state optimal policy is unique, then it must coincide with the greedy policy induced by $Q^\star$. In this case, we say the MDP admits a *unique start-state optimal policy*, which is also strongly optimal.

Finally, writing $a^\star \triangleq \operatorname{argmax}_a Q_h^\star(s,a)$, we define at (s,h), $\Delta_h^\pi(s,a^\star) = \min_{a \neq a^\star} \Delta_h(s,a)$ and we define the *minimal value gap* as

$$\Delta_{\min}^\pi \ \triangleq \ \min_{s,a,h} \Delta_h^\pi(s,a)\,, \tag{A.1}$$

which is postive when π is the strongly globally optimal policy in M. When the MDP is not clear from the context, we write $\Delta_{\min}^\pi(M)$.

Additional notation. Given a (Markov) policy ρ, we denote by w^ρ the visitation probabilities induced on the MDP, i.e., for any s, a, h, $w_h^\rho(s,a)$ is the probability that the state-action (s,a) is visited at episode h when ρ is followed.

The reward space is $\mathcal{R}$ and the transitions space is $\mathcal{P}$. We have $\mathbf{M} = \mathcal{R} \otimes \mathcal{P}$.

Since our results are targeting the first-order asymptotic bound, we let T_0 denote an arbitrary constant that may appear multiple times in our calculations.

$O(f(t))$ denotes an asymptotic upper bound when $t \to \infty$, hiding constants, and $g(t) = o(f(t))$ if and only if $g(t)/f(t) \to 0$ when $t \to \infty$.

Finally, we define the following ℓ_1-distance on $\mathbf{M}$,

$$\|M - M'\|_1 = \sum_{s,a,h} \big|\mu(s,a,h) - \mu'(s,a,h)\big| + \|p(\cdot|s,a,h) - p'(\cdot|s,a,h)\|_1, \quad \forall\, M, M' \in \mathbf{M}\,.$$

B SAMPLE COMPLEXITY LOWER BOUND

We now state a lower bound on the sample complexity of any δ-PAC algorithm for best policy identification, derived from the celebrated information-contraction principle.

Let

$$\mathrm{Alt}(M) \triangleq \big\{M' \in \mathbf{M} : M \ll M' \quad \text{and} \quad \Pi^\star(M) \cap \Pi^\star(M') = \emptyset\big\},$$

that is, the set of MDPs M', so that M is absolutely continuous with respect to M' but have no optimal policy in common with M.

At episode t, an adaptive algorithm selects an exploration policy ρ_t based on past observations. Executing ρ_t in the environment generates the trajectory

$$\{s_h^t, a_h^t, r_h^t, s_{h+1}^t\}_{h=1}^H, \quad s_1^t = s_{\mathrm{init}}, \quad a_h^t \mid s_h^t \sim \pi_t^{\exp}(\cdot \mid s_h^t, h),$$

with transitions $s_{h+1}^t \mid s_h^t, a_h^t \sim p(\cdot \mid s_h^t, a_h^t, h)$ and rewards $r_h^t \mid s_h^t, a_h^t \sim R_h(s_h^t, a_h^t)$. Let $\mathcal{H}_t$ denote the σ-algebra generated by all observations and exploration policies up to the end of episode t, including external randomness $(U_1, \dots, U_t)$ used for tie-breaking and randomization. For two MDPs

$$M = (\mathcal{S}, \mathcal{A}, H, p, \mu, s_{\mathrm{init}}), \quad \widetilde{M} = (\mathcal{S}, \mathcal{A}, H, \tilde{p}, \tilde{\mu}, s_{\mathrm{init}}),$$

we define the KL divergence between their kernels at (s,a,h) as

$$\mathrm{KL}(M \,\|\, \widetilde{M})_{s,a,h} \triangleq \mathrm{KL}\Big(R_h(s,a) \otimes p(\cdot \mid s,a,h) \,\Big\|\, \widetilde{R}_h(s,a) \otimes \widetilde{p}(\cdot \mid s,a,h)\Big).$$

An algorithm is said to be δ-PAC if, for any $M \in \mathbf{M}$, its stopping time τ and recommendation $\hat{\pi}^\tau$ satisfy

$$\Pr_M\big(\tau < \infty,\ \hat{\pi}_\tau \notin \Pi^\star(M)\big) \leqslant \delta.$$

Using the data-processing inequality and change-of-distribution arguments (see Garivier & Kaufmann 2016), we obtain the following fundamental bound:

Lemma 6. *Consider an MDP M with a unique optimal policy. Then, for any δ-PAC algorithm with stopping time τ, it holds that*

$$\sum_{s,a,h} \mathbb{E}_M\left[n_{\tau+1}(s,a,h)\right] \mathrm{KL}(M \,\|\, \widetilde{M})_{s,a,h} \geqslant \log \frac{1}{2.4\delta}, \qquad \forall\, \widetilde{M} \in \mathrm{Alt}(M).$$

Using this result, we prove the sample complexity lower bound. We recall that Ω_M denotes the set of visitation probabilities induced by Markovian policies on M

$$\Omega_M \ \triangleq\ \big\{\{w_h(s,a)\}_{s,a,h} : \exists\, \pi \in \Pi \ \text{ such that } \ w_h(s,a) = \Pr_M^\pi(s_h = s, a_h = a),\ \forall (s,a,h)\big\}. \tag{B.1}$$

Theorem 1. *Any δ-PAC algorithm with stopping time τ satisfies,*

$$\mathbb{E}_M[\tau] \geqslant \Gamma_M^{-1} \log \frac{1}{2.4\delta}, \quad \textit{where}$$
$$\Gamma_M \triangleq \sup_{w \in \Omega_M} \inf_{\widetilde{M} \in \mathrm{Alt}(M)} \left[\sum_{s,a,h} w_h(s,a) \, \mathrm{KL}(M \| \widetilde{M})_{(s,a,h)} \right].$$

Proof. Fix M with a unique optimal policy and let τ be the stopping time of any δ-PAC algorithm. By Lemma 6, for every $\widetilde{M} \in \mathrm{Alt}(M)$,

$$\sum_{s,a,h} \mathbb{E}_M[n_{\tau+1}(s,a,h)] \, \mathrm{KL}(M \, \| \, \widetilde{M})_{s,a,h} \; \geqslant \; \log \frac{1}{2.4\delta} \, . \tag{B.2}$$

We recall that ρ^t is predictable (i.e., measurable wrt $\mathcal{H}_{t-1}$). Therefore

$$\begin{aligned} \mathbb{E}_M[n_{\tau+1}(s,a,h)] &= \mathbb{E}_M \left[\sum_{t \geqslant 1} \mathbf{1}(\tau \geqslant t) \cdot \mathbf{1}\big(s_h^t = s, a_h^t = a\big) \right] \\ &= \mathbb{E}_M \left[\mathbb{E} \left[\sum_{t \geqslant 1} \mathbf{1}(\tau \geqslant t) \cdot \mathbf{1}\big(s_h^t = s, a_h^t = a\big) \middle| \mathcal{H}_{t-1} \right] \right] \end{aligned}$$

which follows from the tower rule of expectations. Next observe that the event $\mathbf{1}(t \leqslant \tau)$ is the complementary of the event $\mathbf{1}(\tau < t)$ which is the same as $\mathbf{1}(\tau \leqslant t-1)$ so it is measurable wrt $\mathcal{H}_{t-1}$. Thus

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}(\tau \geqslant t) \, \mathbf{1}\big(s_h^t = s, a_h^t = a\big) \big| \mathcal{H}_{t-1}\right] &= \mathbf{1}(\tau \geqslant t) \cdot \mathbb{E}\left[\mathbf{1}\big(s_h^t = s, a_h^t = a\big) \big| \mathcal{H}_{t-1}\right] \\ &= \mathbf{1}(\tau \geqslant t) \cdot w_h^{\pi_t^{\exp}}(s,a). \end{aligned}$$

Combining the above displays,

$$\mathbb{E}_M[n_{\tau+1}(s,a,h)] = \mathbb{E}_M \left[\sum_{t=1}^{\tau} w_h^{\pi_t^{\exp}}(s,a) \right],$$

then dividing both sides by $\mathbb{E}_M[\tau]$ yields

$$w_h(s,a) \triangleq \frac{\mathbb{E}_M\left[\sum_{t=1}^{\tau} w_h^{\pi_t^{\exp}}(s,a)\right]}{\mathbb{E}_M[\tau]} = \mathbb{E}_M \left[\frac{1}{\mathbb{E}_M[\tau]} \sum_{t=1}^{\tau} w_h^{\pi_t^{\exp}}(s,a) \right].$$

Next, we remark that using Caratheodory's theorem, A. J. Wagenmaker, Simchowitz, et al. 2022 justifies in the proof of their Lemma F.4 that $\{w_h(s,a)\}_{s,a,h}$ as defined above belongs to Ω_M. Rearranging gives

$$\mathbb{E}_M[\tau] \sum_{s,a,h} w_h(s,a) \, \mathrm{KL}(M \| \widetilde{M})_{s,a,h} \; \geqslant \; \log \frac{1}{2.4\delta}, \forall \; \widetilde{M} \in \mathrm{Alt}(M), \tag{B.3}$$

the result follows by taking on the left-hand side the inf over $\mathrm{Alt}(M)$ followed by a sup over Ω_M. □

C POSTERIOR STOPPING RULE

We discuss the proof of the stopping rule. We prove that under mild conditions, we can prove that a simple posterior stopping algorithm can be used to calibrate an efficient stopping rule.

The stopping rule for this setting relies on the posterior distribution. Letting the empirical MDP $\hat{M}_t$, define the optimal policy in the empirical MDP as $\hat{\pi}_t$, and the value function for policy π as $\hat{V}_{t,0}^{\pi}$.

Letting $B(t,\delta)$ be a function to be defined, and letting $\widetilde{M}_{t,2}, \widetilde{M}_{t,2}, \ldots,$ be *i.i.d* samples from the posteiror distribution over the MDPs, we denote by $\Pi^\star(M)$ the set of optimal policies of M. We introduce the following stopping time :

$$\tau \triangleq \inf\left\{ t \geqslant 1 : \forall k \ \leqslant \ B(t,\delta),\, \hat{\pi}_t \in \Pi^\star(\widetilde{M}_{t,k}) \right\}, \tag{C.1}$$

and after stopping, we recommend the empirical optimal planning $\hat{\pi}^t$. We prove the following lemma. Let us introduce an event $\mathcal{E}_\delta$ that holds with probability at least $(1-\delta)$ for any $\delta \in (0,1)$. Let Alt_π be the set of MDPs for which π is not an optimal policy.

Lemma 7. *We have*

$$\mathrm{Pr}_M(\tau < \infty, \hat{\pi}_\tau \notin \Pi^\star(M)) \leqslant \delta/2 + \sum_{t=1}^{\infty} \mathbb{E}_M\left[\mathbf{1}\big(\mathcal{E}_{\delta/2}, \hat{\pi}_t \notin \Pi^\star(M)\big) \exp\left(-\mathrm{Pr}_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\left(\widetilde{M} \in \mathrm{Alt}_{\hat{\pi}^t}\right) B(t,\delta)\right)\right]$$

Proof. We should expect to show that this stopping rule is δ-PAC for a specific choice of B and possibly an inflation of the posterior. To start, we assume that each MDP admits a unique optimal policy and the posterior is defined over such a space, then observe that

$$\begin{aligned}
\mathrm{Pr}_M(\tau < \infty, \hat{\pi}_\tau \notin \Pi^\star(M)) &\leqslant \mathrm{Pr}_M\left(\tau < \infty, \hat{\pi}_\tau \notin \Pi^\star(M), \left\{ \forall\, k \ \leqslant B(\tau,\delta), \hat{\pi}_\tau \in \Pi^\star(\widetilde{M}^{\tau,k}) \right\}\right) \\
&\leqslant \delta/2 + \sum_{t=1}^{\infty} \mathbb{E}_M\left[\mathbf{1}\big(\mathcal{E}_{\delta/2}, \hat{\pi}_t \notin \Pi^\star(M)\big)\, \mathbb{E}_M\left[\mathbf{1}\Big(\forall\, k \ \leqslant B(t,\delta), \hat{\pi}_t \in \Pi^\star(\widetilde{M}_{t,k})\Big) \Big| \mathcal{H}_{t-1}\right]\right].
\end{aligned}$$

Next, we have to control

$$L_{t,\delta} \triangleq \mathbf{1}\big(\mathcal{E}_{\delta/2}, \hat{\pi}_t \notin \Pi^\star_M\big)\, \mathbb{E}_M\left[\mathbf{1}\Big(\forall\, k \ \leqslant B(t,\delta), \hat{\pi}_t \in \Pi^\star(\widetilde{M}_{t,k})\Big) \mid \mathcal{H}_{t-1}\right],$$

for which a simple rewriting yields

$$\begin{aligned}
L_{t,\delta} &= \mathbf{1}\big(\mathcal{E}_{\delta/2}, \hat{\pi}_t \notin \Pi^\star(M)\big)\, \mathrm{Pr}_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\left(\hat{\pi}_t \in \Pi^\star(\widetilde{M})\right)^{B(t,\delta)} \\
&= \mathbf{1}\big(\mathcal{E}_{\delta/2}, \hat{\pi}_t \notin \Pi^\star(M)\big)\left(1 - \mathrm{Pr}_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\left(\hat{\pi}_t \notin \Pi^\star(\widetilde{M})\right)\right)^{B(t,\delta)} \\
&\leqslant \mathbf{1}\big(\mathcal{E}_{\delta/2}, \hat{\pi}_t \notin \Pi^\star(M)\big) \exp\left\{ -\mathrm{Pr}_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\left(\hat{\pi}_t \notin \Pi^\star(\widetilde{M})\right) B(t,\delta)\right\} \\
&\leqslant \mathbf{1}\big(\mathcal{E}_{\delta/2}, \hat{\pi}_t \notin \Pi^\star(M)\big) \exp\left\{ -\mathrm{Pr}_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\left(\widetilde{M} \in \mathrm{Alt}_{\hat{\pi}_t}\right) B(t,\delta)\right\}.
\end{aligned}$$

□

Thus, calibrating the stopping rule properly requires tuning $B(t,\delta)$ such that the sum in the RHS of Lemma 7 is smaller than $\delta/2$. Preceisly computing $\mathrm{Pr}_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\left(\widetilde{M} \in \mathrm{Alt}_{\hat{\pi}_t}\right)$ involves evaluating high-dimensional integrals over the complex domain $\mathrm{Alt}_{\hat{\pi}_t}$.

D OPTIMAL EXPLORATION VIA ONLINE LEARNING

In this section, we analyze a no-regret RL algorithm for online learning in the setting where rewards change over time and their range is not fixed in advance. We show that such an algorithm generates a sequence of exploration policies that accumulates information optimally. Our main result is a generic regret bound for online RL with changing rewards coupled with a mirror descent update, which is of independent interest beyond the present application.

D.1 Setup and Goal

We analyze a generic version of Algorithm 2 where the policy improvement step uses a generic online learner on the action set $\mathcal{A}$. The KL-based reward kernel fed to the online learner at episode t is

$$r_t^{\mathrm{exp}}(s,a,h) \triangleq \frac{\big(\theta_t(s,a,h) - \hat{\mu}_t(s,a,h)\big)^2}{2} + \sum_{s'} \hat{p}_t(s' \mid s,a,h) \log \frac{\hat{p}_{t-1}(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}, \tag{D.1}$$

where the transition probability clipping ensures $r_t^{\exp}(s,a,h) \in (0, \frac{1}{2} + t^\alpha)$.

To ensure optimal data collection, the exploration algorithm must ultimately guarantee, with high probability, the sequence of exploration policies $(\pi_t^{\exp})_{t \geqslant 1}$ satisfies

$$\sup_{w \in \Omega_M} \sum_{t=1}^{T} \sum_{s,a,h} w_h(s,a)\, r_t^{\exp}(s,a,h) - \sum_{t=1}^{T} \sum_{s,a,h} w_h^{\pi_t^{\exp}}(s,a)\, r_t^{\exp}(s,a,h) \leqslant o(T). \tag{D.2}$$

An online RL learner with changing rewards has been analyzed in Al-Marjani et al. 2023a, but under two restrictive assumptions: the horizon T is known in advance, and the reward range is fixed in $(0,1)$. In our setting, the reward range $(0, \frac{1}{2} + t^\alpha)$ grows with t, requiring a more careful analysis.

D.2 Algorithm

We present a general online RL algorithm, whose pseudo-code is given in Algorithm 3. It maintains SH independent online learners $(\mathcal{L}_{s,h})_{s,h}$ on the simplex $\triangle_A$, one per state-stage pair, combined with an optimistic policy evaluation step to compute the loss fed to each learner. The key features are:

- **Unknown dynamics.** The transition kernel p is unknown and fixed; the agent uses the empirical estimate $\hat{p}_t$ together with Bernstein-type bonuses to ensure optimism.
- **Changing rewards.** The reward kernel $r_t^{\exp}$ is deterministic and revealed to the agent at the end of each episode, but changes with t.
- **Independent per-state policy learning.** The exploration policy $\pi_t^{\exp}(\cdot \mid s,h)$ is updated independently for each (s,h) via the corresponding learner $\mathcal{L}_{s,h}$, initialized at the uniform distribution over $\mathcal{A}$.

Optimistic policy evaluation. At episode t, the agent collects the trajectory $\{(s_h^t, a_h^t, s_{h+1}^t)\}_{h=1}^H$ using $\pi_t^{\exp}$, after which the reward kernel $r_t^{\exp}$ is revealed. The optimistic Q-function for $\pi_t^{\exp}$ is computed by backward induction using the Bernstein-type bonus

$$\mathrm{bonus}_t(s,a,h) \triangleq \sqrt{\frac{2\ \mathrm{Var}_{\hat{p}_{t,h}}\left[\bar{V}_{t,h+1}^{\pi_t^{\exp}}\right](s,a) \cdot \beta^p(n_t(s,a,h), 1/t^3)}{\max(1, n_t(s,a,h))}} + \frac{2\beta^p(n_t(s,a,h), 1/t^3)\, g_h^t}{3\max(1, n_t(s,a,h))}, \tag{D.3}$$

where $g_{t,h} \triangleq \max_s \bar{V}_{t,h+1}(s)$. The optimistic Q-function and its associated value function are defined recursively as

$$\bar{Q}_{t,h}(s,a) \triangleq r_t^{\exp}(s,a,h) + \hat{p}_{t,h}\bar{V}_{t,h+1}^{\pi_t^{\exp}}(s,a) + \mathrm{bonus}_t(s,a,h), \qquad \bar{V}_{t,h}(s) \triangleq \sum_{a \in \mathcal{A}} \pi_t^{\exp}(a \mid s,h)\, \bar{Q}_{t,h}(s,a), \tag{D.4}$$

with $\bar{V}_{t,H+1} \equiv 0$. In the policy improvement phase, each learner $\mathcal{L}_{s,h}$ is updated with $\bar{Q}_{t,h}(s,\cdot)$ and the next policy $\pi_{t+1}^{\exp}$ is computed from the updated learner.

Choice of online learner. We use Bernstein-type bonuses (D.3) together with AdaHedge (De Rooij et al. 2014) as the per-state online learner. AdaHedge is parameter-free, adapts its learning rate on the fly, and guarantees sublinear regret even for unbounded losses, which is essential here since the range of ι_t grows with t. Hedge with a scaled learning rate is also valid but yields less tight finite-time guarantees.

Algorithm 3: POLICY UPDATE HELPER FUNCTION

Function *ComputeExplorationPolicy*$\big(t, \pi_t^{\exp}, r_t^{\exp}, \hat{p}_t, n_t\big)$

Set $\bar{V}_{t,H+1}(s) \equiv 0$ for all $s \in \mathcal{S}$

for *episode stage* $h = H, H-1, \dots, 1$ **do**

foreach $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

$$\bar{Q}_{t,h}(s,a) \leftarrow r_t^{\exp}(s,a,h) + \hat{p}_{t,h}\bar{V}_{t,h+1}(s,a) + \sqrt{\frac{2\,\mathrm{Var}_{\hat{p}_{t,h}}\big[\bar{V}_{t,h+1}\big](s,a)\ \cdot \beta^p(n_t(s,a,h),\,1/t^3)}{\max\left(1, n_t(s,a,h)\right)}} + \frac{2\,\beta^p(n_t(s,a,h),\,1/t^3)}{3\,\max\left(1, n_t(s,a,h)\right)}$$

// Optimistic policy evaluation

$\bar{V}_{t,h}(s) \leftarrow \sum_{a\in\mathcal{A}} \pi_t^{\exp}(a \mid s,h)\,\bar{Q}_{t,h}(s,a) \quad \forall s \in \mathcal{S}$

foreach $(s,h) \in \mathcal{S} \times [H]$ **do** // Policy improvement

Feed learner $\mathcal{L}_{s,h}$ with gain $\bar{Q}_{t,h}(s,\cdot)$

Update $\pi_{t+1}^{\exp}(\cdot \mid s,h)$ from $\mathcal{L}_{s,h}$

return policy $\pi_{t+1}^{\exp}$

D.3 Regret Upper Bound

We present the analysis below. We define the events :

$$\mathcal{E}_D^p(T) \triangleq \left\{\forall\, t \leqslant T\,, \forall\,(s,a,h)\,, n_t(s,a,h)\,\mathrm{KL}(\hat{p}_t(\cdot|s,a,h) \parallel p(\cdot|s,a,h)) \leqslant \beta^p(n_t(s,a,h), 1/T^2)\right\}, \tag{D.5}$$

$$\mathcal{E}_D^{\mathrm{cnt}}(T) \triangleq \left\{\forall\, t \leqslant T\,, \forall\,(s,a,h)\,, n_t(s,a,h) \geqslant \frac{1}{2}\bar{n}_t(s,a,h) - \log(2SAHT^2)\right\}, \tag{D.6}$$

$$\mathcal{E}_D(T) \triangleq \mathcal{E}_D^p(T) \cap \mathcal{E}_D^{\mathrm{cnt}}(T)\,. \tag{D.7}$$

At episode t, the challenger MDP $\widetilde{M}_t \triangleq (\theta_t, \Phi_t)$ is sampled from the inflated posterior, and $\hat{M}_t = (\hat{\mu}_t, \hat{p}_t)$ denotes the empirical MDP. The reward kernel $r_t^{\exp}$ is defined in (D.1).

We prove the following result.

Proposition 8. *Let $(\mathcal{L}_{s,h})_{s,h}$ be some online learners on $\triangle$ such that for any sequence of losses, the learner guarantees some regret $\mathrm{Regret}_{\mathcal{L}}(T)$. Then Algorithm 3 guarantees the following regret bound on the event $\mathcal{E}_D$,*

$$\mathrm{Regret}(T) \leqslant O(T^{\alpha+\frac{1}{2}}\beta^p(T, 1/T^3)) + H\,\mathrm{Regret}_{\mathcal{L}}(T)\,.$$

Proof. Below, we let $Q_{t,h}^{\pi}$ and $V_{t,h}^{\pi}$ denote the state–action and state value functions in the MDP $M_t^{\exp} \triangleq (r_t^{\exp}, p)$. $\bar{Q}_{t,h}^{\pi}, \bar{V}_{t,h}^{\pi}$ denote the state–action and state value functions in the "optimistic MDP" $\bar{M}_t^{\exp} \triangleq (r_t^{\exp} + \mathrm{bonus}_t, p)$. For any fixed comparator ρ,

$$\sum_{t=1}^{T} V_{t,0}^{\rho} - V_{t,0}^{\pi_t^{\exp}} = \sum_{t=1}^{T}\left[V_{t,0}^{\rho} - \bar{V}_{t,0}^{\rho} + \bar{V}_{t,0}^{\rho} - V_{t,0}^{\pi_t^{\exp}} + \bar{V}_{t,0}^{\pi_t^{\exp}} - \bar{V}_{t,0}^{\pi_t^{\exp}}\right]. \tag{D.8}$$

On the event $\mathcal{E}_D(T)$, for $t \in (\sqrt{T}, T)$, the optimism property $\bar{V}_{t,0}^{\rho} \geqslant V_{t,0}^{\rho}$ holds (Lemma 9), justifying that on $\mathcal{E}_D(T)$

$$\sum_{t=1}^{T}\left[V_{t,0}^{\rho} - \bar{V}_{t,0}^{\rho}\right] \leqslant O(T^{\alpha+\frac{1}{2}}).$$

It remains to bound the remaining terms appearing in Eq. (D.8). Let us define

$$(A) = \sum_{t=1}^{T}\left[\bar{V}_{t,0}^{\rho} - \bar{V}_{t,0}^{\pi_t^{\exp}}\right] \qquad (B) = \sum_{t=1}^{T}\left[\bar{V}_{t,0}^{\pi_t^{\exp}} - V_{t,0}^{\pi_t^{\exp}}\right]$$

Bounding (A). By the performance difference lemma,

$$\bar{V}_{t,0}^{\rho} - \bar{V}_{0}^{t,\pi_t^{\exp}} = \mathbb{E}^{\rho}\left[\sum_{h=1}^{H}\left\langle \bar{Q}_{t,h}^{\pi_t^{\exp}}(s_h,\cdot),\, \rho(\cdot \mid s_h,h) - \pi_t^{\exp}(\cdot \mid s_h,h)\right\rangle \,\middle|\, s_1 = s_{\text{init}}\right]. \tag{D.9}$$

The inner product in (D.9) is exactly the loss fed to learner $\mathcal{L}_{s_h,h}$ at round t against comparator $\rho(\cdot \mid s_h,h) = \rho_h(\cdot \mid s_h)$. By the regret guarantee of each learner,

$$\sum_{t=1}^{T}\left\langle \bar{Q}_{t,h}^{\pi_t^{\exp}}(s,\cdot),\, \rho(\cdot \mid s,h) - \pi_t^{\exp}(\cdot \mid s,h)\right\rangle \leqslant \mathrm{Regret}_{\mathcal{L}}(T), \quad \forall (s,h) \in \mathcal{S}\times[H].$$

Combining with (D.9) and summing over h,

$$(A) \leqslant H\,\mathrm{Regret}_{\mathcal{L}}(T). \tag{D.10}$$

Bounding (B). Term (B) controls the total size of the Bernstein bonuses. On $\mathcal{E}_D(T)$, by induction,

$$\bar{V}_{t,0}^{\pi_t^{\exp}} - V_{t,0}^{\pi_t^{\exp}} \leqslant 2\sum_{s,a,h} w_h^{\pi_t^{\exp}}(s,a)\,\mathrm{bonus}_t(s,a,h),$$

where, using $\mathrm{Var}_{\hat{p}_{t,h}}(\bar{V}_{t,h+1})[s,a] \leqslant (g_h^t)^2$,

$$\mathrm{bonus}_t(s,a,h) \leqslant g_{t,h}\left[\sqrt{\frac{2\beta^p(n_t(s,a,h),1/t^3)}{\max\left(1,n_t(s,a,h)\right)}} + \frac{2\beta^p(n_t(s,a,h),1/t^3)}{3\max\left(1,n_t(s,a,h)\right)}\right]. \tag{D.11}$$

We split the sum over (s,a,h) into two parts using the set

$$\mathcal{Z}_t \triangleq \left\{(s,a,h) :\, \bar{n}_t(s,a,h) \geqslant 4\log(2SAHT^2)\right\}, \qquad \bar{n}_t(s,a,h) \triangleq \sum_{k=1}^{t-1}\Big[(1-\varepsilon_k)w_h^{\pi_k^{\exp}}(s,a) + \varepsilon_k w_h^{c_k}(s,a)\Big].$$

On $\mathcal{E}_D(T)$, for $(s,a,h)\in\mathcal{Z}_t$: $n_t(s,a,h) \geqslant \frac{1}{4}\bar{n}_t(s,a,h)$. Moreover, since $\varepsilon_t = t^{-\gamma}$, for $t \geqslant 2^{1/\gamma}$, $1-\varepsilon_{t+1} \geqslant \frac{1}{2}$.

Using $\max_{t\leqslant T,h}[g_{t,h}] \leqslant T^{\alpha}$ and monotonicity of β^p in its second argument,

$$\sum_{t=1}^{T}\sum_{(s,a,h)\in\mathcal{Z}_t} w_h^{\pi_t^{\exp}}(s,a)\frac{g_{t,h}\,\beta^p(n_t(s,a,h),1/t^3)}{\max\left(1,n_t(s,a,h)\right)} \leqslant T^{\alpha}\beta^p(T,1/T^3)\sum_{t=1}^{T}\sum_{(s,a,h)\in\mathcal{Z}_t}\frac{w_h^{\pi_t^{\exp}}(s,a)}{\max\left(1,\frac{1}{4}\bar{n}_t(s,a,h)\right)}.$$

Applying the elliptic potential (Lemma 46) and Cauchy–Schwarz,

$$\sum_{t=1}^{T}\sum_{(s,a,h)\in\mathcal{Z}_t}\frac{\frac{1}{4}(1-\varepsilon_t)w_h^{\pi_t^{\exp}}}{\max\left(1,\frac{1}{4}\bar{n}_t(s,a,h)\right)} \leqslant 4\sum_{s,a,h}\log\left(1+\tfrac{1}{4}\sum_{t=1}^{T+1}(1-\varepsilon_t)w_h^{\pi_t^{\exp}}(s,a)\right)$$

and similarly, for any $\gamma\in(0,1)$

$$\begin{aligned}\sum_{t=1}^{T}\sum_{(s,a,h)\in\mathcal{Z}_t}\frac{w_h^{\pi_t^{\exp}}(s,a)}{\sqrt{\max\left(1,n_t(s,a,h)\right)}} &\leqslant O\left(\sum_{t=1}^{T}\sum_{(s,a,h)\in\mathcal{Z}_t}\left[\frac{(1-\varepsilon_t)w_h^{\pi_t^{\exp}}(s,a)}{\sqrt{\max\left(1,\sum_{k=1}^{t-1}(1-\varepsilon_k)w_h^{\pi_k^{\exp}}(s,a)\right)}}\right]\right) \\ &\leqslant O\Big(\sqrt{T+1}\Big).\end{aligned} \tag{D.12}$$

Combining these bounds,

$$\text{Term } (B)\big|_{\mathcal{Z}_t} \leqslant O\Big(\beta^p(T, 1/T^3)\, T^{\alpha+\frac{1}{2}}\Big) \tag{D.13}$$

For terms outside $\mathcal{Z}_t$, we note that for $(s,a,h) \notin \mathcal{Z}_t$, $\bar{n}_t(s,a,h) < 4\log(2SAHT^2)$, so the total visits to such pairs is bounded. We have,

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{(s,a,h)\notin\mathcal{Z}_t} w_h^{\pi_t^{\exp}}(s,a)\,\text{bonus}_t(s,a,h) \leqslant \\
&\sum_{s,a,h} \max_{t\leqslant T} \text{bonus}_t(s,a,h) \left[\sum_{t=1}^{T} w_h^{\pi_t^{\exp}}(s,a)\mathbf{1}\left(\sum_{k=1}^{t-1}(1-\varepsilon_k) w_h^{\pi_k^{\exp}}(s,a) < 4\log(SAHT^2)\right)\right] \leqslant \\
&\max_{t\leqslant T,s,a,h} \text{bonus}_t(s,a,h)\, O\left(\sum_{t=1}^{T}(1-\varepsilon_t) w_h^{\pi_t^{\exp}}(s,a)\mathbf{1}\left(\sum_{k=1}^{t-1}(1-\varepsilon_k) w_h^{\pi_k^{\exp}}(s,a) < 4\log(SAHT^2)\right)\right) \leqslant \\
&\max_{t\leqslant T,s,a,h} \text{bonus}_t(s,a,h) \cdot O(\log T)
\end{aligned}$$

where the latter inequality holds for $\gamma \in (0,1)$, and we recall that $\varepsilon_t \triangleq t^{-\gamma}$.

Using

$$\max_{t\leqslant T,s,a,h} \text{bonus}_t(s,a,h) \leqslant T^{\alpha}\left[\sqrt{2\beta^p(T,1/T^3)} + \frac{2\beta^p(T,1/T^3)}{3}\right],$$

we have

$$\sum_{t=1}^{T} \sum_{(s,a,h)\notin\mathcal{Z}_t} w_h^{\pi_t^{\exp}}(s,a)\,\text{bonus}_t(s,a,h) \leqslant O(T^{\alpha}\beta^p(T,1/T^3)\log T). \tag{D.14}$$

Combining (D.10), (D.13), and (D.14), and noting that β^p is logarithmic in T so all bonus terms are $O(\sqrt{T})$,

$$\text{Regret}(T) \leqslant H\,\text{Regret}_{\mathcal{L}}(T) + O\Big(\beta^p(T,1/T^3)\,T^{\alpha+\frac{1}{2}}\Big),$$

on the event $\mathcal{E}_D(T)$. □

Lemma 9. *On the event $\mathcal{E}_D(T)$, for any $t \in (\sqrt{T}, T)$ and s, a, h, and any policy ρ, we have $\bar{Q}^{\rho}_{t,h}(s,a) \geqslant Q^{\rho}_{t,h}(s,a)$.*

Proof. We justify that for $t \in (\sqrt{T}, T)$, the bonus used in the optimistic Q-function is sufficient to ensure optimism on the event $\mathcal{E}_D(T)$.

We proceed by induction; assuming $\bar{Q}^{\rho}_{t,h+1}(s,a) \geqslant Q^{\rho}_{t,h+1}(s,a)$ holds for all s, a and some h fixed. Note that this directly implies that $\bar{V}^{\rho}_{t,h+1}(s) \geqslant V^{\rho}_{t,h+1}(s)$ and we have

$$\begin{aligned}
\bar{Q}^{\rho}_{t,h}(s,a) &= r_t^{\exp}(s,a,h) + \hat{p}_{t,h}\bar{V}^{\rho}_{t,h+1}(s,a) + \sqrt{\frac{2\,\text{Var}_{\hat{p}_{t,h}}\big[\overline{V}^{\rho}_{t,h+1}\big](s,a)\cdot\beta^p(n_t(s,a,h),1/t^3)}{\max\left(1,n_t(s,a,h)\right)}} + \\
&\qquad \frac{2\beta^p(n_t(s,a,h),1/t^3)g_{t,h}}{3\max\left(1,n_t(s,a,h)\right)} \\
&\overset{(a)}{\geqslant} r_t^{\exp}(s,a,h) + \hat{p}_{t,h}\bar{V}_{t,h+1}(s,a) + \sqrt{\frac{2\,\text{Var}_{\hat{p}_{t,h}}\big[\overline{V}_{t,h+1}\big](s,a)\cdot\beta^p(n_t(s,a,h),1/T^2)}{\max\left(1,n_t(s,a,h)\right)}} + \\
&\qquad \frac{2\beta^p(n_t(s,a,h),1/T^2)g_{t,h}}{3\max\left(1,n_t(s,a,h)\right)} \\
&\overset{(b)}{\geqslant} r_t^{\exp}(s,a,h) + p_h V^{\star}_{h+1}(s,a) \quad \text{(induction + monotonicity, see e.g. Menard et al. 2021)}
\end{aligned}$$

where the last inequality holds on the event $\mathcal{E}_D(T)$ and we recall that for $t \in (\sqrt{T}, T)$, we have $t^3 \geqslant T^2$, combining with the fact that $\delta \mapsto \beta(x, \delta)$ is decreasing. This concludes the induction step, and as the induction hypothesis trivially holds for $H+1$, we have proved the claimed result.

□

We now state the main result of this section, which is the regret of the exploration algorithm.

Proposition 10. *Running Algorithm 2 on an MDP M with the exploration function* ComputeExplorationPolicy *using AdaHedge as a subroutine generates a sequence of policies $(\pi_t^{\exp})_{t\geqslant 1}$ satisfying on the event $\mathcal{E}_D(T)$*

$$\sup_{w\in\Omega_M} \sum_{t\leqslant T}\sum_{s,a,h} w_h(s,a) r_t^{\exp}(s,a,h) - \sum_{t\leqslant T}\sum_{s,a,h} w_h^{\pi_t^{\exp}}(s,a) r_t^{\exp}(s,a,h) \leqslant O(\beta^p(T, 1/T^3) T^{\alpha+\frac{1}{2}}) \,.$$

Proof. This is a direct consequence of Proposition 8. Indeed, using AdaHedge as subroutine ensures that after T iterations, we have (cf De Rooij et al. 2014,)

$$\mathrm{Regret}_{\mathcal{L}}(T) \leqslant \sigma_T \sqrt{T \log S} + 16\sigma_T(2 + \log(S)/3),$$

where $\sigma_T \triangleq \max_{t\leqslant T, s,h,a} \bar{Q}_{t,h}^{\pi_t^{\exp}}(s,a)$. We note that ,

$$\sigma_T \leqslant O(T^{\alpha}\beta^p(T, 1/T^3)),$$

and the conclusion follows thanks to Proposition 8. □

E SUFFICIENT EXPLORATION AND MODEL ESTIMATION

In this section, we show that every reachable state–action pair is visited sufficiently often to guarantee accurate estimation of the unknown model parameters. Our approach achieves this by leveraging a regret minimization algorithm operating under *changing reward functions*.

The use of regret minimization for exploration in MDPs has been widely studied. However, existing analyses typically rely on additional structural assumptions, such as the MDP being communicating or ergodic (A. Wagenmaker & Jamieson 2022; Al-Marjani et al. 2023a), the availability of lower bounds on visitation probabilities (A. J. Wagenmaker, Simchowitz, et al. 2022), or the use of periodic restarts and re-initialization of the learning algorithm (A. J. Wagenmaker, Simchowitz, et al. 2022). In contrast, our analysis does not require any of these assumptions.

We define the following high-probability events:

$$\mathcal{E}_E^p(T) \triangleq \Big\{\forall t \leqslant T,\ \forall (s,a,h):\ n_t(s,a,h)\ \mathrm{KL}\big(\hat{p}_t(\cdot \mid s,a,h) \,\|\, p(\cdot \mid s,a,h)\big) \leqslant \beta^p(n_t(s,a,h), 1/T^2)\Big\}, \tag{E.1}$$

$$\mathcal{E}_E^{\mu}(T) \triangleq \Big\{\forall t \leqslant T,\ \forall (s,a,h):\ n_t(s,a,h)\ \mathrm{KL}\big(\hat{R}_h^t(s,a) \,\|\, R_h(s,a)\big) \leqslant \beta^{\mu}(n_t(s,a,h), 1/T^2)\Big\}, \tag{E.2}$$

$$\mathcal{E}_E^{\mathrm{cnt}}(T) \triangleq \Big\{\forall t \leqslant T,\ \forall (s,a,h):\ n_t(s,a,h) \geqslant \tfrac{1}{2}\,\bar{n}_t(s,a,h) - \log(2SAHT^2)\Big\}. \tag{E.3}$$

Here,

$$\bar{n}_t(s,a,h) \triangleq \sum_{k=1}^{t-1}\Big[(1-\varepsilon_k)\, w_h^{\pi_k^{\exp}}(s,a) + \varepsilon_k\, w_h^{c_k}(s,a)\Big]. \tag{E.4}$$

We recall that $Z_t \sim \mathrm{Bernoulli}(t^{-\gamma})$ and

$$\tilde{\pi}_t^{\exp}(\cdot|s,h) \leftarrow \begin{cases} c_t(\cdot|s,h), & \text{if } Z_t = 1, \\ \pi_t^{\exp}(\cdot|s,h), & \text{if } Z_t = 0, \end{cases} \quad \forall (s,h) \in \mathcal{S}\times[H].$$

We introduce the event

$$\mathcal{E}_E^Z(T) \triangleq \left\{ \forall (s,a,h),\, \forall t \leqslant T,\, \sum_{k=1}^{t} \varepsilon_k \cdot \mathbf{1}\Big(\tilde{s}^k = s, \tilde{a}^k = a, \tilde{h}^k = h\Big) \geqslant \frac{1}{2SAH} \sum_{k=1}^{t} \varepsilon_k - \log(2SAHT^2) \right\}. \tag{E.5}$$

We further define

$$\mathcal{E}_E(T) \triangleq \mathcal{E}_E^p(T) \cap \mathcal{E}_E^\mu(T) \cap \mathcal{E}_E^{\mathrm{cnt}}(T) \cap \mathcal{E}_E^Z(T) \tag{E.6}$$

E.1 Online Regret of UCBVI with Changing Rewards

We recall that a state–action pair (s,a) is said to be reachable at stage h if $\max_\rho w_h^\rho(s,a) > 0$, i.e., if there exists a policy that visits (s,a) at stage h with positive probability.

We now show that, on event $\mathcal{E}_E(T)$, every reachable state–action pair is visited sufficiently often to ensure the consistency of the empirical estimates used by the algorithm.

To enforce exploration, we introduce a forced exploration mechanism based on a single step of UCBVI with a known reward function defined as the indicator of a randomly selected state–action–stage triplet. Formally, at episode t, a triplet $(\tilde{s}^t, \tilde{a}^t, \tilde{h}^t)$ is sampled uniformly at random, and we define the reward kernel r_t^{f} as

$$r_t^{\mathrm{f}}(s,a,h) \;\triangleq\; \mathbf{1}\Big(s = \tilde{s}^t,\, a = \tilde{a}^t,\, h = \tilde{h}^t\Big)\;. \tag{E.7}$$

Let $M_t^{\mathrm{f}} \triangleq (\mathcal{S}, \mathcal{A}, H, p, r_t^{\mathrm{f}}, s_{\mathrm{init}})$ denote the corresponding MDP. For any policy π, the value at the initial state satisfies

$$V_{t,0}^\pi = w_{\tilde{h}^t}^\pi(\tilde{s}^t, \tilde{a}^t),$$

that is, the visitation probability of $(\tilde{s}^t, \tilde{a}^t)$ at stage $\tilde{h}^t$ under policy π.

Consequently, an optimal policy for M_t^{f} maximizes the visitation probability of the sampled triplet $(\tilde{s}^t, \tilde{a}^t, \tilde{h}^t)$.

Finally, we note that maximizing $w_h^\pi(s,a)$ is equivalent to maximizing $w_h^\pi(s)$, since the action chosen at stage h does not affect the probability of reaching state s at that stage. Therefore,

$$W_h(s,a) \triangleq \max_\pi w_h^\pi(s,a) = \max_\pi w_h^\pi(s).$$

E.1.1 Algorithm

The forced exploration policy is computed via a single optimistic planning step using the empirical transition kernel $\hat{p}_t$. The goal is to construct an optimistic estimate of the optimal Q-function in the MDP $M_t^{\mathrm{f}} \triangleq (r_t^{\mathrm{f}}, p)$.

In this section, let $\bar{Q}_t$ and $\bar{V}_t$ denote the resulting optimistic state–action and value functions. For notational simplicity, we omit the explicit dependence on the reward kernel r_t^{f}.

We define the exploration bonus

$$\mathrm{bonus}_t(s,a,h) \;\triangleq\; \sqrt{\frac{2\,\mathrm{Var}_{\hat{p}_{t,h}}\big[\bar{V}_{t,h+1}\big](s,a)\;\beta^p\big(n_t(s,a,h),\,1/t^3\big)}{\max\big(1, n_t(s,a,h)\big)}} \;+\; \frac{2\,\beta^p\big(n_t(s,a,h),\,1/t^3\big)}{3\,\max\big(1, n_t(s,a,h)\big)}\,. \tag{E.8}$$

We then construct the optimistic Q-function as

$$\bar{Q}_{t,h}(s,a) \;\triangleq\; r_t^{\mathrm{f}}(s,a,h) \;+\; \hat{p}_{t,h}\bar{V}_{t,h+1}(s,a) \;+\; \big(\mathrm{bonus}_t(s,a,h) \wedge 1\big). \tag{E.9}$$

The clipping by 1 ensures consistency with the reward range.

The exploration policy c_t is defined as the greedy policy with respect to $\bar{Q}_t$, namely

$$c_t(s,h) \;\triangleq\; \operatorname*{argmax}_{a \in \mathcal{A}} \bar{Q}_{t,h}(s,a), \quad \forall (s,h) \in \mathcal{S} \times [H].$$

In the sequel, we denote by $V^{\rho}_{t,\cdot}$ the value function in M^{f}_t, and, by $\bar{V}^{\rho}_{t,\cdot}$ the value function in the "optimistic MDP" $\bar{M}^{\mathrm{f}}_t \triangleq (r^{\mathrm{f}}_t + \mathrm{bonus}_t, \hat{p}_t)$. The associated Q-value functions are defined similarly.

Algorithm 4: FORCED EXPLORATION HELPER FUNCTION

Function *ComputeForcedExplorationPolicy*$(t, \hat{p}_t, n_t)$

Sample $(\tilde{s}^t, \tilde{a}^t, \tilde{h}^t)$ uniformly at random from $\mathcal{S} \times \mathcal{A} \times [H]$

Set reward function $r^{\mathrm{f}}_t(s,a,h) \leftarrow \mathbf{1}\Big(s = \tilde{s}^t,\, a = \tilde{a}^t,\, h = \tilde{h}^t\Big)$

Initialize $\bar{V}_{t,H+1}(s) \equiv 0,\ \forall s \in \mathcal{S}$

for $h = H, H-1, \dots, 1$ **do**

  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

    $\bar{Q}_{t,h}(s,a) \leftarrow r^{\mathrm{f}}_t(s,a,h) + \hat{p}_{t,h}\bar{V}_{t,h+1}(s,a) + \mathrm{bonus}_t(s,a,h)$

  $\bar{V}_{t,h}(s) \leftarrow \max_{a\in\mathcal{A}} \bar{Q}_{t,h}(s,a), \quad \forall s \in \mathcal{S}$

// Greedy policy with respect to optimistic Q

for $(s,h) \in \mathcal{S} \times [H]$ **do**

  $c_t(s,h) \leftarrow \mathrm{argmax}_{a\in\mathcal{A}} \bar{Q}_{t,h}(s,a)$

return policy c_t

E.1.2 Regret Upper Bound

We define the well-explored set for a fixed T

$$\mathcal{Z}_t \triangleq \Big\{(s,a,h) :\ \bar{n}_t(s,a,h) \geqslant 4\log(2SAHT^2)\Big\}. \tag{E.10}$$

On the event $\mathcal{E}^{\mathrm{cnt}}_E(T)$, we have for all $(s,a,h) \in \mathcal{Z}_t$:

$$n_t(s,a,h) \geqslant \tfrac{1}{4}\,\bar{n}_t(s,a,h). \tag{E.11}$$

We also introduce for any sequence $(\kappa_t)_{t\geqslant 1}$, the sum operator $\kappa_{1:T} \triangleq \sum_{t=1}^{T}\kappa_t$.

We prove the following result.

Lemma 11. *Let $(S_t)_{t\geqslant 1}$ be a sequence of binary random variables. On the event $\mathcal{E}_E(T)$, the policies $(c_t)_{t\geqslant 1}$ produced by Algorithm 4 satisfy for any $K \in (\sqrt{T}, T)$,*

$$\sum_{t=1}^{K} S_t\varepsilon_t\Big(V^{\star}_{t,0} - V^{c_t}_{t,0}\Big) \ \leqslant\ \sqrt{T^{1-\gamma}} + O(\log T) + O\left(\sqrt{\beta^p(T,1/T^2)}\sqrt{1 + \sum_{t=1}^{K}\kappa_t\, w^{c_t}_h(s,a)}\right),$$

where $\kappa_t = S_t\varepsilon_t$.

Proof. We decompose

$$\sum_{t=1}^{K}\kappa_t\big(V^{\star}_{t,0} - V^{c_t}_{t,0}\big) \ \leqslant\ \underbrace{\sum_{t=1}^{\sqrt{T}}\kappa_t}_{(I)} + \underbrace{\sum_{t=\sqrt{T}+1}^{K}\kappa_t\big(\bar{V}_{t,0} - V^{c_t}_{t,0}\big)}_{(II)}. \tag{E.12}$$

We focus on controlling (II).

For $t \geqslant \sqrt{T}$, we have $1/t^3 \leqslant 1/T^2$. Since $\delta \mapsto \beta^p(\cdot,\delta)$ is non-increasing, the exploration bonus used in Algorithm 4 dominates the one associated with confidence level $1/T^2$. Therefore, when $\mathcal{E}^p_E(T)$ holds, from classical UCBVI analysis technique (see e.g. Azar, Osband, et al. 2017)

$$\bar{Q}_{t,h}(s,a) \ \geqslant\ Q^{\star}_{t,h}(s,a), \quad \forall (s,a,h),\ \forall t \in (\sqrt{T}, T).$$

By backward induction, this implies

$$\bar{V}_{t,h}(s) \;\geqslant\; V^{\star}_{t,h}(s), \quad \forall (s,h).$$

Since c_t is greedy with respect to $\bar{Q}_t$, standard decomposition with value difference lemma yields

$$\bar{V}_{t,0} - V^{c_t}_{t,0} \;\leqslant\; \sum_{s,a,h} 2\, w^{c_t}_h(s,a) \left(\mathrm{bonus}_t(s,a,h) \wedge 1\right).$$

We then obtain

$$(II) \;\leqslant\; \sum_{t=1}^{K} \sum_{s,a,h} 2\,\kappa_t\, w^{c_t}_h(s,a) \left(\mathrm{bonus}_t(s,a,h) \wedge 1\right).$$

For $(s,a,h) \notin \mathcal{Z}_t$, we have $\bar{n}_t(s,a,h) < 4\log(SAHT^2)$, hence

$$\begin{aligned}\sum_{t=1}^{K} \sum_{(s,a,h)\notin\mathcal{Z}_t} 2\kappa_t w^{c_t}_h(s,a)\big(\mathrm{bonus}_t(s,a,h)\wedge 1\big) &\leqslant \sum_{t=1}^{K}\sum_{s,a,h} 2\varepsilon_t w^{c_t}_h(s,a)\mathbf{1}\left(\sum_{k=1}^{t-1}\varepsilon_k w^{c_k}_h(s,a) < 4\log(SAHT^2)\right)\\ &\leqslant 2 + 8\log(SAHT^2). \end{aligned} \tag{E.13}$$

For the well-explored triplets, let us define

$$(III) \;\triangleq\; \sum_{t=1}^{K} \sum_{(s,a,h)\in\mathcal{Z}_t} 2\kappa_t w^{c_t}_h(s,a) \left(\mathrm{bonus}_t(s,a,h) \wedge 1\right).$$

Using Eq. (E.11)

$$\sum_{t=1}^{K} \sum_{(s,a,h)\in\mathcal{Z}_t} \kappa_t w^{c_t}_h(s,a) \frac{\beta^p(n_t(s,a,h),1/T^2)}{\max\left(1, n_t(s,a,h)\right)} \leqslant O\left(\beta^p(T,1/T^2)\log\left(1+\tfrac{1}{4}\sum_{t=1}^{K}\kappa_t w^{c_t}_h(s,a)\right)\right) \tag{E.14}$$

which follows by elliptic potential (Lemma 46), the definition of $\mathcal{Z}_t$, and monotonicity of β^p.

Similarly, since $\bar{V}_{t,h+1}(\cdot) \in (0,1)$, we have $\mathrm{Var}_{\hat{p}_{t,h}}\left[\bar{V}_{t,h+1}\right](s,a) \leqslant 1$. Hence,

$$\begin{aligned}&\sum_{t=1}^{K} \sum_{(s,a,h)\in\mathcal{Z}_t} \kappa_t\, w^{c_t}_h(s,a) \sqrt{\frac{2\,\mathrm{Var}_{\hat{p}_{t,h}}[\bar{V}_{t,h+1}](s,a)\,\beta^p(n_t(s,a,h),1/T^2)}{\max\left(1, n_t(s,a,h)\right)}} \leqslant\\ &\sum_{t=1}^{K} \sum_{(s,a,h)\in\mathcal{Z}_t} \kappa_t\, w^{c_t}_h(s,a) \sqrt{\frac{2\,\beta^p(T,1/T^2)}{\max\left(1, \frac{1}{4}\,\bar{n}_t(s,a,h)\right)}} \leqslant\\ &4\sqrt{2\,\beta^p(T,1/T^2)} \sum_{t=1}^{K}\sum_{s,a,h} \frac{\kappa_t\, w^{c_t}_h(s,a)}{\sqrt{\max\left(1, \frac{1}{4}\sum_{k=1}^{t-1}\kappa_k w^{c_k}_h(s,a)\right)}},\end{aligned}$$

where the last inequality uses

$$\bar{n}_t(s,a,h) \geqslant \sum_{k=1}^{t-1} \kappa_k w^{c_t}_h(s,a),$$

which follows from Eq. (E.4), where $\kappa_k = \varepsilon_k S_k$ and S_k is binary.

Applying Lemma 16 of Kaufmann, Ménard, et al. 2021b, we obtain

$$\sum_{t=1}^{K} \sum_{(s,a,h)\in\mathcal{Z}_t} \kappa_t\, w^{c_t}_h(s,a) \sqrt{\frac{2\,\mathrm{Var}_{\hat{p}_{t,h}}[\bar{V}_{t,h+1}](s,a)\,\beta^p(n_t(s,a,h),1/T^2)}{\max\left(1, n_t(s,a,h)\right)}} \leqslant O\left(\sqrt{\beta^p(T,1/T^2)}\sqrt{1+\sum_{t=1}^{K}\kappa_t\, w^{c_t}_h(s,a)}\right).$$

Combining this bound with Eq. (E.14) yields, on the event $\mathcal{E}_E(T)$,

$$(III) \;\leqslant\; O\left(\sqrt{\beta^p(T,1/T^2)}\sqrt{1+\sum_{t=1}^{K}\kappa_t\, w_h^{c_t}(s,a)}\right) \tag{E.15}$$

Finally, combining Eq. (E.12), Eq. (E.13), and Eq. (E.15) concludes the proof. □

E.2 Sufficient Exploration of Reachable States

In this section, we prove sufficient exploration of reachable state–actions. We recall that

$$W_h(s,a) \triangleq \sup_{\rho} w_h^{\rho}(s,a),$$

is the maximum visitation probability of triplet (s,a,h).

We prove the following lemma.

Lemma 12. *Let $\gamma \in (0,1)$. There exists an explicit finite constant $k_0 < \infty$, such that on the event $\mathcal{E}_E(T)$, for all $T \geqslant k_0$, all $t \in (T^{2/3}, T)$, and all (s,a,h),*

$$n_t(s,a,h) \geqslant \frac{W_h(s,a)}{4SAH} t^{-\gamma}.$$

The explicit expression of k_0 is given in the proof.

Proof. Let us fix a triplet (s,a,h) and $t \in (T^{2/3}, T)$. We note that the claim is trivially satisfied for unreachable triplets. Hence, in the sequel, we assume $W_h(s,a,h) > 0$.

We have

$$\begin{aligned}
\mathbb{E}\big[\mathbf{1}\big(s_h^t = s, a_h^t = a\big)\,\big|\mathcal{H}_{t-1}\big] &= (1-\varepsilon_t)\,\mathbb{E}\big[\mathbf{1}(s_h = s, a_h = a)\,\big|\mathcal{H}_{t-1}, Z_t = 0\big] + \\
&\quad \varepsilon_t\,\mathbb{E}\big[\mathbf{1}\big(s_h^t = s, a_h^t = a\big)\,\big|\mathcal{H}_{t-1}, Z_t = 1\big], \\
&\geqslant \varepsilon_t\, w_h^{c_t}(s,a)\,.
\end{aligned}$$

Next, when $\mathcal{E}_E^{\mathrm{cnt}}(T)$ holds,

$$\begin{aligned}
n_t(s,a,h) &\geqslant \frac{1}{2}\sum_{k=1}^{t-1}\mathbb{E}\big[\mathbf{1}\big(s_h^k = s, a_h^k = a\big)\,\big|\mathcal{H}_{k-1}\big] - \log(6SAHT^2)\,, \\
&\geqslant \frac{1}{2}\sum_{k=1}^{t-1}\varepsilon_k w_h^{c_k}(s,a) - \log(6SAHT^2),
\end{aligned}$$

and the right-hand side is related to the dynamic regret of the UCBVI forced exploration routine as

$$\sum_{k=1}^{t-1}\varepsilon_k w_h^{c_i}(s,a) \;\geqslant\; \sum_{k=1}^{t-1}\varepsilon_k \mathbf{1}\Big(\tilde{s}^k = s, \tilde{a}^k = a, \tilde{h}^k = h\Big)\, V_{t,0}^{c_k}. \tag{E.16}$$

Thanks to Lemma 11, on the event $\mathcal{E}_E(T)$, with $S_k \triangleq \mathbf{1}\Big(\tilde{s}^k = s, \tilde{a}^k = a, \tilde{h}^k = h\Big)$, and $\kappa_k \triangleq \varepsilon_k S_k$

$$\sum_{k=1}^{t-1}\kappa_k V_{k,0}^{c_k} \geqslant \sum_{k=1}^{t-1}\kappa_k V_{k,0}^{\star} - O(\sqrt{T^{1-\gamma}}) - O(\log T) - O\Big(\sqrt{t^{1-\gamma}\cdot\beta^p(T,1/T^2)}\Big)\,.$$

Rearranging the above displays yields

$$n_t(s,a,h) \geqslant \\ \frac{1}{2}\sum_{k=1}^{t-1}\left[\varepsilon_k \mathbf{1}\Big(\tilde{s}^k = s, \tilde{a}^k = a, \tilde{h}^k = h\Big)\right] \max_{\rho} w_h^{\rho}(s,a) - O(\sqrt{T^{1-\gamma}}) - O(\log T) - O\Big(\sqrt{t^{1-\gamma}\cdot\beta^p(T,1/T^2)}\Big).$$

We remark that on the event $\mathcal{E}_E^Z(T)$,

$$\sum_{k=1}^{t-1}\varepsilon_k \cdot \mathbf{1}\Big(\tilde{s}^k = s, \tilde{a}^k = a, \tilde{h}^k = h\Big) \geqslant \frac{1}{2SAH}\sum_{k=1}^{t-1}\varepsilon_k - \log(2SAHT^2)$$

we have for any $t \in (T^{2/3}, T)$, and any triplet (s,a,h)

$$n_t(s,a,h) \quad \geqslant \quad \frac{W_h(s,a)}{2SAH}\varepsilon_{1\dots t-1} - O\left(\sqrt{t^{\frac{3(1-\gamma)}{2}}}\right) - O(\log t) - O\left(\sqrt{t^{1-\gamma}\cdot\beta^p(t^{3/2},1/t^3)}\right). \tag{E.17}$$

Now, let us define

$$T_0(s,a,h) \triangleq \sup\left\{t \geqslant 1 : n_t(s,a,h) \leqslant \frac{W_h(s,a)}{4SAH}t^{1-\gamma}\right\}. \tag{E.18}$$

From Eq. (E.17), and, as $\varepsilon_{1\dots t} \sim \frac{t^{1-\gamma}}{1-\gamma}$, we know that on $\mathcal{E}_E(T)$, $k_0(s,a,h)$ is upper bounded by a deterministic constant $k_0(s,a,h)$ defined as

$$k_0(s,a,h)^{2/3} \triangleq \sup\Bigg\{t \geqslant 1 : \frac{\varepsilon_{1\dots t}}{2SAH}W_h(s,a) - \frac{t^{1-\gamma}}{4SAH}W_h(s,a) \leqslant O\left(\sqrt{t^{\frac{3(1-\gamma)}{2}}}\right) + \\ O(\log t) + O\left(\sqrt{t^{1-\gamma}\cdot\beta^p(t^{3/2},1/t^3)}\right)\Bigg\}.$$

Indeed, the expressions in $O(\cdot)$ can be recovered from the proof to make the above well-defined, moreover, it is finite for any $\gamma \in (0,1)$ and any reachable (s,a,h).

Next we define

$$k_0 \triangleq \max_{(s,a,h):W_h(s,a)>0} k_0(s,a,h),$$

which concludes the proof. □

E.3 Guarantees on Model Estimation

We prove that after a finite number of episodes the empirical optimal policy coincides with the true optimal policy on most episodes, when $\mathcal{E}_E(T)$ holds. The key tool is the following value difference lemma.

Lemma 13 (Value Difference)**.** *For any policy $\pi \in \Pi$,*

$$V_0^{\pi} - \hat{V}_{t,0}^{\pi} = \mathbb{E}_M^{\pi}\left[\sum_{h=1}^{H} b_t(s_h,a_h,h)\right],$$

where $b_t(s,a,h) \triangleq \mu(s,a,h) - \hat{\mu}_t(s,a,h) + \big(p_h - \hat{p}_{t,h}\big)\hat{V}_{t,h+1}^{\pi}(s,a)$.

Proof. By Bellman's equations,

$$V_h^{\pi}(s) - \hat{V}_{t,h}^{\pi}(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\Big[\mu_h(s,a) - \hat{\mu}_t(s,a,h) + p_h V_{h+1}^{\pi}(s,a) - \hat{p}_{t,h}\hat{V}_{t,h+1}^{\pi}(s,a)\Big] \\ = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\Big[b_t(s,a,h) + p_h\big(V_{h+1}^{\pi} - \hat{V}_{t,h+1}^{\pi}\big)(s,a)\Big].$$

Applying this recursion from h down to H and using $s_1 = s_{\text{init}}$ gives the claimed result. □

Combining Lemma 13 with the sufficient exploration guarantee of Section E.2 and standard concentration bounds shows that the empirical value of any policy converges quickly to its true value. In particular, once the right-hand side of Lemma 13 falls below

$$\Delta_{\min}(\Pi) \triangleq V_0^\star - \max_{\pi \in \Pi \setminus \{\pi^\star\}} V_0^\pi, \tag{E.19}$$

the empirical optimal policy $\hat{\pi}_t$ must equal $\pi^\star$. This is formalized in the next lemma.

Lemma 14. *There exists an explicit finite constant $k_0 < \infty$ such that on the event $\mathcal{E}_E(T)$, for all $T > k_0$ and all $t \in (T^{2/3}, T)$,*

$$\hat{\pi}_t = \pi^\star.$$

The constant k_0 is given explicitly in the proof.

Proof. Since $\pi^\star$ is the unique optimal start-state policy for M, it suffices to show that $\hat{V}_{t,0}^{\pi^\star} > \hat{V}_{t,0}^{\pi}$ for all $\pi \neq \pi^\star$.

By Lemma 13, for any $\pi \in \Pi$, $V_0^\pi - \hat{V}_{t,0}^\pi = \mathbb{E}_M^\pi\left[\sum_{h=1}^H b_t(s_h, a_h, h)\right]$.

Since $\hat{V}_{t,h+1}^\pi(s) \leqslant H$ for all (s,h), Pinsker's inequality gives, on $\mathcal{E}_E(T)$,

$$|b_t(s,a,h)| \leqslant \sqrt{\frac{2\beta^\mu(t,1/t^2)}{n_t(s,a,h)}} + \sqrt{\frac{2H^2\beta^p(t,1/t^2)}{n_t(s,a,h)}}.$$

By Lemma 12, on $\mathcal{E}_E(T)$, there exists another $k_0' < \infty$ such that for $T \geqslant k_0'$, and $t \in (T^{2/3}, T)$, and any reachable (s,a,h),

$$n_t(s,a,h) \geqslant \frac{W_h(s,a)}{4SAH} t^{1-\gamma}$$

. Therefore, when $\mathcal{E}_E(T)$ holds, for $T \geqslant k_0'$, and $t \in (T^{2/3}, T)$,

$$|b_t(s,a,h)| \leqslant \sqrt{\frac{8SAH\,\beta^\mu(t,1/t^2)}{W_h(s,a)\,t^{1-\gamma}}} + \sqrt{\frac{8SAH^3\beta^p(t,1/t^2)}{W_h(s,a)\,t^{1-\gamma}}}. \tag{E.20}$$

Since $w_h^\pi(s,a) = 0$ for unreachable triplets (s,a,h),

$$\left|V_0^\pi - \hat{V}_{t,0}^\pi\right| = \left|\sum_{\substack{s,a,h:\\ W_h(s,a)>0}} w_h^\pi(s,a)\, b_t(s,a,h)\right| \leqslant \sum_{\substack{s,a,h:\\ W_h(s,a)>0}} w_h^\pi(s,a)\left[\sqrt{\frac{8SAH\,\beta^\mu(t,1/t^2)}{W_h(s,a)\,t^{1-\gamma}}} + \sqrt{\frac{8SAH^3\beta^p(t,1/t^2)}{W_h(s,a)\,t^{1-\gamma}}}\right].$$

We define

$$T_0^{2/3} \triangleq \max_{\substack{s,a,h:\\ W_h(s,a)>0}} \sup\left\{t \geqslant 1 : \sqrt{\frac{8SAH\,\beta^\mu(t,1/t^2)}{W_h(s,a)\,t^{1-\gamma}}} + \sqrt{\frac{8SAH^3\beta^p(t,1/t^2)}{W_h(s,a)\,t^{1-\gamma}}} \geqslant \frac{\Delta_{\min}(\Pi)}{2}\right\}, \tag{E.21}$$

which is finite since β^μ, β^p are logarithmic in t, $\gamma \in (0,1)$, and $\Delta_{\min}(\Pi) > 0$. We further let $k_0 = \max(k_0', T_0)$.

Thus, when $\mathcal{E}_E$ holds, for $T \geqslant k_0$, $t \in (T^{2/3}, T)$ and any $\pi \neq \pi^\star$,

$$\hat{V}_{t,0}^{\pi^\star} - \hat{V}_0^{t,\pi} \geqslant \Delta_{\min}(\Pi) + \underbrace{\hat{V}_{t,0}^{\pi^\star} - V_0^{\pi^\star}}_{> -\Delta_{\min}(\Pi)/2} + \underbrace{V_0^\pi - \hat{V}_{t,0}^\pi}_{> -\Delta_{\min}(\Pi)/2} > 0,$$

so that $\hat{\pi}_t = \pi^\star$. □

F ANALYSING CONDITIONAL POSTERIOR SAMPLING

In this section, we analyze the guarantees on the inf player (distribution over $\mathrm{Alt}(\hat{M}_T)$) by interpreting it as an instance of continuous exponential weights (CEW). The goal of these results is to demonstrate that conditional posterior sampling, akin to continuous exponential weights, is a no-regret learner of the best challenger MDP.

We recall the log-likelihood ratio between a model (θ, Φ) and the empirical MDP $\hat{M}_T$ after $(T-1)$ episodes

$$\Upsilon_T(\theta, \Phi) \triangleq \sum_{s,a,h} n_T(s,a,h) \left[\frac{\big(\theta(s,a,h) - \hat{\mu}_T(s,a,h)\big)^2}{2} + \sum_{s'} \hat{p}_T(s' \mid s,a,h) \log \frac{\hat{p}_t(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right] . \tag{F.1}$$

The conclusion of this section is to prove the result below.

Theorem 15. *Let $\eta_t \propto t^{-\varsigma}$ with $\varsigma > 2\alpha$. There exists a constant $k_0 < \infty$ and a deterministic function $f(T) = o(T)$ such for any $T \geqslant k_0$, the inequality*

$$\sum_{t=T^{2/3}}^{T-1} \mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))} \left[\ell_t^\alpha(\theta,\Phi)\right] \leqslant$$
$$\inf_{(\theta,\Phi)\in\mathrm{Alt}(\hat{M}_T)} \sum_{t=1}^{T-1} \sum_h \left[\frac{(\hat{\mu}_T(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h))^2}{2} + \log \frac{1}{\Phi(s_{h+1}^t, \mid s_h^t, a_h^t, h)} \right] + f(T) \,, \text{ where}$$

$$\ell_t^\alpha(\theta,\Phi) \triangleq \sum_h \frac{\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h)\big)^2}{2} + \log \frac{1}{\max\left(\Phi(s_{h+1}^t \mid s_h^t, a_h^t, h), e^{-t^\alpha}\right)} \,,$$

holds with probability at least $1 - O(1/T^2)$. Moreover, f can be made explicit from the proof.

F.1 Setup

Before introducing the proof of this result, let us introduce the function over reward parameters and the function over the transition kernels defined for some $\eta > 0$ by

$$\begin{aligned} f_t^\eta(\theta) &\triangleq \exp\left\{ -\sum_{s,a,h} \frac{\eta \cdot n_t(s,a,h)}{2} (\hat{\mu}_t(s,a,h) - \theta(s,a,h))^2 \right\} , \\ g_t^\eta(\Phi) &\triangleq \exp\left\{ \sum_{s,a,h} \sum_{s'} \eta \cdot n_t(s'|s,a,h) \log \Phi(s'|s,a,h) \right\} . \end{aligned} \tag{F.2}$$

Since f_t^η, g_t^η are random variables, all results claimed in the next lemma hold with probability 1.

Lemma 16. *Let $\eta > 0$ and let $M \triangleq (\theta, \Phi) \in \mathbf{M}$ be any MDP with positive transitions. After observing the trajectory $\big((s_h^t, a_h^t, r_h^t, s_{h+1}^t)\big)_{h=1}^H$ at episode t, the following identities hold with probability 1:*

$$\frac{g_{t+1}^\eta(\Phi)}{g_t^\eta(\Phi)} = \exp\left\{ -\eta \sum_{h=1}^H \log \frac{1}{\Phi(s_{h+1}^t \mid s_h^t, a_h^t, h)} \right\}, \tag{F.3}$$

$$\frac{f_{t+1}^\eta(\theta)}{f_t^\eta(\theta)} = \exp\left\{ -\frac{\eta}{2} \sum_{h=1}^H \big(\hat{\mu}_t(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h)\big)^2 + \frac{\eta}{2}\, \sigma_{t+1}(\theta) \right\}, \tag{F.4}$$

where

$$\sigma_{t+1}(\theta) \triangleq \sum_{h=1}^H n_{t+1}(s_h^t, a_h^t, h) \Big[\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h)\big)^2 - \big(\hat{\mu}_{t+1}(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h)\big)^2 \Big]. \tag{F.5}$$

Proof. Noting that any transition $(s' \mid s, a, h)$ not observed at the end of episode t satisfies $n_t(s' \mid s, a, h) = n_{t+1}(s' \mid s, a, h)$, the definitions in Eq. (F.2) give

$$\begin{aligned}
\frac{g^\eta_{t+1}(\Phi)}{g^\eta_t(\Phi)} &= \exp\Bigg\{\sum_h \eta \cdot n_{t+1}(s^t_{h+1}|s^t_h, a^t_h, h) \log \Phi(s^t_{h+1}|s^t_h, a^t_h, h) - \\
&\qquad \sum_h \eta \cdot n_t(s^t_{h+1} \mid s^t_h, a^t_h, h) \log \Phi(s^t_{h+1}|s^t_h, a^t_h, h)\Bigg\}, \\
&\overset{(a)}{=} \exp\left\{\sum_h \eta \log \Phi(s^t_{h+1} \mid s^t_h, a^t_h, h)\right\}, \\
&= \exp\left\{-\eta \sum_h \log \frac{1}{\Phi(s^t_{h+1} \mid s^t_h, a^t_h, h)}\right\},
\end{aligned}$$

where (a) follows since the observed transition $(s^t_h, a^t_h, s^t_{h+1})$ contributes exactly $+1$ to the count.

The proof of the second statement of the lemma follows by noting that

$$\begin{aligned}
\frac{f^\eta_{t+1}(\theta)}{f^\eta_t(\theta)} &= \exp\Bigg\{-\sum_h \frac{\eta \cdot n_{t+1}(s^t_h, a^t_h, h)}{2}(\hat{\mu}_{t+1}(s^t_h, a^t_h, h) - \theta(s^t_h, a^t_h, h))^2 + \\
&\qquad \sum_h \frac{\eta \cdot n_t(s^t_h, a^t_h, h)}{2}(\hat{\mu}_t(s^t_h, a^t_h, h) - \theta(s^t_h, a^t_h, h))^2\Bigg\}, \\
&\overset{(b)}{=} \exp\Bigg\{\sum_h \frac{\eta \cdot n_{t+1}(s^t_h, a^t_h, h)}{2}\Big[(\hat{\mu}_t(s^t_h, a^t_h, h) - \theta(s^t_h, a^t_h, h))^2 - (\hat{\mu}_{t+1}(s^t_h, a^t_h, h) - \theta(s^t_h, a^t_h, h))^2\Big] - \\
&\qquad \sum_h \frac{\eta}{2}(\hat{\mu}_t(s^t_h, a^t_h, h) - \theta(s^t_h, a^t_h, h))^2\Bigg\},
\end{aligned}$$

where (b) follows since $n_{t+1}(s^t_h, a^t_h, h) = n_t(s^t_h, a^t_h, h) + 1$. Indeed, since (s^t_h, a^t_h) is visited at stage h during episode t, the count increases by exactly 1.

With the definition of σ_{t+1}, we have proven the claimed statement. □

F.2 Posterior Density Ratio and Stochastic Hedge Loss

We introduce further notation related to the posterior density. As the transition and rewards are independent, it is a simple exercise to show that the posterior density is simply the product of $g^{\eta_t}_t$ and $f^{\eta_t}_t$.

For $\eta > 0$, we introduce λ^η defined on $\mathbf{M}$ as

$$\lambda_t(\theta, \Phi) \triangleq -\frac{1}{\eta} \log\left\{f^\eta_t(\theta) g^\eta_t(\Phi)\right\}, \tag{F.6}$$

for an MDP $\widetilde{M} \triangleq (\theta, \Phi)$, and we define a density w.r.t. canonical measure of $\mathbf{M}$ proportional to

$$f^\eta_t(\theta) g^\eta_t(\Phi) \cdot \mathbf{1}\Big((\theta, \Phi) \in \mathrm{Alt}(\hat{M}_t)\Big)$$

whose normalizing constant is

$$\Lambda^\eta_t \triangleq \int f^\eta_t(\theta) g^\eta_t(\Phi) \cdot \mathbf{1}\Big((\theta, \Phi) \in \mathrm{Alt}(\hat{M}_t)\Big)\, \mathrm{d}\theta \mathrm{d}\Phi = \int e^{-\eta \lambda_t(\theta, \Phi)} \cdot \mathbf{1}\Big((\theta, \Phi) \in \mathrm{Alt}(\hat{M}_t)\Big)\, \mathrm{d}\theta \mathrm{d}\Phi\,, \tag{F.7}$$

where we recall that $\mathrm{Alt}(\hat{M}_t) \subset \mathbf{M}$ and $\mathrm{d}\theta\mathrm{d}\Phi$ denotes the standard (product) measure on this space.

We denote by $\nu^\eta_t(\cdot)$ the distribution over $\mathbf{M}$ with product density $f^\eta_t \cdot g^\eta_t$ and by $\nu^\eta_t(\cdot \mid \mathcal{X})$ its truncattion to the set $\mathcal{X} \subseteq \mathbf{M}$.

Given $\mathcal{H}_{t-1}$, the probability of sampling an MDP $\widetilde{M} \triangleq (\theta, \Phi)$ from $\nu_t^\eta(\cdot \mid \mathrm{Alt}(\hat{M}_t))$ is proportional to

$$e^{-\eta \lambda_t(\theta,\Phi)} \cdot \mathbf{1}\Big((\theta,\Phi) \in \mathrm{Alt}(\hat{M}_t)\Big).$$

Lemma 17. *Let $\eta > 0$. If $\hat{\pi}_t = \hat{\pi}_{t+1}$ then, we have*

$$\log \frac{\Lambda_{t+1}^\eta}{\Lambda_t^\eta} \leqslant \frac{1}{2} \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^\eta(\cdot|\mathrm{Alt}(\hat{M}_t))} \left[e^{-2\eta \cdot \ell_t^\alpha(\theta,\Phi)}\right] + \frac{1}{2} \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^\eta(\cdot|\mathrm{Alt}(\hat{M}_t))} \left[e^{\eta \cdot \sigma_{t+1}(\theta)}\right].$$

Proof. First, we note that given and MDP M, the set of MDPs M' such that M is not absolutely continuous w.r.t. M' has empty interior. Therefore $\mathrm{Alt}(\hat{M}_t) = \mathrm{Alt}(\hat{M}_{t+1})$ (up to a null set) whenever $\hat{\pi}_t = \hat{\pi}_{t+1}$. We have

$$\begin{aligned}
\log \frac{\Lambda_{t+1}^\eta}{\Lambda_t^\eta} &= \log \frac{\int f_{t+1}^\eta(\theta)\, g_{t+1}^\eta(\Phi) \cdot \mathbf{1}\Big((\theta,\Phi) \in \mathrm{Alt}(\hat{M}_t)\Big)\, \mathrm{d}\theta\, \mathrm{d}\Phi}{\int f_t^\eta(\theta)\, g_t^\eta(\Phi) \cdot \mathbf{1}\Big((\theta,\Phi) \in \mathrm{Alt}(\hat{M}_t)\Big)\, \mathrm{d}\theta\, \mathrm{d}\Phi} \\
&= \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^\eta(\cdot|\mathrm{Alt}(\hat{M}_t))} \left[\frac{f_{t+1}^\eta(\theta)}{f_t^\eta(\theta)} \cdot \frac{g_{t+1}^\eta(\Phi)}{g_t^\eta(\Phi)}\right] \\
&= \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^\eta(\cdot|\mathrm{Alt}(\hat{M}_t))} \Bigg[\exp\Bigg\{-\frac{\eta}{2} \sum_h \Bigg[\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h)\big)^2 + \\
&\qquad 2 \log \frac{1}{\Phi(s_{h+1}^t \mid s_h^t, a_h^t, h)}\Bigg] + \frac{\eta}{2}\, \sigma_{t+1}(\theta)\Bigg\}\Bigg] \\
&\overset{(a)}{\leqslant} \frac{1}{2} \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^\eta(\cdot|\mathrm{Alt}(\hat{M}_t))} \left[\exp\left\{-\eta \sum_h \left[\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h)\big)^2 + 2 \log \frac{1}{\Phi(s_{h+1}^t \mid s_h^t, a_h^t, h)}\right]\right\}\right] \\
&\quad + \frac{1}{2} \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^\eta(\cdot|\mathrm{Alt}(\hat{M}_t))} \left[e^{\eta \cdot \sigma_{t+1}(\theta)}\right],
\end{aligned}$$

where (a) follows from the Cauchy–Schwarz inequality $\log \mathbb{E}[XY] \leqslant \frac{1}{2} \log \mathbb{E}[X^2] + \frac{1}{2} \log \mathbb{E}[Y^2]$.

Intuitively, everything proceeds as if we were running continual exponential weights with learning rate η and stochastic loss

$$l_t(\theta,\Phi) \triangleq \sum_{h=1}^H \left[\frac{\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h)\big)^2}{2} + \log \frac{1}{\Phi(s_{h+1}^t \mid s_h^t, a_h^t, h)}\right]. \tag{F.8}$$

However, to control the magnitude of the stochastic loss, which might be unbounded due to low transition probabilities, we work with the clipped loss

$$\ell_t^\alpha(\theta,\Phi) \triangleq \sum_{h=1}^H \left[\frac{\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h)\big)^2}{2} + \log \frac{1}{\max\big(\Phi(s_{h+1}^t \mid s_h^t, a_h^t, h), e^{-t^\alpha}\big)}\right], \tag{F.9}$$

which satisfies $\ell_t^\alpha(\theta,\Phi) \leqslant l_t(\theta,\Phi)$ by definition. Combining with the calculations above, we have with probability 1,

$$\log \frac{\Lambda_{t+1}^\eta}{\Lambda_t^\eta} \leqslant \frac{1}{2} \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^\eta(\cdot|\mathrm{Alt}(\hat{M}_t))} \Big[e^{-2\eta\, \ell_t^\alpha(\theta,\Phi)}\Big] + \frac{1}{2} \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^\eta(\cdot|\mathrm{Alt}(\hat{M}_t))} \Big[e^{\eta\, \sigma_{t+1}(\theta)}\Big].$$

□

We further define

$$\varphi_{t+1}^\eta \triangleq \frac{1}{2} \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^\eta(\cdot|\mathrm{Alt}(\hat{M}_t))} \Big[e^{\eta\, \sigma_{t+1}(\theta)}\Big], \tag{F.10}$$

which will be later upper bounded in Lemma 21.

F.3 Hedge with Decreasing Learning Rate

We use a decreasing learning rate schedule $(\eta_t)_{t\geqslant 1}$ rather than a constant one, since a constant rate would require the doubling trick and discarding of past data. Throughout, we assume $(\eta_t)_{t\geqslant 1}$ is non-increasing and satisfies $\sum_{t=1}^{T}\eta_t = o(T)$. For generality, we assume $\eta_t = t^{-\varsigma}$.

Lemma 18. *Let $T_0 \leqslant T$ and assume the event $\mathcal{E}_F^{T_0}(T) \triangleq \{\forall\, T_0 \leqslant t \leqslant T,\ \hat{\pi}_t = \hat{\pi}_{T_0}\}$ holds. Define*

$$\psi_{t+1}^{\alpha} \triangleq \frac{\eta_t C_{t,\alpha}^2}{2} + \eta_t^{-1}\varphi_{t+1}^{\eta_t}, \qquad \Psi_T^{\alpha}(T_0) \triangleq \sum_{t=T_0}^{T-1}\psi_{t+1}^{\alpha},$$

where $C_{t,\alpha} \triangleq H\big(\frac{1}{2}+t^{\alpha}\big)$ and $\varphi_{t+1}^{\eta_t}$ is defined in (F.10). Then

$$\sum_{t=T_0}^{T-1}\mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}\big[\ell_t^{\alpha}(\theta,\Phi)\big] \leqslant -\frac{1}{\eta_T}\log\Lambda_T^{\eta_T} + \Psi_T^{\alpha}(T_0) + \frac{1}{\eta_{T_0}}\log\Lambda_{T_0}^{\eta_{T_0}} + \sum_{t=T_0}^{T-1}\left[\frac{1}{\eta_t}\log\Lambda_{t+1}^{\eta_t} - \frac{1}{\eta_{t+1}}\log\Lambda_{t+1}^{\eta_{t+1}}\right].$$

Proof. Thanks to Lemma 17 we have

$$\log\frac{\Lambda_{t+1}^{\eta}}{\Lambda_t^{\eta}} \leqslant \frac{1}{2}\log\mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta}(\cdot|\mathrm{Alt}(\hat{M}_t))}\Big[e^{-2\eta\,\ell_t^{\alpha}(\theta,\Phi)}\Big] + \varphi_{t+1}^{\eta}$$

where φ_{t+1}^{η} is defined in Eq. (F.10). Next, we note that $\ell_t^{\alpha} \in (0, (\frac{1}{2}+t^{\alpha})\cdot H)$, and we define $C_{t,\alpha} \triangleq H\,(\frac{1}{2}+t^{\alpha})$. By Hoeffding's lemma we have for any η,

$$\mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta}(\cdot|\mathrm{Alt}(\hat{M}_t))}\Big[e^{-2\eta\,\ell_t^{\alpha}(\theta,\Phi)}\Big] \leqslant \exp\left\{-2\eta\mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta}(\cdot|\mathrm{Alt}(\hat{M}_t))}[\ell_t^{\alpha}(\theta,\Phi)] + \frac{\eta^2 C_{t,\alpha}^2}{2}\right\}. \tag{F.11}$$

Therefore,

$$\frac{1}{\eta_t}\log\frac{\Lambda_{t+1}^{\eta_t}}{\Lambda_t^{\eta_t}} \leqslant -\mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}\big[\ell_t^{\alpha}(\theta,\Phi)\big] + \underbrace{\frac{1}{2}\eta_t C_{t,\alpha}^2 + \eta_t^{-1}\varphi_{t+1}^{\eta_t}}_{\psi_{t+1}^{\alpha}}. \tag{F.12}$$

Assuming $\mathcal{E}_F^{T_0}(T)$ and summing (F.12) from $t=T_0$ to $T-1$ gives

$$\sum_{t=T_0}^{T-1}\frac{1}{\eta_t}\log\frac{\Lambda_{t+1}^{\eta_t}}{\Lambda_t^{\eta_t}} \leqslant -\sum_{t=T_0}^{T-1}\mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}\big[\ell_t^{\alpha}(\theta,\Phi)\big] + \underbrace{\sum_{t=T_0}^{T-1}\psi_{t+1}^{\alpha}}_{\Psi_T^{\alpha}(T_0)}. \tag{F.13}$$

The left-hand side is not directly telescoping. We decompose it as

$$\begin{aligned}
\sum_{t=T_0}^{T-1}\frac{1}{\eta_t}\log\frac{\Lambda_{t+1}^{\eta_t}}{\Lambda_t^{\eta_t}} &= \sum_{t=T_0}^{T-1}\left[\frac{1}{\eta_t}\log\Lambda_{t+1}^{\eta_t} - \frac{1}{\eta_t}\log\Lambda_t^{\eta_t}\right] \\
&= \sum_{t=T_0}^{T-1}\left[\frac{1}{\eta_{t+1}}\log\Lambda_{t+1}^{\eta_{t+1}} - \frac{1}{\eta_t}\log\Lambda_t^{\eta_t}\right] + \sum_{t=T_0}^{T-1}\left[\frac{1}{\eta_t}\log\Lambda_{t+1}^{\eta_t} - \frac{1}{\eta_{t+1}}\log\Lambda_{t+1}^{\eta_{t+1}}\right] \\
&= \frac{1}{\eta_T}\log\Lambda_T^{\eta_T} - \frac{1}{\eta_{T_0}}\log\Lambda_{T_0}^{\eta_{T_0}} + \sum_{t=T_0}^{T-1}\left[\frac{1}{\eta_t}\log\Lambda_{t+1}^{\eta_t} - \frac{1}{\eta_{t+1}}\log\Lambda_{t+1}^{\eta_{t+1}}\right],
\end{aligned}$$

where the last step uses that the first sum telescopes. Substituting back into Eq. (F.13) and rearranging gives the claimed inequality. □

F.4 Best Model in Hindsight

The idea here is to show a lower bound on this quantity when there is enough mass around the best model. While such bounds are standard to derive for Hedge on convex sets (with convex loss), in our case, although the loss is convex, the set $\mathrm{Alt}(\hat{M}_t)$ is non-convex and a priori not a simple union of convex sets.

Lemma 19. *Let* $\gamma \triangleq \frac{1}{6TSAH(H+1)}$. *We have*

$$\frac{1}{\eta_T}\log\Lambda_T^{\eta_T} \geqslant -\inf_{(\theta,\Phi)\in\mathrm{Alt}(\hat{M}_T)}\sum_{t=1}^{T-1}\sum_h\left[\frac{(\hat{\mu}_T(s_h^t,a_h^t,h)-\theta(s_h^t,a_h^t,h))^2}{2}+\log\frac{1}{\Phi(s_{h+1}^t,|\, s_h^t,a_h^t,h)}\right]$$
$$\frac{2(SAH+S^2AH)\log\gamma}{\eta}-\gamma\frac{TH}{2}\log(eS^2T)+\frac{\log\mathrm{vol}(\mathrm{Alt}(\hat{M}_T)\cap\mathbf{M}^{1/S\cdot\sqrt{T}})}{\eta}-O(1)\,.$$

Proof. The proof of this result relies on a lower bound on the inflated posterior over $\mathrm{Alt}(\hat{M}_T)$. Simplifying notation, let $\eta \triangleq \eta_T$ and we have

$$\Lambda_T^\eta = \int e^{-\eta\lambda_T(\theta,\Phi)}\mathbf{1}\Big(\theta,\Phi)\in\mathrm{Alt}(\hat{M}_T)\Big)\,\mathrm{d}\theta\mathrm{d}\,\Phi\,, \tag{F.14}$$

where λ_T^η is a convex function as defined earlier. We prove that the posterior puts the majority of its mass around the mode. First, we need to prove that there is enough volume around the best model

$$(\theta_T^\star,\Phi_T^\star)\in\operatorname*{argmin}_{(\theta,\Phi)\in\mathrm{cl}(\mathrm{Alt}(\hat{M}_T))}\lambda_T(\theta,\Phi), \tag{F.15}$$

so that the integral in Eq. (F.14) will concentrate around the mode. We define $\widetilde{M}_T^\star = (\theta_T^\star,\Phi_T^\star)$.

First, note that this argmin is well defined as $\mathrm{cl}(\mathrm{Alt}(\hat{M}_T))$ is a compact set. Because the instance in (F.15) might be close to the boundary of $\mathrm{Alt}(\hat{M}_T)$, we build another instance around $\widetilde{M}_T^\star$ which is still contained in the alternative set $\mathrm{Alt}(\hat{M}_T)$, and with a non-negligible enclosing volume around it.

Because $\widetilde{M}_T^\star$ belongs to the closure of the alternative set, there exists a deterministic policy $\tilde{\pi}\in\Pi_{\mathrm{det}}\backslash\{\hat{\pi}_t\}$ such that $\tilde{\pi}$ is globally optimal in $\widetilde{M}_T^\star$.

Next, we invoke Lemma 43 which shows that there exists some MDP $\widetilde{M}' \triangleq (\theta',\Phi')$ such that $\tilde{\pi}$ is the unique global optimal and start-state policy, and its transition and reward kernels satisfy

$$\begin{cases}\Delta(\widetilde{M}') & \geqslant \frac{1}{T}\\ \Phi'(\cdot\mid s,a,h) & \triangleq (1-\gamma_{s,a,h})\Phi_T^\star(\cdot\mid s,a,h)+\gamma_{s,a,h}\Phi_T^\star(\cdot\mid s,\pi_h(s),h)\\ |\theta_T^\star(s,a,h)-\theta'(s,a,h)| & \leqslant d_{s,a,h}\triangleq 2^{H-h+1}\varepsilon\end{cases}, \tag{F.16}$$

where $\gamma_{s,a,h}\in\left[0,\frac{(1+2^{H-h})\varepsilon}{1-(2^{H-h+1}+3)\varepsilon}\right]$.

In particular, $\widetilde{M}'\in\mathrm{Alt}(\hat{M}_T)$. Next, we relate Λ_T^η to the (near optimal) $\lambda_T^\eta(\theta',\Phi')$. Thanks to Lemma 44, for $\gamma \triangleq \frac{1}{6TSAH(H+1)}$, and calling $\widetilde{M}_T \triangleq (\theta_T,\Phi_T)$ the instance built by this lemma. We then have

$$(1-\gamma)\widetilde{M}'+\gamma\,\mathrm{Alt}(\hat{M}_T)\subset\mathrm{Alt}(\hat{M}_T). \tag{F.17}$$

Therefore,

$$\begin{aligned}
\log \Lambda_T^\eta &= \log \int e^{-\eta \lambda_T(\theta,\Phi)} \mathbf{1}\Big((\theta,\Phi) \in \mathrm{Alt}(\hat{M}_T)\Big)\, \mathrm{d}\theta \mathrm{d}\Phi \\
&\geqslant \log \int_{(1-\gamma)\widetilde{M}' + \gamma\, \mathrm{Alt}(\hat{M}_T)} e^{-\eta \lambda_T(\theta,\Phi)} \mathrm{d}\theta \mathrm{d}\Phi \\
&\geqslant \log \int_{\mathrm{Alt}(\hat{M}_T)} \gamma^{S^2AH+SAH} \exp\left\{-\eta \lambda_T((1-\gamma)\cdot(\theta',\Phi') + \gamma\cdot(\theta,\Phi)\right\} \mathrm{d}\theta \mathrm{d}\Phi \\
&\geqslant \log\left(\gamma^{S^2AH+SAH} e^{-\eta(1-\gamma)\cdot\lambda_T(\theta',\Phi')}\right) + \log \int_{\mathrm{Alt}(\hat{M}_T)} e^{-\eta\gamma\cdot\lambda_T(\theta,\Phi)} \mathrm{d}\theta \mathrm{d}\Phi \\
&= -\eta(1-\gamma)\lambda_T(\theta',\Phi') + (S^2AH+SAH)\log\gamma + \log \int_{\mathrm{Alt}(\hat{M}_T)} e^{-\eta\gamma\lambda_T(\theta,\Phi)} \mathrm{d}\theta \mathrm{d}\Phi
\end{aligned} \tag{F.18}$$

which follows by convexity of λ_T. Combining the results above, in particular in Eq. (F.16),

$$\begin{aligned}
\lambda_T(\theta',\Phi') &\leqslant \sum_{t=1}^{T-1}\sum_h \left[\frac{(\hat{\mu}_T(s_h^t,a_h^t,h) - \theta'(s_h^t,a_h^t,h))^2}{2} + \log\frac{1}{(1-\gamma_h)\Phi_T^\star(s_{h+1}^t \mid s_h^t,a_h^t,h)}\right] \\
&\leqslant \sum_{t=1}^{T-1}\sum_h \left[\frac{(\hat{\mu}_T(s_h^t,a_h^t,h) - \theta_T^\star(s_h^t,a_h^t,h))^2}{2} + \log\frac{1}{(1-\gamma_h)\Phi_T^\star(s_{h+1}^t \mid s_h^t,a_h^t,h)}\right] \\
&\quad + TH\left[4^H\varepsilon^2 + 2^H\varepsilon\right] \\
&\leqslant \lambda_T(\theta_T^\star,\Phi_T^\star) + TH\left[4^H\varepsilon^2 + 2^H\varepsilon + \log\frac{1}{1-\gamma_1}\right] \quad \text{(monotonicity of } h\mapsto\gamma_h\text{)}\,.
\end{aligned}$$

Therefore

$$\lambda_T(\theta',\Phi') \leqslant \lambda_T(\theta_T^\star,\Phi_T^\star) + TH\left[4^H\varepsilon^2 + 2^H\varepsilon + \log\frac{1}{1-\gamma_1}\right] \tag{F.19}$$

It remains to control the residual integral

$$\log \int_{\mathrm{Alt}(M)} e^{-\eta\gamma\lambda_T(\theta,\Phi)} \mathrm{d}\theta \mathrm{d}\Phi.$$

We note that if this is too small the bound above will be meaningless. Also, observe that the integraand is unbounded, (due to the transition probabilities) hence to control this quantity we introduce $\mathbf{M}^{1/S\cdot\sqrt{T}}$ to be the set ofs MDPs for which each transition is at least $1/S\cdot\sqrt{T}$. We have

$$\begin{aligned}
\int_{\mathrm{Alt}(M)} e^{-\eta\gamma\lambda_T(\theta,\Phi)} \mathrm{d}\theta \mathrm{d}\Phi &\geqslant \int_{\mathrm{Alt}(M)\cap\mathbf{M}^{1/\sqrt{T}}} e^{-\eta\gamma\lambda_T(\theta,\Phi)} \mathrm{d}\theta \mathrm{d}\Phi \\
&\geqslant \int_{\mathrm{Alt}(\hat{M}_T)\cap\mathbf{M}^{1/S\cdot\sqrt{T}}} \exp\left\{-\eta\gamma \sum_{s,a,h} n_t(s,a,h)\big(\frac{1}{2} + \frac{1}{2}\log(S^2T)\big)\right\} \mathrm{d}\theta \mathrm{d}\Phi \\
&= e^{-\eta\gamma\frac{TH}{2}\log(eS^2T)}\ \mathrm{vol}(\mathrm{Alt}(\hat{M}_T)\cap\mathbf{M}^{1/S\cdot\sqrt{T}})
\end{aligned}$$

Combining the above displays yields,

$$\log \int_{\mathrm{Alt}(M)} e^{-\eta\gamma\lambda_T(\theta,\Phi)} \mathrm{d}\theta \mathrm{d}\Phi \geqslant -\eta\gamma\frac{TH}{2}\log(eS^2T) + \log \mathrm{vol}(\mathrm{Alt}(\hat{M}_T)\cap\mathbf{M}^{1/S\cdot\sqrt{T}}). \tag{F.20}$$

Plugging-in this back into Eq. (F.18) yields

$$\frac{1}{\eta}\log\Lambda_T \geqslant -(1-\gamma)\lambda_T(\theta',\Phi') + \frac{2(SAH+S^2AH)\log\gamma}{\eta} - \gamma\frac{\log(eS^2T)}{2} + \frac{\log \mathrm{vol}(\mathrm{Alt}(\hat{M}_T)\cap\mathbf{M}^{1/S\cdot\sqrt{T}})}{\eta}$$

and combining with Eq. (F.19) yields

$$\frac{1}{\eta}\log\Lambda_T^\eta \geqslant$$
$$-\lambda_T(\theta_T^\star,\Phi_T^\star)+\frac{2(SAH+S^2AH)\log\gamma}{\eta}-\gamma\frac{TH}{2}\log(eS^2T)+\frac{\log\mathrm{vol}(\mathrm{Alt}(\hat{M}_T)\cap\mathbf{M}^{1/S\cdot\sqrt{T}})}{\eta}- \tag{F.21}$$
$$TH\left[4^H\varepsilon^2+2^H\varepsilon+\log\frac{1}{1-\gamma_1}\right].$$

Plugging $\varepsilon=1/T$ in the above, using the the inequality $\log(1+x)\leqslant x$ and recalling that $\gamma_1\in\left[0,\frac{(1+2^{H-1})}{T-(2^H+3)}\right]$ yields

$$H\left[\frac{4^H}{T}+2^H+T\frac{\gamma_1}{1-\gamma_1}\right]=O(1).$$

Rewriting the terms above, we have

$$\frac{1}{\eta}\log\Lambda_T^\eta\geqslant-\lambda_T(\theta_T^\star,\Phi_T^\star)-\frac{2(SAH+S^2AH)\log\gamma}{\eta}-\gamma\frac{TH}{2}\log(eS^2T)+\frac{\log\mathrm{vol}(\mathrm{Alt}(\hat{M}_T)\cap\mathbf{M}^{1/S\cdot\sqrt{T}})}{\eta}-O(1)$$

which achieves to prove the claimed result. □

F.5 Technical and Concentration Lemmas

We show a series of lemmas to complete the analysis of the conditional posterior resampling.

Lemma 20. *We have*

$$\sum_{t=T_0}^{T-1}\left[\frac{1}{\eta_t}\log\Lambda_{t+1}^{\eta_t}-\frac{1}{\eta_{t+1}}\log\Lambda_{t+1}^{\eta_{t+1}}\right]\leqslant\sum_{t=T_0}^{T-1}\left[\frac{\log\mathrm{vol}(\mathrm{Alt}(\hat{M}_{t+1}))}{\eta_t}-\frac{\log\mathrm{vol}(\mathrm{Alt}(\hat{M}_{t+1}))}{\eta_{t+1}}\right].$$

Proof. Let

$$\rho_t(\theta,\Phi)\triangleq\frac{1}{\mathrm{vol}(\mathrm{Alt}(\hat{M}_t))}\mathbf{1}\Big((\theta,\Phi)\in\mathrm{Alt}(\hat{M}_t)\Big)$$

be the uniform measure on $\mathrm{Alt}(\hat{M}_t)$. By definition of Λ_t^η,

$$\frac{1}{\eta}\log\Lambda_t^\eta=\frac{1}{\eta}\log\int\rho_t(\theta,\Phi)e^{-\eta\,\lambda_t(\theta,\Phi)}\mathrm{d}\theta\,\mathrm{d}\Phi+\frac{\log\mathrm{vol}(\mathrm{Alt}(\hat{M}_t))}{\eta}.$$

Introducing $f_t(\eta)\triangleq\frac{1}{\eta}\log\int\rho_t(\theta,\Phi)e^{-\eta\,\lambda_t(\theta,\Phi)}\,\mathrm{d}\theta\,\mathrm{d}\Phi$, we decompose

$$\sum_{t=T_0}^{T-1}\left[\frac{1}{\eta_t}\log\Lambda_{t+1}^{\eta_t}-\frac{1}{\eta_{t+1}}\log\Lambda_{t+1}^{\eta_{t+1}}\right]=$$
$$\sum_{t=T_0}^{T-1}\left[f_{t+1}(\eta_t)-f_{t+1}(\eta_{t+1})\right]+\sum_{t=T_0}^{T-1}\left[\frac{\log\mathrm{vol}(\mathrm{Alt}(\hat{M}_{t+1}))}{\eta_t}-\frac{\log\mathrm{vol}(\mathrm{Alt}(\hat{M}_{t+1}))}{\eta_{t+1}}\right].$$

By Lemma 38, f_{t+1} is non-decreasing. Since $(\eta_t)_t$ is non-increasing, $\eta_t\geqslant\eta_{t+1}$ for all t, so $f_{t+1}(\eta_t)-f_{t+1}(\eta_{t+1})\geqslant 0$, which concludes the proof. □

Lemma 21. *Define $C_{t,\alpha}\triangleq H\big(\frac{1}{2}+t^\alpha\big)$ and assume $\eta_t\propto t^{-\varsigma}$, with $\varsigma>2\alpha$. There exists deterministic f_Ψ such that the event*

$$\mathcal{E}_F^\Psi(T)\triangleq\left\{\Psi_T^\alpha(T_0)\triangleq\sum_{t=T_0}^{T-1}\left[\frac{\eta_tC_{t,\alpha}^2}{2}+\eta_t^{-1}\varphi_{t+1}^{\eta_t}\right]\leqslant f_\Psi(T)\right\}$$

holds with probability $1-O(1/T^2)$. Moreover, f_Φ is given in the proof and satisfies $f_\Psi=o(T)$

Proof. We bound each term in $\Psi_T^\alpha(T_0)$ separately. For the first term, we observe that

$$\sum_{t=T_0}^{T-1} \eta_t C_{t,\alpha}^2 \leqslant O\left(\sum_{t=1}^{T-1} t^{2\alpha} t^{-\varsigma}\right) = o(T),$$

when $\varsigma > 2\alpha$.

To control the second term in Ψ, recalling the definition of $\varphi_{t+1}^{\eta_t}$ from (F.10) and $\sigma_{t+1}(\theta)$ from (F.5), we have

$$\begin{aligned}
\sigma_{t+1}(\theta) &\leqslant \sum_h 2\, n_{t+1}(s_h^t, a_h^t, h)\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \theta(s_h^t, a_h^t, h)\big)\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \hat{\mu}_{t+1}(s_h^t, a_h^t, h)\big) \\
&= \underbrace{\sum_h 2\, n_{t+1}(s_h^t, a_h^t, h)\, \hat{\mu}_t(s_h^t, a_h^t, h)\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \hat{\mu}_{t+1}(s_h^t, a_h^t, h)\big)}_{\mathcal{J}_{t+1}} \\
&\quad \underbrace{- \sum_h 2\, n_{t+1}(s_h^t, a_h^t, h)\, \theta(s_h^t, a_h^t, h)\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \hat{\mu}_{t+1}(s_h^t, a_h^t, h)\big)}_{\kappa_{t+1}(\theta)}.
\end{aligned}$$

The term $\mathcal{J}_{t+1}$ is independent of (θ, Φ) and is controlled by Azuma–Hoeffding (see Lemma 22 below). For $\kappa_{t+1}(\theta)$, applying Hoeffding's lemma with the bound $|\kappa_{t+1}(\theta)| \leqslant \sqrt{\phi_{t+1}}$, where

$$\phi_{t+1} \triangleq \left(\sum_h 2\, n_{t+1}(s_h^t, a_h^t, h)\big|\hat{\mu}_t(s_h^t, a_h^t, h) - \hat{\mu}_{t+1}(s_h^t, a_h^t, h)\big|\right)^2,$$

gives

$$\eta_t^{-1} \log \mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}\left[e^{-\eta_t \kappa_{t+1}(\theta)}\right] \leqslant -\mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}[\kappa_{t+1}(\theta)] + \frac{1}{2}\eta_t \phi_{t+1}.$$

Summing over t and applying concentration to both $\mathcal{J}_{t+1}$ and ϕ_{t+1} (via Lemmas 22, 23, and 24) shows that $\sum_{t=K}^{T-1} \eta_t^{-1} \varphi_{t+1}^{\eta_t} = o(T)$ on an event with probability larger than $1 - O(1/T^2)$. The conclusion then follows. □

Lemma 22. *There exists a deterministic function $f_{\mathcal{J}}$ such that the event*

$$\mathcal{E}_F^{\mathcal{J}}(T)\left\{\mathcal{J}_T \triangleq \sum_{t=1}^{T}\sum_{h=1}^{H} 2\, n_{t+1}(s_h^t, a_h^t, h)\, \hat{\mu}_t(s_h^t, a_h^t, h)\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \hat{\mu}_{t+1}(s_h^t, a_h^t, h)\big) \leqslant f_{\mathcal{J}}(T)\right\},$$

holds with probability at least $1 - O(1/T^2)$. Moreover $f_{\mathcal{J}}$ is given in the proof and satisfies $f_{\mathcal{J}} = o(T)$.

Proof. Define $S_t(s, a, h) \triangleq \sum_{k\in[t-1]} \mathbf{1}\big(s_h^k = s,\, a_h^k = a\big)\, R_h^k$, where $R_h^k \mid s_h^k, a_h^k \sim \mu(s_h^k, a_h^k, h) + \zeta_h^k$ and $\zeta_h^k \sim \mathcal{N}(0,1)$ is independent noise. A direct computation gives

$$\begin{aligned}
n_{t+1}(s_h^t, a_h^t, h)\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \hat{\mu}_{t+1}(s_h^t, a_h^t, h)\big) &= S_t(s_h^t, a_h^t, h) - S_{t+1}(s_h^t, a_h^t, h) + \hat{\mu}_t(s_h^t, a_h^t, h) \\
&= \hat{\mu}_t(s_h^t, a_h^t, h) - \mu(s_h^t, a_h^t, h) + \zeta_h^t,
\end{aligned}$$

so that

$$\mathcal{J}_T = \underbrace{2\sum_{t=1}^{T}\sum_{h=1}^{H} \hat{\mu}_t(s_h^t, a_h^t, h)\big(\hat{\mu}_t(s_h^t, a_h^t, h) - \mu(s_h^t, a_h^t, h)\big)}_{(I)} + \underbrace{2\sum_{t=1}^{T}\sum_{h=1}^{H} \hat{\mu}_t(s_h^t, a_h^t, h)\, \zeta_h^t}_{(II)}. \tag{F.22}$$

Term (II). Since $\hat{\mu}_t$ is $\mathcal{H}_{t-1}$-measurable, (s_h^t, a_h^t, h) is independent of ζ_h^t which is itself independent of $\mathcal{H}_{t-1}$, and centered. Therefore, the partial sums form a martingale with subgaussian increments. By Azuma–Hoeffding's inequality, the event

$$\left\{(II) \leqslant 2\sqrt{2\log(T^2)\sum_{t=1}^{T}\sum_{h=1}^{H} \hat{\mu}_t(s_h^t, a_h^t, h)^2}\right\}$$

holds with probability at least $1 - 1/T^2$.

Term (I). By Cauchy–Schwarz,

$$\begin{aligned}(I) &\leqslant 2\sum_{h=1}^{H}\sqrt{\sum_{t=1}^{T}\frac{\hat{\mu}_t(s_h^t,a_h^t,h)^2}{\max\left(1,n_t(s_h^t,a_h^t,h)\right)}}\cdot\sqrt{\sum_{t=1}^{T}\max\left(1,n_t(s_h^t,a_h^t,h)\right)\left(\hat{\mu}_t(s_h^t,a_h^t,h)-\mu(s_h^t,a_h^t,h)\right)^2}\\ &\leqslant 2\sum_{h=1}^{H}\sqrt{\sum_{t=1}^{T}\frac{\hat{\mu}_t(s_h^t,a_h^t,h)^2}{\max\left(1,n_t(s_h^t,a_h^t,h)\right)}}\cdot\sqrt{T\,\beta^{\mu}(T,1/T^2)},\end{aligned}$$

where the second inequality uses the self-normalized concentration event

$$\mathcal{E}_F^{\mu}(T)\triangleq\left\{\forall t\leqslant T,\ \sum_{s,a,h}n_t(s,a,h)\,\mathrm{KL}(\hat{R}_h^t(s,a)\parallel R_h(s,a))\leqslant\beta^{\mu}(T,1/T^2)\right\},$$

which holds with probability $1-1/T^2$. Since rewards are bounded in $(0,1)$, the elliptic potential yields

$$\sum_{t=1}^{T}\frac{\hat{\mu}_t(s_h^t,a_h^t,h)^2}{\max\left(1,n_t(s_h^t,a_h^t,h)\right)}\leqslant 4\log(T+1),$$

giving $(I)\leqslant 4H\sqrt{2T\,\beta^{\mu}(T,1/T^2)\log(T+1)}$.

Combining the bounds on (I) and (II), and using that $\sum_{t=1}^{T}\sum_h\hat{\mu}_t(s_h^t,a_h^t,h)^2\leqslant HT$ together with $\beta^{\mu}(T,1/T^2)=O(\log T)$, we obtain $\mathfrak{J}_T\leqslant O\big(\sqrt{T\log T}\big)=o(T)$ on an event holding with probability at least $1-1/T^2$. □

Lemma 23. *There exists a deterministic sequence $\phi(T)=o(T)$ such that the event*

$$\mathcal{E}_F^{\phi}(T)\triangleq\left\{\sum_{t=1}^{T}\sum_{h=1}^{H}\eta_t\Big[2\,n_{t+1}(s_h^t,a_h^t,h)\big|\hat{\mu}_t(s_h^t,a_h^t,h)-\hat{\mu}_{t+1}(s_h^t,a_h^t,h)\big|\Big]^2\leqslant\phi(T)\right\}$$

holds with probability at least $1-O(1/T^2)$, where $\phi(T)$ is given explicitly in the proof.

Proof. Using the identity established in the proof of Lemma 22,

$$n_{t+1}(s_h^t,a_h^t,h)\big(\hat{\mu}_t(s_h^t,a_h^t,h)-\hat{\mu}_{t+1}(s_h^t,a_h^t,h)\big)=\hat{\mu}_t(s_h^t,a_h^t,h)-\mu(s_h^t,a_h^t,h)+\zeta_h^t,$$

and the inequality $(x+y)^2\leqslant 2x^2+2y^2$, we obtain

$$\Big[2\,n_{t+1}(s_h^t,a_h^t,h)\big|\hat{\mu}_t(s_h^t,a_h^t,h)-\hat{\mu}_{t+1}(s_h^t,a_h^t,h)\big|\Big]^2\leqslant 8\big(\hat{\mu}_t(s_h^t,a_h^t,h)-\mu(s_h^t,a_h^t,h)\big)^2+8\big(\zeta_h^t\big)^2.\tag{F.23}$$

Using the self-normalized concentration event

$$\mathcal{E}_F^{\mu}(T)\triangleq\left\{\forall t\leqslant T,\ \sum_{s,a,h}n_t(s,a,h)\,\mathrm{KL}(\hat{R}_h^t(s,a)\parallel R_h(s,a))\leqslant\beta^{\mu}(T,1/T^2)\right\},$$

it follows

$$\begin{aligned}\sum_{t=1}^{T}\big(\hat{\mu}_t(s_h^t,a_h^t,h)-\mu(s_h^t,a_h^t,h)\big)^2&=\sum_{t=1}^{T}\frac{\max\left(1,n_t(s_h^t,a_h^t,h)\right)\left(\hat{\mu}_t(s_h^t,a_h^t,h)-\mu(s_h^t,a_h^t,h)\right)^2}{\max\{1,n_t(s_h^t,a_h^t,h)\}}\\ &\leqslant\sum_{t=1}^{T}\frac{2\beta^{\mu}(T,1/T^2)}{\max\left(1,n_t(s_h^t,a_h^t,h)\right)},\end{aligned}$$

which thus holds with probability at least $1 - 1/T^2$.

For the χ^2 terms, by a union bound over $SAHT$ events, the following

$$\mathcal{E}_F^{\zeta}(T) \triangleq \left\{\forall\, t \leqslant T,\ \forall\,(s,a,h),\ |\zeta_h^t| \leqslant \sqrt{2\log(2SAHT^3)}\right\}$$

holds with probability at least $1 - 1/T^2$.

On $\mathcal{E}_F^{\mu}(T) \cap \mathcal{E}_F^{\zeta}(T)$, which holds with probability at least $1-2/T^2$, substituting both bounds into (F.23), summing over t and h, and applying the elliptic potential bound $\sum_{t=1}^{T} \frac{1}{\max\left(1, n_t(s_h^t, a_h^t, h)\right)} \leqslant 4\log(T+1)$ gives

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \eta_t \left[2\, n_{t+1}(s_h^t, a_h^t, h)\left|\hat{\mu}_t - \hat{\mu}_{t+1}\right|((s_h^t, a_h^t, h))\right]^2$$
$$\leqslant 16H\log(2SAHT^3)\sum_{t=1}^{T}\eta_t + 64H\,\beta^{\mu}(T, 1/T^2)\log(T+1),$$

which is $o(T)$ since $\sum_{t=1}^{T}\eta_t = o(T)$ and $\beta^{\mu}(T, 1/T^2) = O(\log T)$. □

Lemma 24. *There exists a deterministic sequence $d(T) = o(T)$ such that the event*

$$\mathcal{E}_F^{D}(T) \triangleq \left\{\sum_{t=K}^{T} -\mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}\left[\kappa_{t+1}(\theta)\right] \leqslant d(T)\right\}$$

holds with probability at least $1 - 1/T^2$, where $d(T)$ is given explicitly in the proof.

Proof. Recalling the definition of $\kappa_{t+1}(\theta)$ from the proof of Lemma 21 and introducing

$$u_t(s,a,h) \triangleq \mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}\left[\theta(s,a,h)\right],$$

we have

$$\begin{aligned} D_T(T_0) &= \sum_{t=T_0}^{T-1}\sum_{h=1}^{H} 2\, n_{t+1}(s_h^t, a_h^t, h)\, u_t(s_h^t, a_h^t, h)\left(\hat{\mu}_{t+1}(s_h^t, a_h^t, h) - \hat{\mu}_t(s_h^t, a_h^t, h)\right) \\ &= \sum_{t=T_0}^{T-1}\sum_{h=1}^{H} 2\left[\mu(s_h^t, a_h^t, h) - \hat{\mu}_t(s_h^t, a_h^t, h) - \zeta_h^t\right] u_t(s_h^t, a_h^t, h), \end{aligned}$$

where the second equality uses the identity from Lemma 22. We bound the two resulting terms separately.

Since $u_t(s,a,h) \leqslant 1$, Cauchy–Schwarz and the elliptic potential (Lemma 46) give, on

$$\mathcal{E}_F^{\mu}(T) \triangleq \left\{\forall t \leqslant T,\ \sum_{s,a,h} n_t(s,a,h)\,\mathrm{KL}(\hat{R}_h^t(s,a) \parallel R_h(s,a)) \leqslant \beta^{\mu}(T, 1/T^2)\right\},$$

the following

$$\begin{aligned} \sum_{t=T_0}^{T-1}\sum_{h=1}^{H}\left(\mu - \hat{\mu}_t\right)(s_h^t, a_h^t, h)\cdot u_t(s_h^t, a_h^t, h) &\leqslant \sqrt{\sum_{t,h}\frac{u_t(s_h^t, a_h^t, h)^2}{\max\left(1, n_t(s_h^t, a_h^t, h)\right)}}\cdot \\ &\quad \sqrt{\sum_{t,h}\max\left(1, n_t(s_h^t, a_h^t, h)\right)\left(\mu - \hat{\mu}_t\right)^2(s_h^t, a_h^t, h)} \\ &\leqslant \sqrt{4H\log(T+1)}\cdot\sqrt{2T\,\beta^{\mu}(T, 1/T^2)}. \end{aligned}$$

The sequence $(-\zeta_h^t\, u_t(s_h^t, a_h^t, h))_{t,h}$ is a martingale difference sequence. By Azuma–Hoeffding's inequality, the event

$$\left\{ \sum_{t=T_0}^{T-1} \sum_{h=1}^{H} -\zeta_h^{t+1}\, u_t(s_h^t, a_h^t, h) \leqslant \sqrt{2\log(T^2) \sum_{t,h} u_t(s_h^t, a_h^t, h)^2} \right\}$$

holds with probability at least $1 - 1/T^2$.

Combining both bounds and using $u_t \leqslant 1$ gives, on an event of probability at least $1 - 2/T^2$,

$$D_T(T_0) \leqslant 2\sqrt{8HT\, \beta^\mu(T, 1/T^2) \log(T+1)} + 2\sqrt{2HT \log(T^2)} = o(T),$$

which defines the event $\mathcal{E}_F^D(T)$. □

F.6 Proof of Theorem 15

We can now prove Theorem 15 with the previous intermediate results.

Let us define

$$\mathcal{E}_F(T) = \mathcal{E}_E(T) \cap \mathcal{E}_F^\mu(T) \cap \mathcal{E}_F^\phi(T) \cap \mathcal{E}_F^D(T) \cap \mathcal{E}_F^\vartheta(T) \cap \mathcal{E}_F^\Psi(T), \tag{F.24}$$

where $\mathcal{E}_E(T)$ is defined in Appendix E, and the other events are defined above. By Lemma 14, there exists $k_0 < \infty$ such that for $T \geqslant k_0$, when $\mathcal{E}_E(T)$ holds,

$$\hat{\pi}_t = \pi^\star, \quad \forall t \in (T^2/3, T).$$

Setting $T_0 \triangleq T^{2/3}$ in Lemma 18, $\mathcal{E}_E(T) \Longrightarrow \mathcal{E}_F^{T^{2/3}}(T)$, so that

$$\begin{aligned} \sum_{t=T^{2/3}}^{T-1} \mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}\left[\ell_t^\alpha(\theta,\Phi)\right] \;\leqslant\; & -\frac{1}{\eta_T}\log\Lambda_T^{\eta_T} + \Psi_T^\alpha(T^{2/3}) + \frac{1}{\eta_{T^{2/3}}}\log\Lambda_{T^{2/3}}^{\eta_{T^{2/3}}} + \\ & \sum_{t=T^{2/3}}^{T-1}\left[\frac{1}{\eta_t}\log\Lambda_{t+1}^{\eta_t} - \frac{1}{\eta_{t+1}}\log\Lambda_{t+1}^{\eta_{t+1}}\right]. \end{aligned}$$

Thanks to Lemma 21, Lemma 20, and Lemma 19, and noting that $\mathcal{E}_F(T)$ holds with probability at least $1 - O(1/T^2)$ concludes the proof.

G SADDLE-POINT CONVERGENCE

In this section, we prove the saddle-point convergence result that characterizes the optimality of the sampling rule in terms of posterior contraction. We work on the intersection of the following concentration events:

$$\begin{aligned} \mathcal{E}_G^p(T) &\triangleq \left\{ \forall\, t \leqslant T :\; \sum_{s,a,h} n_t(s,a,h)\, \mathrm{KL}(\hat{p}_t(\cdot \mid s,a,h) \,\|\, p(\cdot \mid s,a,h)) \leqslant \beta^p(T, 1/T^2) \right\}, \\ \mathcal{E}_G^\mu(T) &\triangleq \left\{ \forall\, t \leqslant T :\; \sum_{s,a,h} n_t(s,a,h)\, \mathrm{KL}\Big(\hat{R}_h^t(s,a) \,\|\, R_h(s,a)\Big) \leqslant \beta^\mu(T, 1/T^2) \right\}, \\ \mathcal{E}_G^{\mu+p}(T) &\triangleq \left\{ \forall\, t \leqslant T :\; \sum_{s,a,h} n_t(s,a,h)\, \mathrm{KL}\Big(\hat{R}_h^t(s,a) \otimes \hat{p}_h^t(\cdot \mid s,a) \,\|\, R_h(s,a) \otimes p(\cdot \mid s,a,h)\Big) \leqslant \beta^{\mu+p}(T, 1/T^2) \right\}. \end{aligned} \tag{G.1}$$

Explicit thresholds $\beta^p, \beta^\mu, \beta^{\mu+p}$ ensuring high-probability coverage follow from self-normalized concentration (see e.g. Kaufmann & Koolen 2021). Each event holds with probability at least $1 - 1/T^2$, so their intersection holds with probability at least $1 - 3/T^2$.

We recall the log-likelihood ratio between a model (θ, Φ) and the empirical MDP $\hat{M}_t$:

$$\Upsilon_t(\theta, \Phi) \triangleq \sum_{s,a,h} n_t(s,a,h) \left[\frac{\big(\theta(s,a,h) - \hat{\mu}_t(s,a,h)\big)^2}{2} + \sum_{s'} \hat{p}_t(s' \mid s,a,h) \log \frac{\hat{p}_t(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right]. \tag{G.2}$$

The GLR (Generalized Likelihood Ratio) is defined as

$$\mathrm{GLR}(T) \triangleq \inf_{(\theta,\Phi) \in \mathrm{Alt}(\hat{M}_T)} \Upsilon_T(\theta, \Phi) \tag{G.3}$$

The main result of this section is the following.

Proposition 25. *Let $\gamma \in (0,1)$, $\alpha \in (0, \frac{1}{2})$, and $\varsigma > 2\alpha$ and $\eta_t = O(t^{-\varsigma})$. There exists $k_0 < \infty$ such that for any $T > k_0$,*

$$\mathrm{GLR}(T) \geqslant T \cdot \Gamma_M - O(\log T) - O(T^{1+\alpha-\gamma}) - o(T) - O\Big(T^{\alpha + \frac{\gamma+1}{2}} \sqrt{\log T}\Big) \,,$$

holds with probability at least $1 - O(1/T^2)$.

Proof. We note that

$$\begin{aligned}
\mathrm{GLR}(T) \;&\triangleq\; \inf_{(\theta,\Phi) \in \mathrm{Alt}(\hat{M}_T)} \Upsilon_T(\theta, \Phi) \,, \\
&= \inf_{(\theta,\Phi) \in \mathrm{Alt}(\hat{M}_T)} \sum_{s,a,h} n_T(s,a,h) \left[\frac{(\theta_h(s,a) - \hat{\mu}_T(s,a,h))^2}{2} + \sum_{s'} \hat{p}_T(s' \mid s,a,h) \log \frac{1}{\Phi(s' \mid s,a,h)} \right] \\
&\quad + \sum_{s,a,h} \sum_{s'} n_T(s,a) \hat{p}_T(s' \mid s,a,h) \log \hat{p}_T(s' \mid s,a,h) \,.
\end{aligned}$$

We recall that in the equation above, for unobserved transitions, $0 \cdot \log 0 = 0$. Next, since $n_T(s,a,h)\hat{p}_T(s' \mid s,a,h) = n_T(s' \mid s,a,h)$ (and this is always true, even at initialization), we have

$$\sum_{s,a,h} \sum_{s'} n_T(s,a,h) \hat{p}_T(s' \mid s,a,h) \log \frac{1}{\Phi(s' \mid s,a,h)} = \sum_{t \in [T-1]} \sum_{h} \log \frac{1}{\Phi(s^t_{h+1} \mid s^t_h, a^t_h, h)}.$$

Similarly we have

$$\sum_{s,a,h} n_T(s,a,h) \frac{\big(\theta(s,a,h) - \hat{\mu}_T(s,a,h)\big)^2}{2} = \sum_{t \in [T-1]} \sum_{h} \frac{(\theta(s^t_h, a^t_h, h) - \hat{\mu}_T(s^t_h, a^t_h, h))^2}{2}.$$

Combining these yields

$$\begin{aligned}
\Upsilon_T(\theta, \Phi) = \sum_{t \in [T-1]} \left[\sum_{h} \frac{(\theta(s^t_h, a^t_h, h) - \hat{\mu}_T(s^t_h, a^t_h, h))^2}{2} + \log \frac{1}{\Phi(s^t_{h+1} \mid s^t_h, a^t_h, h)} \right] + \\
\sum_{t \in [T-1]} \sum_{h} \log \hat{p}_T(s^t_{h+1} \mid s^t_h, a^t_h, h) \,.
\end{aligned}$$

Moreover, the above is well-defined since $\forall t < T, p_T(s^t_{h+1} \mid s^t_h, a^t_h, h) > 0$.

Next, we invoke Theorem 15 using the regret properties of the conditional posterior sampling algorithm, we have with probability larger than $1 - O(1/T^2)$,

$$\begin{aligned}
&\inf_{(\theta,\Phi) \in \mathrm{Alt}(\hat{M}_T)} \sum_{t \in [T-1]} \sum_{h} \left[\frac{\big(\theta(s^t_h, a^t_h, h) - \hat{\mu}_T(s^t_h, a^t_h, h)\big)^2}{2} + \log \frac{1}{\Phi(s^t_{h+1} \mid s^t_h, a^t_h, h)} \right] \geqslant \\
&\sum_{t=T^{2/3}}^{T-1} \mathbb{E}_{(\theta,\Phi) \sim \nu_t^{\eta_t}(\cdot \mid \mathrm{Alt}(\hat{M}_t))} \big[\ell_t^{\alpha}(\theta, \Phi) \big] \;-\; o(T) \,,
\end{aligned} \tag{G.4}$$

which yields

$$\mathrm{GLR}(T) \geqslant \sum_{t=1}^{T-1} \mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}\left[\ell_t^\alpha(\theta,\Phi)\right] + \sum_{t\in[T-1]}\sum_h \log \hat{p}_T(s_{h+1}^t \mid s_h^t, a_h^t, h) - o(T), \tag{G.5}$$

where, since $\ell_t^\alpha \in (0, H(\frac{1}{2}+t^\alpha))$ and $\alpha < \frac{1}{2}$, the first $T^{2/3}$ terms of the sum contribute for $O(T^{\frac{2+2\alpha}{3}}) = o(T)$.

By Lemma 31, with probability at least $1 - O(1/T^2)$, we have

$$\sum_{t\in[T-1]} \ell_t^\alpha(\theta_t,\Phi_t) \leqslant \sum_{t\in[T-1]} \mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot|\mathrm{Alt}(\hat{M}_t))}\left[\ell_t^\alpha(\theta,\Phi)\right] + \sqrt{2\log(T^2)\sum_{t=1}^{T} C_{t,\alpha}^2}, \tag{G.6}$$

with $C_{t,\alpha} = H(\frac{1}{2}+t^\alpha)$. We recall that $\widetilde{M}_t \triangleq (\theta_t,\Phi_t)$ is the challenger MDP sampled at time t in the pseudocode of Algorithm 1.

Substituting (G.6) into (G.5) and recalling the definition of ℓ_t^α from (F.9), we obtain

$$\begin{aligned}
\mathrm{GLR}(T) &\geqslant \sum_{t\in[T-1]}\sum_{h=1}^{H}\left[\frac{\left(\hat{\mu}_t(s_h^t,a_h^t,h)-\theta_t(s_h^t,a_h^t,h)\right)^2}{2} + \log\frac{1}{\max\left(\Phi_t(s_{h+1}^t \mid s_h^t,a_h^t,h), e^{-t^\alpha}\right)}\right] \\
&\quad + \sum_{t\in[T-1]}\sum_{h=1}^{H}\log\hat{p}_T(s_{h+1}^t \mid s_h^t,a_h^t,h) - o(T) \\
&= \sum_{t\in[T-1]}\sum_{h=1}^{H}\left[\frac{\left(\hat{\mu}_t(s_h^t,a_h^t,h)-\theta_t(s_h^t,a_h^t,h)\right)^2}{2} + \log\frac{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t,a_h^t,h),\varepsilon\right)}{\max\left(\Phi_t(s_{h+1}^t \mid s_h^t,a_h^t,h), e^{-t^\alpha}\right)}\right] \\
&\quad + \sum_{t\in[T-1]}\sum_{h=1}^{H}\log\frac{\hat{p}_T(s_{h+1}^t \mid s_h^t,a_h^t,h)}{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t,a_h^t,h),\varepsilon\right)} - o(T),
\end{aligned} \tag{G.7}$$

where the equality rearranges the logarithms by adding and subtracting $\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t,a_h^t,h),\varepsilon\right)$.

Thanks to Lemma 29, with probability larger than $1 - O(1/T^2)$.

$$\begin{aligned}
&\sum_{t=1}^{T-1}\sum_{s,a,h} w_h^{\tilde{\pi}_t^{\exp}}(s,a)\left[\frac{\left(\hat{\mu}_t(s,a,h)-\theta_t(s,a,h)\right)^2}{2} + \sum_{s'} p(s' \mid s,a,h)\log\frac{\max\left(\hat{p}_t(s' \mid s,a,h),\varepsilon\right)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}\right] \\
&- \sum_{t=1}^{T-1}\sum_{h=1}^{H}\left[\frac{\left(\hat{\mu}_t(s_h^t,a_h^t,h)-\theta_t(s_h^t,a_h^t,h)\right)^2}{2} + \log\frac{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t,a_h^t,h),\varepsilon\right)}{\max\left(\Phi_t(s_{h+1}^t \mid s_h^t,a_h^t,h), e^{-t^\alpha}\right)}\right] \leqslant \\
&\sqrt{2\sum_{t=1}^{T-1} H^2(t^\alpha+\tfrac{1}{2})^2\log(T^2)}
\end{aligned} \tag{G.8}$$

Combining (G.7) with (G.8) gives

$$\begin{aligned}
\mathrm{GLR}(T) &\geqslant \sum_{t=1}^{T-1}\sum_{s,a,h} w_h^{\tilde{\pi}_t^{\exp}}(s,a)\left[\frac{\left(\hat{\mu}_t(s,a,h)-\theta_t(s,a,h)\right)^2}{2} + \sum_{s'} p(s' \mid s,a,h)\log\frac{\max\left(\hat{p}_t(s' \mid s,a,h),\varepsilon\right)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}\right] \\
&\quad + \sum_{t=1}^{T-1}\sum_{h=1}^{H}\log\frac{\hat{p}_T(s_{h+1}^t \mid s_h^t,a_h^t,h)}{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t,a_h^t,h),\varepsilon\right)} - o(T).
\end{aligned} \tag{G.9}$$

Next, we relate the true transition $p(s' \mid s,a,h)$ to the empirical $\hat{p}_t(s' \mid s,a,h)$ in the logarithm above. We then invoke Lemma 26, to show that with probability at least $1 - O(1/T^2)$,

$$\begin{aligned}\sum_{t=1}^{T-1}\sum_{s,a,h} w_h^{\tilde{\pi}_t^{\exp}}(s,a)\sum_{s'} p(s' \mid s,a,h)\log\frac{\max\left(\hat{p}_t(s' \mid s,a,h),\varepsilon\right)}{\max\left(\Phi_t(s' \mid s,a,h),e^{-t^\alpha}\right)} \geqslant \\ \sum_{t=1}^{T-1}\sum_{s,a,h} w_h^{\tilde{\pi}_t^{\exp}}(s,a)\sum_{s'} \hat{p}_t(s' \mid s,a,h)\log\frac{\max\left(\hat{p}_t(s' \mid s,a,h),\varepsilon\right)}{\max\left(\Phi_t(s' \mid s,a,h),e^{-t^\alpha}\right)} - O\Big(T^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log T}\Big) - o(T)\end{aligned} \tag{G.10}$$

Substituting (G.10) into (G.9) and noting that

$$\hat{p}_t(s' \mid s,a,h)\log\max\left(\hat{p}_t(s' \mid s,a,h),\varepsilon\right) \geqslant \hat{p}_t(s' \mid s,a,h)\log\hat{p}_t(s' \mid s,a,h)$$

gives

$$\begin{aligned}\mathrm{GLR}(T) \geqslant \sum_{t=1}^{T}\sum_{s,a,h} w_h^{\tilde{\rho}^t}(s,a)\left[\frac{\big(\hat{\mu}_t(s,a,h)-\theta_t(s,a,h)\big)^2}{2} + \sum_{s'}\hat{p}_t(s' \mid s,a,h)\log\frac{\max\left(\hat{p}_t(s' \mid s,a,h),\varepsilon\right)}{\max\left(\Phi_t(s' \mid s,a,h),e^{-t^\alpha}\right)}\right] \\ + \sum_{t=1}^{T}\sum_{h=1}^{H}\log\frac{\hat{p}_T(s_{h+1}^t \mid s_h^t,a_h^t,h)}{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t,a_h^t,h),\varepsilon\right)} - O\Big(T^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log T}\Big) - o(T).\end{aligned} \tag{G.11}$$

Recalling the definition of the reward kernel

$$r_t^{\exp}(s,a,h) \triangleq \frac{\big(\hat{\mu}_t(s,a,h)-\theta_t(s,a,h)\big)^2}{2} + \sum_{s'}\hat{p}_t(s,a,s',h)\log\frac{\hat{p}_t(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h,e^{-t^\alpha}\right)}, \tag{G.12}$$

and substituting into (G.11), we obtain, with probability $1 - O(1/T^2)$

$$\mathrm{GLR}(T) \geqslant \sum_{t=1}^{T}\sum_{s,a,h} w_h^{\tilde{\pi}_t^{\exp}}(s,a)\, r_t^{\exp}(s,a,h) + \sum_{t=1}^{T}\sum_{h=1}^{H}\log\frac{\hat{p}_T(s_{h+1}^t \mid s_h^t,a_h^t,h)}{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t,a_h^t,h),\varepsilon\right)} - \tag{G.13}$$

$$O\Big(T^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log T}\Big) - o(T). \tag{G.14}$$

Decomposing the mixed policy $\tilde{\pi}_t^{\exp} = (1-\varepsilon_t)\pi_t^{\exp} + \varepsilon_t c_t$,

$$\begin{aligned}\sum_{t=1}^{T-1}\sum_{s,a,h} w_h^{\tilde{\pi}_t^{\exp}}(s,a)\, r_t^{\exp}(s,a,h) = \sum_{t=1}^{T-1}\sum_{s,a,h} w_h^{\pi_t^{\exp}}(s,a)\, r_t^{\exp}(s,a,h) + \\ \sum_{t=1}^{T-1}\sum_{s,a,h}\varepsilon_t\big[w_h^{c_t}(s,a) - w_h^{\pi_t^{\exp}}(s,a)\big] r_t^{\exp}(s,a,h).\end{aligned} \tag{G.15}$$

By Lemma 27, the correction term satisfies

$$\left|\sum_{t=1}^{T-1}\sum_{s,a,h}\varepsilon_t\big[w_h^{c_t}(s,a) - w_h^{\pi_t^{\exp}}(s,a)\big] r_t^{\exp}(s,a,h)\right| \leqslant \frac{2}{1+\alpha-\gamma}\, T^{1+\alpha-\gamma}. \tag{G.16}$$

Substituting (G.15) and (G.16) into (G.14),

$$\begin{aligned}\mathrm{GLR}(T) \geqslant \sum_{t=1}^{T}\sum_{s,a,h} w_h^{\pi_t^{\exp}}(s,a)\, r_t^{\exp}(s,a,h) + \sum_{t=1}^{T-1}\sum_{h=1}^{H}\log\frac{\hat{p}_T(s_{h+1}^t \mid s_h^t,a_h^t,h)}{\hat{p}_t(s_{h+1}^t \mid s_h^t,a_h^t,h)} - \\ O\big(T^{1+\alpha-\gamma}\big) - O\Big(T^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log T}\Big) - o(T).\end{aligned} \tag{G.17}$$

By Lemma 30, with probability larger than $1 - O(1/T^2)$,

$$\left| \sum_{t=1}^{T-1} \sum_{h=1}^{H} \log \frac{\hat{p}_T(s_{h+1}^t \mid s_h^t, a_h^t, h)}{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t, a_h^t, h), \varepsilon\right)} \right| \leqslant O\Big(T^{\alpha + \frac{\gamma+1}{2}} \sqrt{\log T}/\varepsilon\Big) \tag{G.18}$$

Combining (G.17) and (G.18) gives

$$\mathrm{GLR}(T) \geqslant \sum_{t=1}^{T-1} \sum_{s,a,h} w_h^{\pi_t^{\exp}}(s,a)\, r_t^{\exp}(s,a,h) - O\big(T^{1+\alpha-\gamma}\big) - O\Big(T^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log T}\Big) - o(T). \tag{G.19}$$

Using the no-regret property of the online RL learner (Proposition 10), we have for $\alpha \in (0, \frac{1}{2})$

$$\mathrm{GLR}(T) \quad \geqslant \quad \sup_{w \in \Omega_M} \sum_{t \in [T-1]} \sum_{s,a,h} w_h(s,a) r_t^{\exp}(s,a,h) - O(T^{1+\alpha-\gamma}) - O\Big(T^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log T}\Big) - o(T).$$

Next, we invoke Lemma 28 to relate the empirical quantities in $r_t^{\exp}$ to their actual values. Indeed Lemma 28 shows that on an event that holds with probability $1 - O(1/T^2)$

$$\begin{aligned} &\sum_{t=1}^{T-1} \sum_{s,a,h} w_h(s,a)\, r_t^{\exp}(s,a,h) \geqslant \\ &\sum_{t=1}^{T-1} \sum_{s,a,h} w_h(s,a) \left[\frac{\big(\theta_t(s,a,h) - \mu(s,a,h)\big)^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)} \right] + o(T) \end{aligned} \tag{G.20}$$

Therefore,

$$\begin{aligned} \mathrm{GLR}(T) \quad \geqslant \quad & \sup_{w \in \Omega_M} \sum_{t=1}^{T-1} \sum_{s,a,h} w_h(s,a) \left[\frac{\big(\theta_t(s,a,h) - \mu(s,a,h)\big)^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)} \right] \\ & - O(T^{1+\alpha-\gamma}) - o(T) - O\Big(T^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log T}\Big) \end{aligned}$$

The final step to conclude the proof invokes Lemma 32 to prove that there exists $k_0 < \infty$ such that for $T \geqslant k_0$, we have for any $w \in \Omega_M$

$$\begin{aligned} &\sum_{s,a,h} w_h(s,a) \left[\frac{\big(\theta_t(s,a,h) - \mu(s,a,h)\big)^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)} \right] \geqslant \\ &\inf_{(\theta,\Phi) \in \mathrm{Alt}(M)} \sum_{s,a,h} w_h(s,a) \left[\frac{\big(\theta(s,a,h) - \mu(s,a,h)\big)^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right] - O(1/t). \end{aligned}$$

We then have

$$\begin{aligned} \mathrm{GLR}(T) \quad \geqslant \quad & \sup_{w \in \Omega_M} \sum_{t=1}^{T-1} \inf_{(\theta,\Phi) \in \mathrm{Alt}(M)} \sum_{s,a,h} w_h(s,a) \left[\frac{(\theta(s,a,h) - \mu(s,a,h))^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\Phi_h(s,a,s')} \right] \\ & - O(\log T) - O(T^{1+\alpha-\gamma}) - o(T) - O\Big(T^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log T}\Big) \\ = \quad & T \cdot \Gamma_M - O(\log T) - O(T^{1+\alpha-\gamma}) - o(T) - O\Big(T^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log T}\Big)\ , \end{aligned}$$

which concludes the proof.

□

G.1 Concentration Lemmas

We now prove the concentrations lemmas used in this section.

Lemma 26. *There exists $k_0 < \infty$ such that for $T \geqslant k_0$, with probability $1 - O(1/T^2)$, the following holds:*

$$\sum_{t=1}^{T} \sum_{s,a,h} w_h^{\tilde{\pi}_t^{\exp}}(s,a) \sum_{s'} p(s' \mid s,a,h) \log \frac{\hat{p}_t(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)} \geqslant$$
$$\sum_{t=1}^{T} \sum_{s,a,h} w_h^{\tilde{\pi}_t^{\exp}}(s,a) \sum_{s'} \hat{p}_t(s' \mid s,a,h) \log \frac{\hat{p}_t(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)} - O\Big(T^{\alpha + \frac{\gamma+1}{2}} \sqrt{\log T}\Big) - o(T).$$

Proof. For fixed (t,s,a,s',h), letting $L_t(s,a,s') \triangleq \log \frac{\hat{p}_t(s,a,s',h)}{\max(\Phi_t(s'|s,a,h), e^{-t^\alpha})}$,

$$\big|p_h(s,a,s') - \hat{p}_t(s,a,s',h)\big| \cdot |L_t(s,a,s')| \leqslant \|\hat{p}_t(\cdot \mid s,a,h) - p_h(\cdot \mid s,a,h)\|_1 \cdot t^\alpha,$$

since $|L_t| \leqslant t^\alpha$. Thanks to Lemma 12, there exists $k_0 < \infty$, such that for $T \geqslant k_0$ for all $t \in (T^{2/3}, T)$ and all reachable (s,a,h), if $\mathcal{E}_G^p(T)$ holds, then

$$\|\hat{p}_t(\cdot \mid s,a,h) - p(\cdot \mid s,a,h)\|_1 \leqslant O\Big(\sqrt{t^{\gamma-1}\beta^p(t^2, 1/t^3)}\Big),$$

so the per-(t,s,a,h) error is at most $O\Big(\sqrt{t^{2\alpha+\gamma-1}\beta^p(t^2,1/t^3)}\Big)$. The first $T^{2/3}$ terms contributes for $O(T^{\frac{2+2\alpha}{3}}) = o(T)$ (as $\alpha \in (0, \frac{1}{2})$). Summing over t, s, a, h and weighting by $w_h^{\tilde{\pi}_t^{\exp}}$, the error term is at most

$$O\Big(T^{\alpha + \frac{\gamma+1}{2}} \sqrt{\beta^p(T^2, 1/T^3)}\Big).$$

The conclusion follows as $\beta^p(T^2, 1/T^3) = O(\log T)$. □

Lemma 27. *We have*

$$\left|\sum_{t=1}^{T} \sum_{s,a,h} \varepsilon_t \Big[w_h^{c_t}(s,a) - w_h^{\pi_t^{\exp}}(s,a)\Big] r_t^{\exp}(s,a,h)\right| \leqslant O(T^{1-\gamma+\alpha}).$$

Proof. Since $|r_t^{\exp}(s,a,h)| \leqslant \frac{1}{2} + t^\alpha$ for all (s,a,h) and $\varepsilon_t = t^{-\gamma}$, the triangle inequality gives

$$\begin{aligned}
\left|\sum_{t=1}^{T} \sum_{s,a,h} \varepsilon_t \big[w_h^{c_t}(s,a) - w_h^{\pi_t^{\exp}}(s,a)\big] r_t^{\exp}(s,a,h)\right| &\leqslant \sum_{t=1}^{T} \varepsilon_t \left|\sum_{s,a,h} \big[w_h^{c_t}(s,a) - w_h^{\pi_t^{\exp}}(s,a)\big] r_t^{\exp}(s,a,h)\right| \\
&\leqslant 2 \sum_{t=1}^{T} t^{-\gamma} \left(t^\alpha + \frac{1}{2}\right) \\
&\leqslant O(T^{1-\gamma+\alpha}),
\end{aligned}$$

where the second inequality uses that $\sum_{s,a,h} |w_h^{c_t}(s,a) - w_h^{\pi_t^{\exp}}(s,a)| \leqslant 2$, and the last step bounds the sums by integrals. □

Lemma 28. *There exists $k_0 < \infty$ such that for $T \geqslant k_0$, for any valid state-action allocation $w \in \Omega_M$, the following ,*

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{s,a,h} w_h(s,a)\, r_t^{\exp}(s,a,h) \geqslant \\
&\sum_{t=1}^{T} \sum_{s,a,h} w_h(s,a) \left[\frac{\big(\theta_t(s,a,h) - \mu(s,a,h)\big)^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}\right] + \\
&\quad O\left(T^{\alpha + \frac{1+\gamma}{2}} \sqrt{\log T}\right)
\end{aligned} \tag{G.21}$$

holds with probability larger than $1 - O(1/T^2)$.

Proof. Since $w_h(s,a)=0$ for any unreachable (s,a,h), we may restrict attention to reachable triplets throughout. For each (s,a,s',h), define

$$C_t(s,a,s',h) \triangleq \hat{p}_t(s' \mid s,a,h)\log\frac{\hat{p}_t(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)} - p(s' \mid s,a,h)\log\frac{p(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}.$$

If $p(s' \mid s,a,h)=0$ then $C_t(s,a,s',h)\geqslant 0$ and the term only improves the bound. For $p(s' \mid s,a,h)>0$, the function $x\mapsto x\log\frac{x}{\varepsilon}$ is convex, so

$$\begin{aligned} C_t(s,a,s',h) &\geqslant \left(1+\log\frac{p_h(s,a,s')}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}\right)\left(\hat{p}_t(s' \mid s,a,h) - p(s' \mid s,a,h)\right) \\ &\geqslant -\left(1+t^\alpha+|\log p(s' \mid s,a,h)|\right)\left|\hat{p}_t(s' \mid s,a,h) - p(s' \mid s,a,h)\right|. \end{aligned}$$

Thanks to Lemma 12, there exists $k_0<\infty$, such that for $T\geqslant k_0$ for all $t\in(T^{2/3},T)$ and all reachable (s,a,h), if $\mathcal{E}_G^p(T)$ holds, then

$$\|\hat{p}_t(\cdot \mid s,a,h) - p(\cdot \mid s,a,h)\|_1 \leqslant O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right),$$

so

$$C_t(s,a,s',h) \geqslant -\left(1+t^\alpha+|\log p_h(s,a,s')|\right)O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right).$$

Applying a similar resaoning to the rewards (bounded in $(0,1)$) yield

$$\left(\theta_t(s,a,h)-\hat{\mu}_t(s,a,h)\right)^2 \geqslant \left(\theta_t(s,a,h)-\mu(s,a,h)\right)^2 + \left(\hat{\mu}_t(s,a,h)-\mu(s,a,h)\right)^2 - 2\left|\hat{\mu}_t(s,a,h)-\mu(s,a,h)\right|,$$

which, again, due to Lemma 12 yields

$$\left|\hat{\mu}_t(s,a,h)-\mu(s,a,h)\right| \leqslant O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right).$$

Summing over t, s, a, s', h and weighting by $w_h(s,a)$,

$$\begin{aligned} &\sum_{t=1}^{T}\sum_{s,a,h} w_h(s,a)\, r_t^{\exp}(s,a,h) \geqslant \\ &\sum_{t=1}^{T}\sum_{s,a,h} w_h(s,a)\left[\frac{\left(\theta_t(s,a,h)-\mu(s,a,h)\right)^2}{2} + \sum_{s'} p(s' \mid s,a,h)\log\frac{p(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}\right] - \\ &\quad\sum_{t\leqslant T} t^\alpha O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right). \end{aligned}$$

Further noting that

$$\sum_{t\leqslant T} t^\alpha O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right) \leqslant O\left(T^{\alpha+\frac{1+\gamma}{2}}\sqrt{\log T}\right)$$

concludes the proof. □

Lemma 29. *Fix $\varepsilon\in(0,1)$. The event*

$$\begin{aligned} \mathcal{E}_w(T) \triangleq \Bigg\{ &\sum_{t=1}^{T}\sum_{s,a,h} w_h^{\tilde{\pi}_t^{\exp}}(s,a)\left[\frac{\left(\hat{\mu}_t(s,a,h)-\theta_t(s,a,h)\right)^2}{2} + \sum_{s'} p(s' \mid s,a,h)\log\frac{\max\left(\hat{p}_t(s' \mid s,a,h),\varepsilon\right)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}\right] \\ &- \sum_{t=1}^{T}\sum_{h=1}^{H}\left[\frac{\left(\hat{\mu}_t(s_h^t,a_h^t,h)-\theta_t(s_h^t,a_h^t,h)\right)^2}{2} + \log\frac{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t,a_h^t,h),\varepsilon\right)}{\max\left(\Phi_t(s_{h+1}^t \mid s_h^t,a_h^t,h), e^{-t^\alpha}\right)}\right] \leqslant \qquad \text{(G.22)} \\ &\sqrt{2\sum_{t=1}^{T} H^2(t^\alpha+\tfrac{1}{2})^2\log(T^2)}\Bigg\} \qquad \text{(G.23)} \end{aligned}$$

holds with probability at least $1-1/T^2$.

Proof. Let us define, for each t and (s, a, s', h),

$$\Psi_t(s,a,s',h) \triangleq \frac{\big(\hat{\mu}_t(s,a,h) - \theta_t(s,a,h)\big)^2}{2} + \log \frac{\max\left(\hat{p}_t(s' \mid s,a,h), \varepsilon\right)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)},$$

and note that $|\Psi_t(s,a,s',h)| \leqslant t^\alpha + \frac{1}{2}$. Since $\hat{\mu}_t$ is $\mathcal{H}_{t-1}$-measurable and (θ_t, Φ_t) depends only on $\mathcal{H}_{t-1}$ plus additional independent randomization, we augment the filtration to $\widetilde{\mathcal{H}}_{t-1} \triangleq \mathcal{H}_{t-1} \cup \{\theta_t, \Phi_t\}$. Conditioned on $\widetilde{\mathcal{H}}_{t-1}$, the only source of randomness is the trajectory at episode t, generated under $\tilde{\pi}_t^{\exp}$, so

$$\mathbb{E}\left[\sum_{h=1}^{H} \Psi_t(s_h^t, a_h^t, s_{h+1}^t, h) \,\middle|\, \widetilde{\mathcal{H}}_{t-1}\right] = \sum_{h=1}^{H} \sum_{s,a,s'} w_h^{\tilde{\pi}_t^{\exp}}(s,a)\, p(s' \mid s,a,h)\, \Psi_t(s,a,s',h).$$

Therefore the process

$$m_t \triangleq \sum_{k=1}^{t} \left[\sum_{h=1}^{H} \Psi_k(s_h^k, a_h^k, s_{h+1}^k, h) - \sum_{h=1}^{H} \sum_{s,a,s'} w_h^{\tilde{\pi}_k^{\exp}}(s,a)\, p(s' \mid s,a,h)\, \Psi_k(s,a,s',h)\right]$$

is a martingale with, increments bounded in absolute value by $H(t^\alpha + \frac{1}{2})$. By Azuma–Hoeffding's maximal inequality,

$$\Pr\left(m_{T-1} > \sqrt{2\sum_{t=1}^{T} H^2 (t^\alpha + \tfrac{1}{2})^2 \log(T^2)}\right) \leqslant \frac{1}{T^2},$$

which gives the claimed event $\mathcal{E}_w(T)$. □

Lemma 30. *Let $\varepsilon > 0$. With probability at least $1 - O(1/T^2)$,*

$$\left|\sum_{t=1}^{T-1}\sum_{h=1}^{H} \log \frac{\hat{p}_T(s_{h+1}^t \mid s_h^t, a_h^t, h)}{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t, a_h^t, h), \varepsilon\right)}\right| \leqslant O\left(T^{\alpha + \frac{1+\gamma}{2}} \sqrt{\log T}/\varepsilon\right).$$

Proof. Since $\hat{p}_T(s_{h+1}^t \mid s_h^t, a_h^t, h) \geqslant 1/T > 0$ (the transition is observed at episode $t \leqslant T$) and $\max(\hat{p}_t, \varepsilon) \geqslant \varepsilon > 0$, every logarithm is finite. By the triangle inequality,

$$\left|\sum_{t=1}^{T-1}\sum_{h=1}^{H} \log \frac{\hat{p}_T(s_{h+1}^t \mid s_h^t, a_h^t, h)}{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t, a_h^t, h), \varepsilon\right)}\right| \leqslant \sum_{t=1}^{T-1}\sum_{h=1}^{H} \left|\log \frac{\hat{p}_T(s_{h+1}^t \mid s_h^t, a_h^t, h)}{\max\left(\hat{p}_t(s_{h+1}^t \mid s_h^t, a_h^t, h), \varepsilon\right)}\right|.$$

For $T_0 = o(T)$, the contribution from $t < T_0$ is $o(T)$ since $\max(\hat{p}_t, \varepsilon) \geqslant \varepsilon$, giving $|\log(\hat{p}_T/\varepsilon)| \leqslant \log(1/\varepsilon)$ per term, times T_0.

Thanks to Lemma 12, there exists $k_0 < \infty$, such that for $T \geqslant k_0$ for all $t \in (T^{2/3}, T)$ and all (s,a,h), if $\mathcal{E}_G^p(T)$ holds, then

$$\|\hat{p}_t(\cdot \mid s,a,h) - p(\cdot \mid s,a,h)\|_1 \leqslant O\left(\sqrt{t^{\gamma-1}\beta^p(t^2, 1/t^3)}\right).$$

We assume $t \geqslant T^{2/3}$. We split into two cases.

In the first case, assume $\hat{p}_t(s_{h+1}^t \mid s_h^t, a_h^t, h) \geqslant \varepsilon$. Then $\max(\hat{p}_t, \varepsilon) = \hat{p}_t$, and since $\gamma < \frac{1}{2}$ both $\hat{p}_T$ and $\hat{p}_t$ lie in $[(1 - t^{\gamma-1/2})p_h,\ (1 + t^{\gamma-1/2})p_h]$, so

$$\left|\log \frac{\hat{p}_T}{\hat{p}_t}\right| \leqslant \log \frac{1 + O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right)}{1 - O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right)} \leqslant O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right)$$

In the second case, assuming $\hat{p}_t(s_{h+1}^t \mid s_h^t, a_h^t, h) < \varepsilon$, then $\max(\hat{p}_t, \varepsilon) = \varepsilon$. Since $\hat{p}_t < \varepsilon$, the concentration bound gives

$$p(s_{h+1}^t \mid s_h^t, a_h^t, h) \leqslant \hat{p}_t(s_{h+1}^t \mid s_h^t, a_h^t, h) + O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right) < \varepsilon + O\left(\sqrt{t^{\gamma-1}\beta^p(t^2,1/t^3)}\right).$$

Applying the concentration bound again to $\hat{p}_T$, yields

$$\hat{p}_T(s_{h+1}^t \mid s_h^t, a_h^t, h) \leqslant p(s_{h+1}^t \mid s_h^t, a_h^t, h) + O\Big(\sqrt{T^{\gamma-1}\beta^p(T^2, 1/T^3)}\Big) \leqslant \varepsilon + O\Big(\sqrt{t^{\gamma-1}\beta^p(t^2, 1/t^3)}\Big).$$

Therefore, using $\log(1+x) \leqslant x$,

$$\log \frac{\hat{p}_T(s_{h+1}^t \mid s_h^t, a_h^t, h)}{\varepsilon} \leqslant$$
$$\log \frac{\varepsilon + O\Big(\sqrt{t^{\gamma-1}\beta^p(t^2, 1/t^3)}\Big)}{\varepsilon} = \log\left(1 + \frac{O\Big(\sqrt{t^{\gamma-1}\beta^p(t^2, 1/t^3)}\Big)}{\varepsilon}\right) \leqslant$$
$$\frac{O\Big(\sqrt{t^{\gamma-1}\beta^p(t^2, 1/t^3)}\Big)}{\varepsilon} = O\Big(\sqrt{t^{\gamma-1}\beta^p(t^2, 1/t^3)}/\varepsilon\Big),$$

where the last step uses that ε is a fixed positive constant.

In both cases the contribution per (t, h) is $O\Big(\sqrt{t^{\gamma-1}\beta^p(t^2, 1/t^3)}\Big)$. Summing gives

$$\sum_{t=T_0}^{T-1}\sum_{h=1}^{H}\left|\log\frac{\hat{p}_T}{\max(\hat{p}_t, \varepsilon)}\right| \leqslant O\left(T^{\alpha+\frac{1+\gamma}{2}}\sqrt{\log T}/\varepsilon\right)$$

□

Lemma 31. *The event*

$$\mathcal{E}_G^{\mathrm{conv}}(T) \triangleq \left\{\sum_{t\in[T-1]}\Big[\ell_t^\alpha(\theta_t, \Phi_t) - \mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot\mid \mathrm{Alt}(\hat{M}_t))}\big[\ell_t^\alpha(\theta, \Phi)\big]\Big] \leqslant \sqrt{2\log(T^2)\sum_{t=1}^{T} C_{t,\alpha}^2}\right\} \tag{G.24}$$

holds with probability at least $1 - 1/T^2$, *where* $C_{t,\alpha} \triangleq H\big(\frac{1}{2} + t^\alpha\big)$. *In particular, the right-hand side is* $O\Big(\sqrt{T^{1+2\alpha}\log T}\Big)$, *which is* $o(T)$ *whenever* $\alpha < \frac{1}{2}$.

Proof. Recalling the definition of ℓ_t^α from (F.9), its increments satisfy $\ell_t^\alpha(\theta, \Phi) \in (0, C_{t,\alpha})$ for all (θ, Φ). Conditioned on $\mathcal{H}_{t-1}$, the only source of randomness in $\ell_t^\alpha(\theta_t, \Phi_t)$ is the sample $(\theta_t, \Phi_t) \sim \nu_t^{\eta_t}(\cdot \mid \mathrm{Alt}(\hat{M}_t))$, so

$$\mathbb{E}[\ell_t^\alpha(\theta_t, \Phi_t) \mid \mathcal{H}_{t-1}] = \mathbb{E}_{(\theta,\Phi)\sim\nu_t^{\eta_t}(\cdot\mid \mathrm{Alt}(\hat{M}_t))}\big[\ell_t^\alpha(\theta, \Phi)\big].$$

Therefore the process

$$m_t \triangleq \sum_{k=1}^{t}\Big[\ell_k^\alpha(\theta_k, \Phi_k) - \mathbb{E}_{(\theta,\Phi)\sim\nu_k^{\eta_k}(\cdot\mid \mathrm{Alt}(\hat{M}_k))}\big[\ell_k^\alpha(\theta, \Phi)\big]\Big]$$

is a martingale with increments bounded in $(-C_{t,\alpha}, C_{t,\alpha})$. By Azuma–Hoeffding's maximal inequality,

$$\Pr\left(m_{T-1} > \sqrt{2\log(T^2)\sum_{t=1}^{T} C_{t,\alpha}^2}\right) \leqslant \frac{1}{T^2},$$

which gives the claimed event. □

The result below proves that the clipping of the transition probabilities on the sampled instance has a negligible impact on the loss of the regret minimizer.

Lemma 32. *There exists a constant $k_0 < \infty$ such that for all $t \geqslant k_0$*

$$\sum_{s,a,h} w_h(s,a)\left[\frac{\big(\theta_t(s,a,h)-\mu(s,a,h)\big)^2}{2}+\sum_{s'} p(s' \mid s,a,h)\log\frac{p(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}\right] \geqslant$$
$$\inf_{\widetilde{M}\triangleq(\theta,\Phi)\in \mathrm{Alt}(M)} \sum_{s,a,h} w_h(s,a)\left[\frac{(\theta(s,a,h)-\mu(s,a,h))^2}{2}+\sum_{s'} p(s' \mid s,a,h)\log\frac{p(s' \mid s,a,h)}{\Phi_h(s' \mid s,a,h)}\right] - O(1/t),$$

holds with probability 1.

Proof. To ease notation, we introduce

$$d_w^\alpha((\theta_t,\Phi_t),(\mu,p)) \triangleq \sum_{s,a,h} w_h(s,a)\left[\frac{(\theta_t(s,a,h)-\mu(s,a,h))^2}{2}+\sum_{s'} p(s' \mid s,a,h)\log\frac{p(s' \mid s,a,h)}{\max\left(\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}\right].$$

Since $\widetilde{M}_t \triangleq (\theta_t,\Phi_t) \in \mathrm{Alt}(M)$ there exists a deterministic policy $\tilde{\pi} \neq \pi^\star$ such that $\tilde{\pi}$ is a global optimal policy in $\widetilde{M}_t$.

We invoke Lemma 43, which provides an MDP $M' \triangleq (\tilde{p},\tilde{\mu})$ for which $\tilde{\pi}$ is the only global optimal policy and there is a unique optimal action per stage and state. Assuming $\frac{1}{t} \leqslant \frac{1}{3}\frac{1}{1+2^{H-1}}$ i.e., $t \geqslant 3(1+2^{H-1})$ then we have by the construction in Lemma 43 with $\varepsilon = 1/t$,

$$\left\{\begin{array}{ll} \tilde{p}_h(\cdot \mid s,a) & \triangleq (1-\gamma_{s,a,h})\Phi_t(\cdot \mid s,a,h)+\gamma_{s,a,h}\Phi_t(\cdot \mid s,\pi_h(s),h) \\ |\tilde{\mu}(s,a,h)-\theta_t(s,a,h)| & \leqslant d_{s,a,h} \triangleq 2^{H-h+1}/t \end{array}\right\}$$

where $\gamma_{s,a,h} \in \left[0, \frac{(1+2^{H-h})\varepsilon}{1-(2^{H-h+1}+3)\varepsilon}\right]$. Therefore, we have

$$\begin{aligned} d_w^\alpha((\tilde{\mu},\tilde{p}),(\mu,p)) &\leqslant \sum_{s,a,h} w_h(s,a)\Bigg[\frac{(\tilde{\mu}(s,a,h)-\mu(s,a,h))^2}{2}+ \\ &\qquad \sum_{s'} p(s' \mid s,a,h)\log\frac{p(s' \mid s,a,h)}{\max\left((1-\gamma_h)\Phi_t(s' \mid s,a,h), e^{-t^\alpha}\right)}\Bigg] \\ &\leqslant d_w^\alpha((\theta_t,\Phi_t),(\mu,p)) + \sum_{s,a,h} w_h(s,a)\left[\frac{4^{H-h+1}}{2t^2}+\frac{2^{H-h+1}}{t}\log\frac{1}{1-\gamma_h}\right] \\ &\leqslant d_w^\alpha((\theta_t,\Phi_t),(\mu,p)) + \frac{4^H}{2t^2}+\frac{2^H}{t}+\frac{\gamma_1}{1-\gamma_1} \quad \text{(by monotonicity of } h \mapsto \gamma_h \text{ and } \log(1+x) \leqslant x) \\ &\leqslant d_w^\alpha((\theta_t,\Phi_t),(\mu,p)) + O(1/t). \end{aligned}$$

Now we invoke Lemma 45 to build an instance $M'_\gamma \triangleq (\tilde{\mu},$
$tilde p^\gamma)$ close to $(\tilde{\mu},\tilde{p})$, but for which $\tilde{\pi}$ is still the unique optimal policy and such that each transition probability is larger than some γ. We invoke Lemma 45 with $\gamma = 1/t$. This instance is obtained by a mixture of $(\tilde{\mu},\tilde{p})$ with $(\tilde{\mu},u)$ with weights $1-\gamma,\gamma$, where u is the uniform transition kernel.

We then have thanks to $\log(1+x) \leqslant x$,

$$\begin{aligned} d_w^\alpha((\tilde{\mu},\tilde{p}^\gamma),(\mu,p)) &\leqslant \sum_{s,a,h} w_h(s,a)\left[\frac{(\tilde{\mu}(s,a,h)-\mu(s,a,h))^2}{2}+\sum_{s'} p(s' \mid s,a,h)\log\frac{p(s' \mid s,a,h)}{\max\left((1-\gamma)\tilde{p}(s' \mid s,a,h), e^{-t^\alpha}\right)}\right] \\ &\leqslant d_w^\alpha((\tilde{\mu},\tilde{p}),(\mu,p)) + \frac{\gamma}{1-\gamma} \\ &= d_w^\alpha((\tilde{\mu},\tilde{p}),(\mu,p)) + \frac{1/t}{1-1/t} \\ &= d_w^\alpha((\tilde{\mu},\tilde{p}),(\mu,p)) + O(1/t)\,. \end{aligned}$$

Thus, combining the two displays above, we have

$$d_w^\alpha((\theta_t, \Phi_t), (\mu, p)) \geqslant d_w^\alpha((\tilde{\mu}, \tilde{p}^\gamma), (\mu, p)) - O(1/t).$$

However, the transition probabilities in $\tilde{p}^\gamma$ are larger than γ/S since $\tilde{p}^\gamma$ is the mixture with the uniform transition kernel with weight γ. Thus, assuming t is such that $1/(tS) > e^{-t^\alpha}$, we have for all (s, a, h, s')

$$\tilde{p}^\gamma(s' \mid s, a, h) \geqslant \frac{1}{tS} \geqslant e^{-t^\alpha}$$

Therefore,

$$d_w^\alpha((\tilde{\mu}, \tilde{p}^\gamma), (\mu, p)) \quad = \quad \sum_{s,a,h} w_h(s, a) \left[\frac{(\theta_h(s, a) - \mu(s, a, h))^2}{2} + \sum_{s'} p(s' \mid s, a, h) \log \frac{p(s' \mid s, a, h)}{\tilde{p}^\gamma(s' \mid s, a, h)} \right].$$

Thus, since Lemma 45 ensures that $\tilde{\pi} \neq \pi^\star$ is the unique optimal policy for each stage and state, we have $(\tilde{\mu}, \tilde{p}^\gamma) \in \mathrm{Alt}(M)$. Putting the above results together, we have

$$\begin{aligned} d_w^\alpha((\theta_t, \Phi_t), (\mu, p)) &\geqslant \sum_{s,a,h} w_h(s, a) \left[\frac{(\theta(s, a, h) - \mu(s, a, h))^2}{2} + \sum_{s'} p(s' \mid s, a, h) \log \frac{p(s' \mid s, a, h)}{\tilde{p}^\gamma(s' \mid s, a, h)} \right] - O(1/t) \\ &\geqslant \inf_{(\theta,\Phi) \in \mathrm{Alt}(M)} \sum_{s,a,h} w_h(s, a) \left[\frac{(\theta(s, a, h) - \mu(s, a, h))^2}{2} + \sum_{s'} p(s' \mid s, a, h) \log \frac{p(s' \mid s, a, h)}{\Phi(s' \mid s, a, h)} \right] - \\ &\quad O(1/t). \end{aligned}$$

□

H POSTERIOR CONVERGENCE

In this section, we prove the results related to the posterior contraction rate. First, we prove the lower bound on the posterior anti-concentration rate. More precisely, we prove the result below.

Theorem 2. *Fix an MDP $M \in \mathbf{M}$ with optimal policy $\pi^\star$ and let ν_t denote the posterior distribution defined in Eq. (1). For any adaptive algorithm for best policy identification, with probability one,*

$$\limsup_{t \to \infty} -\tfrac{1}{t} \log \mathrm{Pr}_{\widetilde{M} \sim \nu_t \,|\, \mathcal{H}_{t-1}}\big(\pi^\star \notin \Pi^\star(\widetilde{M})\big) \;\leqslant\; \Gamma_M \,.$$

Let $(\pi_t^{\mathrm{exp}})_{t \geqslant 1}$ be the sequence of exploration policies used by an adaptive algorithm, i.e., at episode t, the learner selects a policy π_t^{exp} depending on past observations summarized in $\mathcal{H}_{t-1}$.

We that recall

$$\Upsilon_t(\theta, \Phi) \triangleq \sum_{s,a,h} n_t(s, a, h) \left[\frac{\big(\hat{\mu}_t(s, a, h) - \theta(s, a, h)\big)^2}{2} + \sum_{s'} \hat{p}_t(s' \mid s, a, h) \log \frac{\hat{p}_t(s' \mid s, a, h)}{\Phi(s' \mid s, a, h)} \right], \tag{H.1}$$

is the log-likelihood ratio between the empirical MDP $\hat{M}_t \triangleq (\hat{\mu}_t, \hat{p}_t)$ and the MDP $\widetilde{M} \triangleq (\theta, \Phi) \in \mathbf{M}$.

H.1 Asymptotic Limit of Likelihood Ratio

Before proving the result above, we show the following intermediate proposition.

Proposition 33. *For any adaptive algorithm for best policy identification, with probability one,*

$$\limsup_{t \to \infty} \; \frac{1}{t} \left[\inf_{(\theta,\Phi) \in \mathrm{Alt}(M)} \Upsilon_t(\theta, \Phi) \right] \leqslant \Gamma_M.$$

Proof. First, let us justify that the empirical frequency vector $(\frac{n_t(s,a,h)}{t-1})_{s,a,h}$ is an almost valid state–action allocation for M.

For this, let us define the t-indexed stochastic process

$$\begin{aligned} m_t(s,a,h) &\triangleq \sum_{k=1}^{t} \left[\mathbf{1}\big(s_h^k = s, a_h^k = a\big) - w_h^{\pi_k^{\mathrm{exp}}}(s,a) \right], \\ &= n_{t+1}(s,a,h) - \sum_{k \leqslant t} w_h^{\pi_k^{\mathrm{exp}}}(s,a)\,, \end{aligned}$$

where $w^{\pi_t^{\mathrm{exp}}}$ is the state–action visitation measure induced by π_t^{exp} on M. Note that $(m_t(s,a,h))_t$ is a martingale and its increments are bounded in $(-1,1)$. Thanks to Azuma-Hoeffding, the event

$$\mathcal{E}_H^{s,a,h}(T) \triangleq \left\{ \forall\, t \leqslant T, \left| n_{t+1}(s,a,h) - \sum_{k \leqslant t} w_h^{\pi_k^{\mathrm{exp}}}(s,a) \right| \leqslant \sqrt{2T \log(2T^2 SAH)} \right\} \tag{H.2}$$

holds with probability at least $1 - 1/(SAHT^2)$ so that

$$\mathcal{E}_H(T) \triangleq \bigcap_{s,a,h} \mathcal{E}_H^{s,a,h}(T), \tag{H.3}$$

holds with probability at least $1 - 1/T^2$. Therefore, thanks to Borell-Cantelli's lemma, there exits $\widetilde{T}$ such that with probability 1, for $T \geqslant \widetilde{T}, \mathcal{E}_H(T)$ holds which implies that for any s, a, h and $T \geqslant \widetilde{T}$,

$$\left| \tfrac{n_{T+1}(s,a,h)}{T} - \frac{1}{T} \sum_{t \leqslant T} w_h^{\pi_t^{\mathrm{exp}}}(s,a) \right| \leqslant \sqrt{\frac{2\log(2T^2SAH)}{T}}\,, \tag{H.4}$$

and further remark that, as the space of valid-action allocation is convex (see e.g.,A. J. Wagenmaker, Chen, et al. 2022), $\frac{1}{T}\sum_{t\in[T]} w_h^{\pi_t^{\mathrm{exp}}}$ is a valid state-action allocation for M.

Since transition probabilities may be arbitrarily small, the KL divergence in Υ_t is potentially unbounded. To handle this, we restrict to alternatives with bounded transitions. Concretely, observe that the model obtained by setting all transitions to uniform and all rewards to zero on states visited by the optimal policy of M lies in $\mathrm{Alt}(M)$, as M is trivially absolutely continuous with respect to it. Therefore, for any $\varepsilon > \log S$, the restricted alternative set

$$\mathrm{Alt}(M,\varepsilon) \triangleq \left\{ (\theta,\Phi) \in \mathrm{Alt}(M) : \forall\,(s,a,h,s'),\ \Phi(s' \mid s,a,h) \geqslant e^{-\varepsilon} \right\}$$

is non-empty. Since $\mathrm{Alt}(M,\varepsilon) \subset \mathrm{Alt}(M)$,

$$\begin{aligned} &\inf_{(\theta,\Phi)\in\mathrm{Alt}(M)} \Upsilon_t(\theta,\Phi) \leqslant \\ &\inf_{(\theta,\Phi)\in\mathrm{Alt}(M,T^{1/4})} \sum_{s,a,h} n_T(s,a,h) \left[\frac{\big(\hat{\mu}_T(s,a,h) - \theta(s,a,h)\big)^2}{2} + \sum_{s'} \hat{p}_T(s' \mid s,a,h) \log \frac{\hat{p}_T(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right]. \end{aligned} \tag{H.5}$$

By definition of $\mathrm{Alt}(M, T^{1/4})$, every (θ,Φ) in this set satisfies

$$\forall\,(s,a,h,s'): \quad \log \frac{1}{\Phi(s' \mid s,a,h)} \leqslant T^{1/4}, \tag{H.6}$$

which allows concentration arguments to be applied in the next step.

We define the event $\mathcal{E}_H^{\mu+p}(T) \triangleq \mathcal{E}_H^{\mu}(T) \cap \mathcal{E}_H^{p}(T)$, where

$$\mathcal{E}_H^{p}(T) \triangleq \left\{ \forall\, t \leqslant T : \sum_{s,a,h} n_t(s,a,h)\, \mathrm{KL}(\hat{p}_t(\cdot \mid s,a,h) \,\|\, p(\cdot \mid s,a,h)) \leqslant \beta^p(T, 1/T^2) \right\}, \tag{H.7}$$

$$\mathcal{E}_H^{\mu}(T) \triangleq \left\{ \forall\, t \leqslant T : \sum_{s,a,h} n_t(s,a,h)\, \mathrm{KL}\Big(\hat{R}_h^t(s,a) \,\|\, R_h(s,a)\Big) \leqslant \beta^{\mu}(T, 1/T^2) \right\}. \tag{H.8}$$

On $\mathcal{E}^{\mu}(T)$ we have,

$$\begin{aligned} \frac{\big(\hat{\mu}_T(s,a,h) - \theta(s,a,h)\big)^2}{2} &\leqslant \frac{(\mu(s,a,h) - \theta(s,a,h))^2}{2} + \frac{(\mu(s,a,h) - \hat{\mu}_T(s,a,h))^2}{2} + |\mu(s,a,h) - \hat{\mu}_T(s,a,h)| \\ &\leqslant \frac{\big(\mu(s,a,h) - \theta(s,a,h)\big)^2}{2} + \frac{\beta^{\mu}(T,1/T^2)}{n_T(s,a,h)} + \sqrt{\frac{2\beta^{\mu}(T,1/T^2)}{n_T(s,a,h)}}. \end{aligned} \tag{H.9}$$

Next, for any $q \in \triangle$, we let $\mathcal{S}_T^+ \triangleq \{s' : \hat{p}_T(s' \mid s,a,h) > 0\}$. For $s' \notin \mathcal{S}_T^+$, the convention $0 \log 0 = 0$ means the left-hand side contributes 0. By convexity of $x \mapsto x \log x$,

$$\begin{aligned} &\hat{p}_T(s' \mid s,a,h) \log \frac{\hat{p}_T(s' \mid s,a,h)}{q(s')} - p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{q(s')} \leqslant \\ &\left(1 + \log \frac{\hat{p}_T(s' \mid s,a,h)}{q(s')}\right) \big(\hat{p}_T(s' \mid s,a,h) - p(s' \mid s,a,h)\big), \quad s' \in \mathcal{S}_T^+. \end{aligned}$$

Summing over $s' \in \mathcal{S}_T^+$ yields

$$\begin{aligned} &\sum_{s'} \hat{p}_T(s' \mid s,a,h) \log \frac{\hat{p}_T(s' \mid s,a,h)}{q(s')} \leqslant \\ &\sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{q(s')} + \sum_{s' \in \mathcal{S}_T^+} \left(1 + \log \frac{\hat{p}_T(s' \mid s,a,h)}{q(s')}\right) \big(\hat{p}_T(s' \mid s,a,h) - p(s' \mid s,a,h)\big). \end{aligned} \tag{H.10}$$

We note that for $s' \in \mathcal{S}_T^+$, either the transition s' has been observed at least once from (s,a,h), or (s,a,h) has never been visited and $\hat{p}_T(s' \mid s,a,h) = 1/S$, by initialization. In all cases $\hat{p}_T(s' \mid s,a,h) \geqslant \min(1/T, 1/S)$. Therefore,

$$\left|1 + \log \frac{\hat{p}_T(s' \mid s,a,h)}{q(s')}\right| \leqslant 1 + \log(ST).$$

Substituting into (H.10) and using the triangle inequality,

$$\begin{aligned} \sum_{s'} \hat{p}_T(s' \mid s,a,h) \log \frac{\hat{p}_T(s' \mid s,a,h)}{q(s')} \leqslant \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{q(s')} + \\ \big(1 + \log(ST)\big) \left\|\hat{p}_T(\cdot \mid s,a,h) - p(\cdot \mid s,a,h)\right\|_1. \end{aligned} \tag{H.11}$$

Applying Pinsker's inequality on $\mathcal{E}_H^p(T)$,

$$\sum_{s'} \hat{p}_T(s' \mid s,a,h) \log \frac{\hat{p}_T(s' \mid s,a,h)}{q(s')} \leqslant \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{q(s')} + \big(1 + \log(ST)\big) \sqrt{\frac{2\beta^p(T,1/T^2)}{n_T(s,a,h)}}. \tag{H.12}$$

Substituting (H.9) and (H.12) into (H.5), applying Cauchy–Schwarz to the error terms, and using $\sum_{s,a,h} n_T(s,a,h) = TH$,

$$\begin{aligned}
&\inf_{(\theta,\Phi)\in \mathrm{Alt}(M)} \Upsilon_t(\theta,\Phi) \leqslant \\
&\inf_{(\theta,\Phi)\in \mathrm{Alt}(M,T^{1/4})} \sum_{s,a,h} n_T(s,a,h) \left[\frac{\big(\mu(s,a,h)-\theta(s,a,h)\big)^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right] + \\
&\quad SAH\beta^{\mu}(T,1/T^2) + \sqrt{2SAH^2T\,\beta^{\mu}(T,1/T^2)} \;+\; \big(1+\log(ST)\big)\sqrt{2SAH^2T\,\beta^{p}(T,1/T^2)}.
\end{aligned} \tag{H.13}$$

Since β^{μ} and β^{p} are at most logarithmic in T, the three error terms in (H.13) are all $o(T)$.

We then invoke Eq. (H.2), which proves that on the event $\mathcal{E}_H(T)$ we have

$$\begin{aligned}
&\inf_{(\theta,\Phi)\in \mathrm{Alt}(M,T^{1/4})} \sum_{s,a,h} n_T(s,a,h) \left[\frac{(\mu(s,a,h)-\theta(s,a,h))^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right] \leqslant \\
&\qquad \inf_{(\theta,\Phi)\in \mathrm{Alt}(M,T^{1/4})} \sum_{s,a,h} \sum_{t\in[T]} w_h^{\pi_t^{\exp}}(s,a) \left[\frac{(\mu(s,a,h)-\theta(s,a,h))^2}{2} + \right. \\
&\qquad\qquad \left. \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right] + SAH\sqrt{2T\log(2T^2SAH)}\big(\frac{1}{2}+T^{1/4}\big).
\end{aligned}$$

Combining with Eq. (H.13) yields

$$\begin{aligned}
&\inf_{(\theta,\Phi)\in \mathrm{Alt}(M)} \Upsilon_t(\theta,\Phi) \leqslant \\
&\inf_{(\theta,\Phi)\in \mathrm{Alt}(M,T^{1/4})} \sum_{s,a,h} \sum_{t\in[T-1]} w_h^{\pi_t^{\exp}}(s,a) \left[\frac{(\mu_h(s,a)-\theta_h(s,a))^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right] + \\
&\left[\big(1+\log(ST)\big)\sqrt{2SAH^2T\beta^{p}(T,1/T^2)} \right] \\
&+ \left[SAH\beta^{\mu}(T,1/T^2) + \sqrt{2SAH^2T\beta^{\mu}(T,1/T^2)} \right] + SAH\sqrt{2T\log(2T^2SAH)}\big(\frac{1}{2}+T^{1/4}\big)\,.
\end{aligned}$$

Next, noting that β^{μ}, β^{p}'s dependency on T is at most logarithmic (see e.g. Tirinzoni, Al-Marjani, et al. 2023), the second order term in the equation above is $o(T)$, which we rewrite as: on the event $\mathcal{E}_H(T)$,

$$\begin{aligned}
&\inf_{(\theta,\Phi)\in \mathrm{Alt}(M)} \Upsilon_t(\theta,\Phi) \leqslant \\
&\inf_{(\theta,\Phi)\in \mathrm{Alt}(M,T^{1/4})} \sum_{s,a,h} \sum_{t\in[T-1]} w_h^{\pi_t^{\exp}}(s,a) \left[\frac{\big(\mu(s,a,h)-\theta(s,a,h)\big)^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right] \;+\; o(T) \\
&= (T-1)\cdot \inf_{(\theta,\Phi)\in \mathrm{Alt}(M,T^{1/4})} \sum_{s,a,h} \frac{1}{T-1} \sum_{t\in[T-1]} w_h^{\pi_t^{\exp}}(s,a) \left[\frac{\big(\mu(s,a,h)-\theta(s,a,h)\big)^2}{2} + \right. \\
&\qquad \left. \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\Phi_h(s,a,s')} \right] + o(T) \\
&\leqslant (T-1)\;\cdot \sup_{w\in\Omega_M} \inf_{(\theta,\Phi)\in \mathrm{Alt}(M,T^{1/4})} \sum_{s,a,h} w_h(s,a) \left[\frac{(\mu(s,a,h)-\theta(s,a,h))^2}{2} + \sum_{s'} p(s' \mid s,a,h) \log \frac{p(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)} \right] + \\
&\qquad o(T)
\end{aligned}$$

which follows as, from the convexity of Ω_M we have $\frac{1}{T-1}\sum_{t=1}^{T-1} w.^{\pi_t^{\exp}} \in \Omega_M$. Finally, remark that the sequence $(u_T)_{T\geqslant 1}$ defined by

$$u_T \triangleq \sup_{w\in\Omega_M} \inf_{(\theta,\Phi)\in \mathrm{Alt}(M,T^{1/4})} \sum_{s,a,h} w_h(s,a)\left[\frac{\big(\mu(s,a,h)-\theta(s,a,h)\big)^2}{2} + \sum_{s'} p(s' \mid s,a,h)\log\frac{p(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)}\right]$$

is non-increasing and $\inf\{u_T : T\geqslant 1\} = \Gamma_M$. Thus, noting $\Pr(\mathcal{E}_H(T)\cap \mathcal{E}_H^{\mu+p}(T)) \geqslant 1 - O(1/T^2)$, thanks to Borel-Cantelli's lemma, we have with probability one,

$$\limsup_{T\to\infty} \frac{1}{T}\left[\inf_{(\theta,\Phi)\in\mathrm{Alt}(M)} \Upsilon_T(\theta,\Phi)\right] \leqslant \Gamma_M.$$

□

H.2 Proof of Theorem 2

Proof. The result follows from Proposition 33, Lemma 19, and an anti-concentration bound on the posterior. We proceed in two steps: first expressing the posterior probability of $\mathrm{Alt}(M)$ in terms of the GLR, then lower-bounding it.

On can verify that the posterior probability of $\mathrm{Alt}(M)$ is proportional to

$$\Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\Big(\widetilde{M}\in\mathrm{Alt}(M)\Big) \propto \int_{\mathrm{Alt}(M)} e^{-\Upsilon_t(\theta,\Phi)}\mathrm{d}\theta\,\mathrm{d}\Phi. \tag{H.14}$$

Defining

$$\Lambda_t \triangleq \int_{\mathrm{Alt}(M)} e^{-\Upsilon_t(\theta,\Phi)}\mathrm{d}\theta\,\mathrm{d}\Phi, \tag{H.15}$$

we obtain

$$\Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\Big(\widetilde{M}\in\mathrm{Alt}(M)\Big) = \frac{\Lambda_t}{\displaystyle\int_{\mathbf{M}} e^{-\Upsilon_t(\theta,\Phi)}\mathrm{d}\theta\,\mathrm{d}\Phi}. \tag{H.16}$$

Since Υ_t is non-negative (it is a sum of KL divergences and squared differences) and $\mathbf{M}$ is bounded,

$$\int_{\mathbf{M}} e^{-\Upsilon_t(\theta,\Phi)}\mathrm{d}\theta\,\mathrm{d}\Phi \leqslant \mathrm{vol}(\mathbf{M}).$$

Substituting into (H.16),

$$\Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\Big(\widetilde{M}\in\mathrm{Alt}(M)\Big) \geqslant \frac{\Lambda_t}{\mathrm{vol}(\mathbf{M})}. \tag{H.17}$$

By Lemma 19 (with $\eta = 1$), with probability one,

$$\begin{aligned}\log\Lambda_t &\geqslant -\inf_{(\theta,\Phi)\in\mathrm{Alt}(M)} \sum_{k=1}^{t-1}\sum_{h=1}^{H}\left[\frac{\big(\hat{\mu}_t(s_h^k,a_h^k,h)-\theta(s_h^k,a_h^k,h)\big)^2}{2} + \log\frac{\hat{p}_t(s_{h+1}^k \mid s_h^k,a_h^k,h)}{\Phi(s_{h+1}^k \mid s_h^k,a_h^k,h)}\right] - o(t)\\ &= -\inf_{(\theta,\Phi)\in\mathrm{Alt}(M)} \sum_{s,a,h} n_t(s,a,h)\left[\frac{\big(\hat{\mu}_t(s,a,h)-\theta(s,a,h)\big)^2}{2} + \sum_{s'}\hat{p}_t(s' \mid s,a,h)\log\frac{\hat{p}_t(s' \mid s,a,h)}{\Phi(s' \mid s,a,h)}\right] - o(t).\end{aligned} \tag{H.18}$$

Combining (H.17) and (H.18),

$$-\log\Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_t}\Big(\widetilde{M}\in\mathrm{Alt}(M)\Big) \leqslant \inf_{(\theta,\Phi)\in\mathrm{Alt}(M)} \Upsilon_t(\theta,\Phi) + \log\mathrm{vol}(\mathbf{M}) - o(t).$$

Since $\log \mathrm{vol}(\mathbf{M}) = O(1)$, dividing by t and taking the lim sup, Proposition 33 gives, with probability one,

$$\limsup_{t\to\infty} -\frac{1}{t}\log \Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\left(\widetilde{M}\in \mathrm{Alt}(M)\right) \leqslant \limsup_{t\to\infty}\frac{1}{t}\left[\inf_{(\theta,\Phi)\in\mathrm{Alt}(M)} \Upsilon_t(\theta,\Phi)\right] \leqslant \Gamma_M,$$

which completes the proof of Theorem 2. □

We now prove the main result of this section, showing that PIPS reaches the optimal posterior contraction rate with probability one.

Theorem 3. *Consider an MDP $M\in\mathbf{M}$ with a unique optimal policy $\pi^\star$. Let $\gamma\in(0,1)$, $\alpha\in(0,1/2)$, $\alpha<\gamma$, $\varsigma\in(0,1)$, with $\varsigma>2\alpha$. Run with parameterss γ,α and learning rate $\eta_t = O(t^{-\varsigma})$, letting ν_t denote the non-inflated posterior distribution, Algorithm 1 satisfies with probability one*

$$\limsup_{t\to\infty} -\frac{1}{t}\log \Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}\big(\pi^\star\notin\Pi^\star(\widetilde{M})\big) \;=\; \Gamma_M.$$

Proof. Thanks to Proposition 25, with probability at least $1-O(1/t^2)$, for t large enough

$$\mathrm{GLR}(t) \geqslant t\cdot\Gamma_M - O(\log t) - O(t^{1+\alpha-\gamma}) - o(t) - O(t^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log t}).$$

Thanks to Lemma 36 we have for

$$\Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}(\widetilde{M}\in\mathrm{Alt}(M)) \leqslant h(t)e^{-\mathrm{GLR}(t))},$$

where h is polynomial in t, where due to Lemma 12, we will have $\mathrm{Alt}(\hat{M}_t)=\mathrm{Alt}(M)$ (up to a zero measure set), for t large enough.

Therefore, with probability $1-O(1/t^2)$,

$$\log\Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}(\widetilde{M}\in\mathrm{Alt}(M)) \leqslant -t\Gamma_M + \log h(t) + O(\log t) + O(t^{1+\alpha-\gamma}) + o(t) + O(t^{\alpha+\frac{\gamma+1}{2}}\sqrt{\log t}),$$

then, thanks to Borel-Cantelli's lemma, with probability one,

$$\liminf_{t\to\infty} -\frac{1}{t}\log\Pr_{\widetilde{M}\sim\nu_t|\mathcal{H}_{t-1}}(\widetilde{M}\in\mathrm{Alt}(M)) \geqslant \Gamma_M,$$

which combined with Theorem 2 completes the proof. □

I SAMPLE COMPLEXITY GUARANTEES

In this section, we establish sample complexity guarantees regardless of the algorithm's correctness for the specified threshold, showing that when $B(t,\delta)$ is correctly calibrated, the posterior sampling rule combined with the sampling rule of PIPS attains the optimal expected sample complexity.

Theorem 4. *Let $\gamma\in(0,1)$, $\alpha\in(0,1/2)$, $\alpha<\gamma$, $\varsigma\in(0,1)$, with $\varsigma>2\alpha$. If $B(t,\delta)$ satisfies $\limsup_{\delta\to0}\frac{\log B(t,\delta)}{\eta_t\log\frac{1}{\delta}}\leqslant 1$, then, PIPS coupled with the stopping time τ_δ described in (5) and run with parameterss γ,α and learning rate $\eta_t = O(t^{-\varsigma})$, satisfies*

$$\mathbb{E}_M\left[\tau_\delta\right] \leqslant \Gamma_M^{-1}\log\frac{1}{\delta} + o\big(\log\tfrac{1}{\delta}\big)\,,\ \textit{and}$$
$$\Pr_M\left(\limsup_{\delta\to0}\frac{\tau_\delta}{\log\frac{1}{\delta}}\leqslant\Gamma_M^{-1}\right) = 1\,.$$

Proof. Let $(\mathcal{E}(t))_{t\geqslant1}$ be some events that will be explicit below.

We have

$$\begin{aligned}
\mathbb{E}_M[\tau_\delta] &\leqslant 1+\sum_{t=1}^{\infty}\mathrm{Pr}_M(\tau_\delta > t)\,,\\
&\leqslant 1+\sum_{t=1}^{\infty}\mathrm{Pr}_M(\exists\, k \leqslant B(t,\delta) : \hat{\pi}_t \notin \Pi^\star(\widetilde{M}_{t,k}))\,,\\
&\leqslant 1+\sum_{t=1}^{\infty}\mathbb{E}_M\left[\mathbf{1}(\mathcal{E}(t))\,\mathrm{Pr}_M(\exists\, k \leqslant B(t,\delta) : \hat{\pi}_t \notin \Pi^\star(\widetilde{M}_{t,k})|\mathcal{H}_{t-1})\right]+\sum_{t=1}^{\infty}\mathrm{Pr}_M(\mathcal{E}(t)^c)\,,\\
&\leqslant 1+\sum_{t=1}^{\infty}\mathrm{Pr}_M(\mathcal{E}(t)^c)+\sum_{t=1}^{\infty}\mathbb{E}_M\left[B(t,\delta)\mathbf{1}(\mathcal{E}(t))\,\mathrm{Pr}_{\widetilde{M}\sim\nu_t^{\eta_t}|\mathcal{H}_{t-1}}\left(\hat{\pi}_t \notin \Pi^\star(\widetilde{M})\right)\right],
\end{aligned}$$

which follows since conditionally on $\mathcal{H}_{t-1}$, $\widetilde{M}_{t,1}, \widetilde{M}_{t,2}, \dots$ are i.i.d. samples from distribution with conditional density $\nu_t^{\eta_t}(\cdot)$.

Thus taking k_0 such that $\hat{\pi}_t = \pi^\star$ for any $t \geqslant k_0$, we have

$$\mathbb{E}_M[\tau_\delta] \leqslant 1+k_0+\sum_{t=1}^{\infty}\mathrm{Pr}_M(\mathcal{E}(t)^c)+\sum_{t=1}^{\infty}\mathbb{E}\left[B(t,\delta)\mathrm{Pr}_{\widetilde{M}\sim\nu_t^{\eta_t}|\mathcal{H}_{t-1}}\left(\widetilde{M}\in \mathrm{Alt}(M)\right)\right].$$

Next, thanks to Lemma 36 we have

$$\mathrm{Pr}_{\widetilde{M}\sim\nu_t^{\eta_t}|\mathcal{H}_{t-1}}(\widetilde{M}\in\mathrm{Alt}(M)) \leqslant f(t)e^{-\eta_t\,\mathrm{GLR}(t)},$$

where f is at most polynomial in t.

By Proposition 25, there exists an event $\mathcal{E}(t)$, which holds with probability $1-O(1/t^2)$ such that we have

$$\mathrm{GLR}(t) \geqslant \Gamma_M\cdot t - h(t)$$

where h is a deterministic sublinear function of t satisfying $h(t)=o(t)$.

Combining the above results, we have

$$\mathbf{1}(\mathcal{E}(t))\,\mathrm{Pr}_{\widetilde{M}\sim\nu_t^{\eta_t}|\mathcal{H}_{t-1}}(\widetilde{M}\in\mathrm{Alt}(M)) \leqslant f(t)\exp\left(-\Gamma_M t\eta_t+\eta_t h(t)\right).$$

Thus we define

$$T(\delta) \triangleq \sup\left\{t\geqslant 1 : -\log(B(t,\delta))-\log f(t)+\eta_t t\cdot\Gamma_M-\eta_t h(t)\leqslant \log(t\log(t)^{1+\sigma})\right\},$$

so we have

$$\mathbb{E}_M[\tau_\delta] \leqslant T_0+T(\delta)+\sum_{t>2}\frac{1}{t\log(t)^{1+\sigma}}, \tag{I.1}$$

where $T_0<\infty$ is a constant.

We note that

$$\sum_{t>2}\frac{1}{t\log(t)^{1+\sigma}}<\infty$$

for $\sigma>0$.

Therefore, it remains to control the quantity $T(\delta)$ and for this remark that

$$\begin{aligned}
T(\delta) &= \sup\left\{t\geqslant 1 \middle| t\Gamma_M\eta_t-\log(B(t,\delta))-\log f(t)-\eta_t h(t)\leqslant\log(t\log(t)^{1+\sigma})\right\},\\
&= \sup\left\{t\geqslant 1\middle| t\leqslant\Gamma_M^{-1}\left(\frac{1}{\eta_t}\log(B(t,\delta))+\frac{1}{\eta_t}\log f(t)+h(t)-\frac{1}{\eta_t}\log(t\log(t)^{1+\sigma})\right)\right\},
\end{aligned}$$

Now we remark in the above the leading term in δ dependency is the term in $\frac{1}{\eta_t}\log(B(t,\delta))$. □

Bounding $T(\delta)$ First note that f is at most polynomial and $h(t)$ is sublinear in t, as well as $\frac{1}{\eta_t}$ is sublinear in t. Thus,

$$q(t) \triangleq \frac{1}{\eta_t}\log B(t,\delta) + \frac{1}{\eta_t}\log f(t) + h(t) - \frac{1}{\eta_t}\log\big(t\log(t)^{1+\sigma}\big)$$

is sublinear in t. Hence, there exists $\varepsilon \in (0,1)$ such that $q(t) = o_{t\to\infty}(t^\varepsilon)$. Next, observe that $q(\log(1/\delta)^{1/\varepsilon}) = o(\log(1/\delta))$. Thus, we have for $t_\delta = \log(1/\delta)^{1/\varepsilon}$

$$\limsup_{\delta\to 0} \frac{\frac{1}{\eta_{t_\delta}}\log(B(t_\delta,\delta)) + q(t_\delta)}{\log\frac{1}{\delta}} \leqslant 1.$$

Let us introduce

$$b(t_\delta) \triangleq \frac{1}{\eta_{t_\delta}}\log(B(t_\delta,\delta)) + q(t_\delta)\,.$$

Let $\delta_{\min} \in (0,1)$ be defined as

$$\begin{aligned}\delta_{\min} &\triangleq \inf\Big\{\delta\in(0,1) \mid b(t_\delta) > \log(1/\delta)^{1/\varepsilon}\Gamma_M\Big\}\\ &= \inf\big\{\delta\in(0,1) \mid \Gamma_M^{-1} b(t_\delta) > t_\delta\big\}\,,\end{aligned}$$

which is well defined as $\limsup_{\delta\to 0}\frac{b(t_\delta)}{\log\frac{1}{\delta}} \leqslant 1$ and $\varepsilon\in(0,1)$. Let $T_{\max} = \log(1/\delta_{\min})^{1/\varepsilon}$. For all $t \geqslant T_{\max}$, there exists $(0,1)\ni\delta' \leqslant \delta_{\min}$ such that $\lfloor t_{\delta'}\rfloor = t$ and $\Gamma_M^{-1}b(t_{\delta'}) < t_{\delta'}$. Therefore, for all $\delta\leqslant\delta_{\min}$

$$T(\delta) \leqslant \log(1/\delta)^{1/\varepsilon}\,. \tag{I.2}$$

Moreover, by definition $T(\delta)\leqslant\Gamma_M^{-1}b(T(\delta))$ and b is increasing for small δ, it follows that

$$T(\delta) \leqslant \Gamma_M^{-1}\cdot b(\log(1/\delta)^{1/\varepsilon}). \tag{I.3}$$

Combining (I.1), (I.2) and (I.3), it follows that for all $\delta\leqslant\delta_{\min}$,

$$\mathbb{E}_M[\tau_\delta] \leqslant \Gamma_M^{-1}\cdot b(\log(1/\delta)^{1/\varepsilon}) + O(1). \tag{I.4}$$

Finally, since $\limsup_{\delta\to 0}\frac{b(\log(1/\delta)^{1/\varepsilon})}{\log\frac{1}{\delta}} \leqslant 1$, we conclude that

$$\limsup_{\delta\to 0}\frac{\mathbb{E}_M[\tau_\delta]}{\log(1/\delta)} \leqslant \Gamma_M^{-1}.$$

Almost-sure upper bound The almost-sure bound on the sample complexity is derived from an application of Borel-Cantelli's lemma. Indeed, introducing for any T the event

$$E(T) \triangleq \left\{\sum_{t=1}^T \mathbf{1}(\tau_\delta > t) - \sum_{t=1}^T \Pr(\tau_\delta > t \mid \mathcal{H}_{t-1}) \leqslant \sqrt{2T\log T^2}\right\}\,,$$

we have when $E(T)$ holds

$$\begin{aligned}\min(\tau_\delta, T) &\leqslant 1 + \sum_{t=1}^T \mathbf{1}(\tau_\delta > t)\\ &\leqslant 1 + \sum_{t=1}^T \Pr(\tau_\delta > t \mid \mathcal{H}_{t-1}) + \sqrt{2T\log T^2}\\ &\leqslant T(\delta) + \sqrt{2T\log T^2} + T_0\,,\end{aligned}$$

where the last inequality follows from the derivations above on the event $\mathcal{E}(T)$ and holds with probability larger than $1 - O(1/T^2)$. Next, assume $\mathcal{E}(T)$ holds. Then, if

$$T \geqslant T'(\delta) \triangleq \sup\Big\{t\geqslant 1 : t < T_0 + T(\delta) + \sqrt{2t\log t^2}\Big\}\,,$$

and $E(T)$ holds then the stopping rule would trigger before episode T. Thanks to Azuma-Hoeffding $E(T)$ holds with probability at least $1-1/T^2$ for any $T \geqslant 1$. Thus $E(T)^c$ holds with probability at most $1/T^2$, similarly for $\mathcal{E}(T)^c$ and, as $\sum_{T\geqslant 1} \frac{1}{T^2} < \infty$, we invoke Borel-Cantelli's lemma to justify that $\Pr(\limsup_{T\to\infty}(E(T)\cap\mathcal{E}(T))^c) = 0$ so with probability 1, there exists $\widetilde{T}$ (possibly random) such that for $T \geqslant \widetilde{T}$, $E(T)\cap\mathcal{E}(T)$ holds.

From the calculations above, we remark that the leading term in $T'(\delta)$ is at most of order $\log(1/\delta)$. Let $\delta'_{\min}$ be such that $\forall\, \delta \leqslant \delta'_{\min}$,

$$\lceil \log(1/\delta)^{3/2} \rceil > T'(\delta),$$

which is well-defined. Let $\delta \leqslant \delta'_{\min}$ such that additionally $\lceil \log(1/\delta)^{3/2} \rceil > \widetilde{T}$. We then have

$$\begin{aligned}
\tau_\delta &\leqslant T_0 + T(\delta) + \sqrt{2\lceil \log(1/\delta)^{3/2} \rceil \log\lceil \log(1/\delta)^{3/2} \rceil^2} \\
&\leqslant T_0 + T(\delta) + \log(1/\delta)^{3/4} \sqrt{8 \log(1+\log(1/\delta)^{3/2})}\,.
\end{aligned}$$

Finally recalling that

$$\limsup_{\delta\to 0} \frac{T(\delta)}{\log\frac{1}{\delta}} \leqslant \Gamma_M^{-1}\,.$$

We conclude by noting that

$$\begin{aligned}
\limsup_{\delta\to 0} \frac{\tau_\delta}{\log\frac{1}{\delta}} &\leqslant \limsup_{\delta\to 0} \frac{T(\delta)}{\log\frac{1}{\delta}} + \lim_{\delta\to 0} \frac{T_0 + \log(1/\delta)^{3/4}\sqrt{8\log(1+\log(1/\delta)^{3/2})}}{\log\frac{1}{\delta}} \\
&\leqslant \Gamma_M^{-1}\,.
\end{aligned}$$

J CONCENTRATION RESULTS

J.1 Self-noramlized Concentration

We prove some concentration results used in other sections. We define the following concentration events of the self-normalized quantities related to the GLR

$$\begin{aligned}
\mathcal{E}^p(\delta) &\triangleq \left\{ \forall t \geqslant 1,\ \sum_{s,a,h} n_t(s,a,h)\, \mathrm{KL}(\hat{p}_t(\cdot \mid s,a,h) \parallel p(\cdot \mid s,a,h)) \leqslant \beta^p(t,\delta) \right\} \\
\mathcal{E}^\mu(\delta) &\triangleq \left\{ \forall t \geqslant 1,\ \sum_{s,a,h} n_t(s,a,h)\, \mathrm{KL}(\hat{R}_h^t(s,a) \parallel R_h(s,a)) \leqslant \beta^\mu(t,\delta) \right\} \\
\mathcal{E}^{\mu+p}(\delta) &\triangleq \left\{ \forall t \geqslant 1,\ \sum_{s,a,h} n_t(s,a,h)\, \mathrm{KL}(\hat{R}_h^t(s,a) \otimes \hat{p}_t(\cdot \mid s,a,h) \parallel R_h(s,a) \otimes p(\cdot \mid s,a,h)) \leqslant \beta^{\mu+p}(t,\delta) \right\}
\end{aligned}$$

It is possible to choose $\beta^p, \beta^\mu, \beta^{\mu+p}$ with logarithmic dependency to ensure that each event hold with probability $1-\delta$ (cf e.g., Kaufmann & Koolen 2021). Since our focus is on characterizing the sample complexity in the asymptotic regime, we do not make these terms explicit. The interested reader may refer to Kaufmann & Koolen 2021 for explicit expressions of β^p, β^μ, and $\beta^{\mu+p}$.

Moreover, the event

$$\mathcal{E}^{\mathrm{cnt}}(T) \triangleq \left\{ \forall t \geqslant 1,\ \forall (s,a,h) :\ n_t(s,a,h) \geqslant \tfrac{1}{2}\, \bar{n}_t(s,a,h) - \log(2SAH/\delta) \right\}.$$

holds with probability at least $1-\delta$, thanks to Lemma 34.

Lemma 34 (Dann, Lattimore, et al. 2017)**.** *Let $X_1, X_2, \ldots$ be a sequence of Bernoulli random variables adapted to a filtration $\{\mathcal{H}_t\}_{t\geqslant 1}$ and such that for any i, $\Pr(X_i = 1 \mid \mathcal{H}_{i-1}) = P_i$. We have*

$$\Pr\Big(\exists n \geqslant 1 : \sum_{t\leqslant n} X_t \leqslant \frac{1}{2} \sum_{t\leqslant n} P_t - W \Big) \leqslant \exp(-W)\,.$$

J.2 Gaussian and Dirichlet Concentration

We leverage Lemma 35 to prove concentration of distribution over MDPs in $\mathbf{M}$.

Lemma 35 (Lu & W. V. Li 2009)**.** *Let $X \sim \mathcal{N}(\theta, \Sigma)$ be a multivariate normal vector in dimension d. For any convex set $C \subset \mathbb{R}^d$*

$$\Pr_{X\sim\mathcal{N}(\theta,\Sigma)}(X \in C) \leqslant \frac{1}{2} e^{-\inf_{\lambda\in C} \frac{1}{2}\|\lambda-\theta\|^2_{\Sigma^{-1}}}.$$

We note that from the proof of Lemma 35 in Lu & W. V. Li 2009, the result extends to multivariate normals truncated to to a convex set.

We prove the following crucial concentration.

Lemma 36. *Let $P \triangleq (P(s,a,h))_{s,a,h}$ be a distribution whose marginals $(P(s,a,h))_{s,a,h}$ are independent $\mathrm{Dir}((\alpha(s' \mid s,a,h)+1)_{s'\in\mathcal{S}})$ and $R \triangleq (R(s,a,h))_{s,a,h}$ where each $R(s,a,h)$ is an independent $\mathcal{N}(\mu(s,a,h), \sigma^2(s,a,h))$, possibly truncated to a convex set $\mathcal{C}$. Let $\mathcal{X} \subset \mathbf{M}$ such that for any $q \in \mathcal{P}$, $\mathcal{X}_q \triangleq \left\{\theta \in (0,1)^{SAH} : (\theta, q) \in \mathcal{X}\right\}$ is convex. Introducing $(\kappa_h)_h$ satisfying $\kappa(s' \mid s,a,h)\alpha(s,a,h) = \alpha(s' \mid s,a,h)\,, \forall (s,a,h,s')$, we have*

$$\begin{aligned}
&\Pr((R,Q) \in \mathcal{X}) \\
&\leqslant \frac{1}{2}\left(\prod_{s,ah} \max\left(1, S\alpha(s,a,h)^{S-1} e^{-\mathrm{Ent}(\kappa(\cdot|s,a,h))}\right)\right) \exp\Bigg\{-\inf_{(\theta,q)\in\mathcal{X}}\Bigg[\sum_{s,a,h}\frac{(\mu(s,a,h)-\theta(s,a,h))^2}{2\sigma^2(s,a,h)} + \\
&\quad \sum_{s,a,h}\alpha(s,a,h)\,\mathrm{KL}(\kappa(\cdot \mid s,a,h)\|q(\cdot \mid s,a,h))\Bigg]\Bigg\}.
\end{aligned}$$

Proof. We have $\mathcal{X} \subset \mathbf{M} \triangleq (0,1)^{SAH} \times \triangle^{SAH}$ and by assumption, for any $q \in \triangle^{SAH}$, the set $\mathcal{X}_q = \{r \in \mathcal{R} : (r,q) \in \mathcal{X}\}$ is convex. Next, P has the same distribution as a multivariate normal in dimension SAH with independent marginals. Noting that $\Pr((R,P) \in \mathcal{X}) = \mathbb{E}\left[\Pr((R,P) \in \mathcal{X} \mid P)\right]$, and, since R, P are independent, leveraging Lemma 35 we have

$$\begin{aligned}
\Pr((R,P) \in \mathcal{X}) &= \mathbb{E}\left[\Pr((R,P) \in \mathcal{X} \middle| P)\right] \\
&\leqslant \frac{1}{2}\mathbb{E}\left[\exp\left\{-\inf_{\theta \text{ s.t } (\theta,P)\in\mathcal{X}} \sum_{s,a,h}\frac{(\mu(s,a,h)-\theta(s,a,h))^2}{2\sigma^2(s,a,h)}\right\}\right] \qquad \text{(J.1)}
\end{aligned}$$

which invokes Lemma 35 since given $P \in \mathcal{P}$, $\mathcal{X}_P$ as defined above is convex and R is a multivariate normal vector.

We recall that for $\alpha \in \mathbb{R}^d$, the density of $\mathrm{Dir}(\alpha)$ is proportional to

$$e^{-\sum_i(\alpha_i-1)\log\frac{1}{x_i}}.$$

Next, we have for any (s,a,h), letting $q(\cdot \mid s,a,h) \in \triangle$,

$$\begin{aligned}
\sum_{s'} \alpha(s' \mid s,a,h) \log \frac{1}{q(s' \mid s,a,h)} &= \alpha(s,a,h) \sum_{s'} \kappa(s' \mid s,a,h) \log \frac{1}{q(s' \mid s,a,h)} \\
&= \alpha(s,a,h) \sum_{s'} \kappa(s' \mid s,a,h) \log \frac{\kappa(s' \mid s,a,h)}{q(s' \mid s,a,h)} - \\
&\quad \alpha(s,a,h) \sum_{s'} \kappa(s' \mid s,a,h) \log \kappa(s' \mid s,a,h) \\
&= \alpha(s,a,h) \sum_{s'} \kappa(s' \mid s,a,h) \log \frac{\kappa(s' \mid s,a,h)}{q(s' \mid s,a,h)} - \\
&\quad \sum_{s'} \alpha(s' \mid s,a,h) \log \kappa(s' \mid s,a,h), \\
&= \alpha(s,a,h) \operatorname{KL}(\kappa(\cdot \mid s,a,h) \,\|\, q(\cdot \mid s,a,h)) - \sum_{s'} \alpha(s' \mid s,a,h) \log \kappa(s' \mid s,a,h) \,.
\end{aligned}$$

Thus, the density of $\operatorname{Dir}(1+\alpha(\cdot \mid s,a,h))$ is also proportional to

$$e^{-\alpha(s,a,h) \operatorname{KL}(\kappa(s' \cdot \mid s,a,h) \,\|\, q(s' \cdot \mid s,a,h)))} .$$

Combining with Eq.(J.1) we have

$$\begin{aligned}
\Pr((R,P) \in \mathcal{X}) \leqslant \frac{1}{2N} \int_{\mathcal{P}} \exp\left\{ - \inf_{\theta \text{ s.t } (\theta,q) \in \mathcal{X}} \sum_{s,a,h} \frac{(\mu(s,a,h) - \theta(s,a,h))^2}{2\sigma^2(s,a,h)} \right\} \cdot \\
\exp\left\{ - \sum_{s,a,h} \alpha(s,a,h) \operatorname{KL}(\kappa(\cdot \mid s,a,h) \,\|\, q(\cdot \mid s,a,h)) \right\} \mathrm{d}q \,,
\end{aligned}$$

where

$$N \triangleq \int_{\mathcal{P}} \exp\left\{ - \sum_{s,a,h} \alpha(s,a,h) \operatorname{KL}(\kappa(\cdot \mid s,a,h) \,\|\, q(\cdot \mid s,a,h)) \right\} \mathrm{d}q \,. \tag{J.2}$$

Next,

$$\begin{aligned}
\Pr((R,P) \in \mathcal{X}) \quad \leqslant \quad \frac{\operatorname{vol}(\mathcal{P})}{2N} \cdot \exp\left\{ - \inf_{q \in \mathcal{P}} \left[\inf_{\theta \text{ s.t. } (\theta,q) \in \mathcal{X}} \sum_{s,a,h} \frac{(\mu(s,a,h) - \theta(s,a,h))^2}{2\sigma^2(s,a,h)} + \right. \right. \\
\left. \left. \sum_{s,a,h} \alpha_h(s,a) \operatorname{KL}(\kappa(\cdot \mid s,a,h) \,\|\, q(\cdot \mid s,a,h)) \right] \right\} ,
\end{aligned}$$

then applying Lemma 40 yields

$$\begin{aligned}
\inf_{q \in \mathcal{P}} \left[\inf_{\theta \text{ s.t } (\theta,q) \in \mathcal{X}} \sum_{s,a,h} \frac{(\mu(s,a,h) - \theta(s,a,h))^2}{2\sigma^2(s,a,h)} + \sum_{s,a,h} \alpha(s,a,h) \operatorname{KL}(\kappa(\cdot \mid s,a,h) \,\|\, q(\cdot \mid s,a,h)) \right] = \\
\inf_{(\theta,q) \in \mathcal{X}} \left[\sum_{s,a,h} \frac{(\mu(s,a,h) - \theta(s,a,h))^2}{2\sigma^2(s,a,h)} + \sum_{s,a,h} \alpha(s,a,h) \operatorname{KL}(\kappa(\cdot \mid s,a,h) \,\|\, q(\cdot \mid s,a,h)) \right] ,
\end{aligned}$$

thus we have

$$\begin{aligned}
\Pr((R,Q) \in \mathcal{X}) \leqslant \frac{\operatorname{vol}(\mathcal{P})}{2N} \exp\Bigg\{ - \inf_{(\theta,q) \in \mathcal{X}} \Bigg[\sum_{s,a,h} \frac{(\mu(s,a,h) - \theta(s,a,h))^2}{2\sigma^2(s,a,h)} + \\
\sum_{s,a,h} \alpha(s,a,h) \operatorname{KL}(\kappa(\cdot \mid s,a,h) \,\|\, q(\cdot \mid s,a,h)) \Bigg] \Bigg\},
\end{aligned}$$

We remark that N rewrites as

$$\prod_{s,a,h}\left[\int_{\triangle} e^{-\alpha(s,a,h)\,\mathrm{KL}(\kappa(\cdot|s,a,h)\|q)}\mathrm{d}q\right],$$

then we invoke Lemma 37 which yields

$$\int_{\triangle} e^{-\alpha(s,a,h)\,\mathrm{KL}(\kappa(\cdot|s,a,h)\|q)}\mathrm{d}q \geqslant \min\left\{\frac{e^{\mathrm{Ent}(\kappa(\cdot|s,a,h))}\alpha(s,a,h)^{-(S-1)}}{S!}, \frac{1}{(S-1)!}\right\},$$

where $\mathrm{Ent}(p)$ denotes the entropy.

Combining the above displays and noting that $\mathrm{vol}(\triangle) = \frac{1}{(S-1)!}$ and $\mathrm{vol}(\mathcal{P}) = (\mathrm{vol}(\triangle))^{SAH}$ yields

$$\begin{aligned}
&\Pr((R,Q)\in\mathcal{X})\\
&\leqslant \frac{1}{2}\left(\prod_{s,ah}\max\left(1, S\alpha(s,a,h)^{S-1}e^{-\mathrm{Ent}(\kappa(\cdot|s,a,h))}\right)\right)\exp\Bigg\{-\inf_{(\theta,q)\in\mathcal{X}}\Bigg[\sum_{s,a,h}\frac{(\mu(s,a,h)-\theta(s,a,h))^2}{2\sigma^2(s,a,h)}+\\
&\quad\sum_{s,a,h}\alpha(s,a,h)\,\mathrm{KL}(\kappa(\cdot\mid s,a,h)\|q(\cdot\mid s,a,h))\Bigg]\Bigg\}.
\end{aligned}$$

□

The following lemma bounds the normalizing constant of the posterior Dirichlet distribution when its density is expressed in terms of KL divergence.

Lemma 37. *Let $q\in\triangle$ (the probability simplex of $\mathbb{R}^d$) and denote by $\mathrm{Ent}(q)\triangleq\sum_i -q_i\log q_i$ the entropy of q. We have*

$$\int_{\triangle} e^{-n\,\mathrm{KL}(q\,\|\,x)}\mathrm{d}x \geqslant \min\left\{\frac{e^{\mathrm{Ent}(q)}n^{-(d-1)}}{d!}, \frac{1}{(d-1)!}\right\}.$$

Proof. Below $\mathrm{d}x$ denotes a Hausdorff-measure differential element and, (only) below, Γ is the gamma function. For $n=0$, the bound follows as $\int_{\triangle}\mathrm{d}x = \mathrm{Vol}(\triangle) = \frac{1}{\Gamma(d)} = \frac{1}{(d-1)!}$. Next, we assume $n\geqslant 1$. Let $\zeta\in(0,1)$ and observe that

$$\begin{aligned}
\int_{\triangle} e^{-n\,\mathrm{KL}(q\,\|\,x)}dx &\geqslant \int_{(1-\zeta)q+\zeta\triangle} e^{-n\,\mathrm{KL}(q\,\|\,x)}\mathrm{d}\mathrm{d}x\\
&= \zeta^{d-1}\int_{\triangle} e^{-n\,\mathrm{KL}(q\,\|\,(1-\zeta)q+\zeta x)}\mathrm{d}x\\
&\geqslant \zeta^{d-1}\int_{\triangle} e^{-n(1-\zeta)\,\mathrm{KL}(q\,\|\,q)-n\zeta\,\mathrm{KL}(q\,\|\,x)}\mathrm{d}x
\end{aligned}$$

which follows by convexity of $x\mapsto\mathrm{KL}(q\,\|\,x)$. Combining with $\mathrm{KL}(q\,\|\,q)=0$ and letting $\zeta=1/n$ it follows that

$$\int_{\triangle} e^{-n\,\mathrm{KL}(q\,\|\,x)}dx \geqslant n^{-(d-1)}\int_{\triangle} e^{-\,\mathrm{KL}(q\|x)}\mathrm{d}x.$$

Next, we bound the right-hand integral. We observe that

$$\mathrm{KL}(q\|x) = \sum_i q_i\log\frac{q_i}{x_i} = -\mathrm{Ent}(q) - \sum_i q_i\log x_i\,,$$

where $\mathrm{Ent}(q)\triangleq\sum_i -q_i\log q_i$. Combining above displays yield

$$\int_{\triangle} e^{-\,\mathrm{KL}(q\|x)}\mathrm{d}x = e^{\mathrm{Ent}(q)}\int_{\triangle}\prod_i x_i^{q_i}\mathrm{d}x$$

Next, observe that $\int_\triangle \prod_i x_i^{q_i} \mathrm{d}x$ is the normalizing constant of $\mathrm{Dir}(1+q)$ thus

$$\begin{aligned}\int_\triangle \prod_i x_i^{q_i} \mathrm{d}x &= \frac{\prod_i \Gamma(1+q_i)}{\Gamma(\sum_i 1+q_i)} \\ &= \frac{\prod_i \Gamma(1+q_i)}{\Gamma(d+1)} = \frac{\prod_i \Gamma(1+q_i)}{d!},\end{aligned}$$

where Γ is the gamma function. As proven in Z.-H. Yang et al. 2017, $\Gamma(1+x) \geqslant x^x$ for any $0 \leqslant x \leqslant 1$. Combining the above displays, we have

$$\begin{aligned}\int_\triangle e^{-\mathrm{KL}(q \,\|\, x)} \mathrm{d}x &\geqslant \frac{e^{\mathrm{Ent}(q)}}{d!} \prod_i \frac{\Gamma(1+q_i)}{q_i^{q_i}} \\ &\geqslant \frac{e^{\mathrm{Ent}(q)}}{d!} \prod_i \frac{q_i^{q_i}}{q_i^{q_i}} = \frac{e^{\mathrm{Ent}(q)}}{d!}\,.\end{aligned}$$

Therefore

$$\int_\triangle e^{-n\,\mathrm{KL}(q \,\|\, x)} \mathrm{d}x \geqslant \frac{e^{\mathrm{Ent}(q)} n^{-(d-1)}}{d!}\,. \qquad \square$$

K TECHNICAL RESULTS

We state different technical results used in the other sections.

The following result is known and appears in Krichene et al. 2015; we include a proof here for completeness.

Lemma 38. *Let ρ be a density wrt canonical measure of $\mathcal{X}$ and $g : \mathcal{X} \to \mathbb{R}$*

$$\eta \mapsto \frac{1}{\eta} \log \int \rho(x) \exp(-\eta g(x)) \mathrm{d}x$$

is non-decreasing.

Proof. Let $\Lambda_\eta \triangleq \int \rho(x) \exp(-\eta g(x)) \mathrm{d}x$ and $f(\eta) = \frac{1}{\eta} \log \Lambda_\eta$ and let the density $w_\eta(x) = \frac{\rho(x) \exp(-\eta g(x))}{\Lambda_\eta}$ and $X \sim w_\eta$. By derivation under the integral sign

$$\begin{aligned}f'(\eta) &= -\frac{1}{\eta^2} \log \Lambda_\eta + \frac{1}{\eta} \frac{\int -g(x)\rho(x) \exp\left(-\eta g(x)\right) \mathrm{d}x}{\int \rho(x) \exp\left(-\eta g(x)\right) \mathrm{d}x} \\ &= -\frac{1}{\eta^2} \log \Lambda_\eta + \frac{1}{\eta^2} \mathbb{E}_X \left[\log \exp(-\eta g(X))\right] \\ &= \frac{1}{\eta^2} \mathbb{E}_X \left[\log \frac{\exp(-\eta g(X))}{\Lambda_\eta}\right] \\ &= \frac{1}{\eta^2} \mathbb{E}_X \left[\log \frac{w_\eta(X)}{\rho(X)}\right] \\ &= \frac{1}{\eta^2} \mathrm{KL}(\nu_\eta, \nu_\rho) \geqslant 0\,,\end{aligned}$$

where ν_η is the distribution with density w_η and ν_ρ the distribution with distribution ρ. $\square$

Lemma 39. *For any mesaurable set $A \subset \mathbf{M}$ there exists an aboluste constant $k_0 < \infty$ such that for $t \geqslant k_0$,*

$$\mathrm{vol}(A \cap \mathbf{M}^{1/S \cdot \sqrt{t}}) \geqslant \frac{1}{2} \mathrm{vol}(A)\,.$$

Proof. The sequence $a_t \triangleq \mathrm{vol}(A \cap \mathbf{M}^{1/S \cdot \sqrt{t}})$ is non-decreasing and upper bounded by $\mathrm{vol}(A)$, which is also its supremum, thus it converges to this value, and the conclusion follows from this observation. $\square$

Lemma 40. *Let Δ be a subset of $\mathcal{X}\times\mathcal{Y}$ and $f:\mathcal{S}\to\mathbb{R}$ be such that for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$, $f(x,y)=g(x)+h(y)$ with $g:\mathcal{X}\to\mathbb{R}$ and $h:\mathcal{Y}\to\mathbb{R}$. Then*

$$\inf_{x\in\mathcal{X}}\left[g(x)+\inf_{\substack{y\in\mathcal{Y}\\ s.t\ (x,y)\in\Delta}} h(y)\right]=\inf_{(x,y)\in\Delta} f(x,y)$$

Proof.

$$\begin{aligned}\inf_{(x,y)\in\Delta} f(x,y) &= \inf_{x\in\mathcal{X}}\inf_{\substack{y\in\mathcal{Y}\\ \text{s.t }(x,y)\in\Delta}} f(x,y)\\ &= \inf_{x\in\mathcal{X}}\inf_{\substack{y\in\mathcal{Y}\\ \text{s.t }(x,y)\in\Delta}} [g(x)+h(y)]\\ &= \inf_{x\in\mathcal{X}}\left[g(x)+\inf_{\substack{y\in\mathcal{Y}\\ \text{s.t }(x,y)\in\Delta}} h(y)\right].\end{aligned}$$

□

Lemma 41. *Let $M,\widetilde{M}\in\mathbf{M}$ satisfy $\|M-\widetilde{M}\|_1\leqslant\gamma$. Then for any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ and any $\pi\in\Pi_{\mathrm{det}}$,*

$$\left|\tilde{Q}_h^\pi(s,a)-Q_h^\pi(s,a)\right|\leqslant(H-h+1)\,\gamma.$$

Proof. Conditioning on $s_h=s$, $a_h=a$ and following π thereafter, the value difference lemma (Lemma 13) applied between M and $\widetilde{M}$ gives

$$Q_h^\pi(s,a)-\tilde{Q}_h^\pi(s,a)=\mathbb{E}_M^\pi\left[\sum_{k=h}^{H}\delta_k(s_k,a_k)\,\middle|\,s_h=s,\,a_h=a\right],$$

where

$$\delta_k(s,a)\triangleq\big(\mu(s,a,k)-\tilde{\mu}(s,a,k)\big)+\big(p_k(\cdot\mid s,a)-\tilde{p}_k(\cdot\mid s,a)\big)^\top\tilde{V}_{k+1}^\pi(\cdot).$$

Since rewards lie in $(0,1)$, $\|\tilde{V}_{k+1}^\pi\|_\infty\leqslant H-k$, so

$$|\delta_k(s,a)|\leqslant|\mu(s,a,k)-\tilde{\mu}(s,a,k)|+(H-k)\|p_k(\cdot\mid s,a)-\tilde{p}_k(\cdot\mid s,a)\|_1\leqslant(H-k+1)\,\gamma_{s,a,k},$$

where $\gamma_{s,a,k}\triangleq|\mu(s,a,k)-\tilde{\mu}(s,a,k)|+\|p_k(\cdot\mid s,a)-\tilde{p}_k(\cdot\mid s,a)\|_1$. Therefore,

$$\begin{aligned}\left|Q_h^\pi(s,a)-\tilde{Q}_h^\pi(s,a)\right| &\leqslant\sum_{k=h}^{H}\mathbb{E}_M^\pi[|\delta_k(s_k,a_k)|\,|\,s_h=s,\,a_h=a]\\ &\leqslant\sum_{s',a'}\sum_{k=h}^{H}(H-k+1)\gamma_{s',a',k}\\ &\leqslant(H-h+1)\,\|M-\widetilde{M}\|_1\\ &\leqslant(H-h+1)\,\gamma.\end{aligned}$$

□

The purpose of the following lemma is to show that if $\widetilde{M}\in\mathrm{Alt}(M)$ does not have enough enclosing volume around it, then it is a difficult instance, i.e., the sup-optimality gaps are small.

Lemma 42. *Let $M\in\mathbf{M}$ and let $\pi\in\Pi_{det}$ be its unique globally optimal policy. Denote by $\Delta^\pi(M)$ the smallest value gap in M for policy π, and let*

$$A^\pi\;\coloneqq\;\big\{M'\in\mathbf{M}:\pi\text{ is globally optimal in }M'\big\}.$$

If the ℓ_1-ball of radius $\frac{\delta}{2(H+1)}$ around M is not entirely contained in A^π, i.e.,

$$\left\{M'\in\mathbf{M}:\|M-M'\|_1\leqslant\frac{\delta}{2(H+1)}\right\}\;\not\subset\;A^\pi,$$

then the smallest gap of M must satisfy

$$\Delta^\pi(M) \leqslant \delta\,.$$

Proof. Fix $\delta, \delta' \geqslant 0$. Assume that $\Delta^\pi(M) > \delta'$, and let $M' \in \mathbf{M}$ satisfy $\|M - M'\|_1 \leqslant \delta$.

Denote by $\widetilde{Q}_h$ and $\widetilde{V}_h$ the state–action and state value functions in M', and by Q_h and V_h their counterparts in M.

Then, for any action $a \neq \pi_h(s)$, we have

$$\begin{aligned}
\widetilde{Q}_h^\pi(s,\pi_h(s)) - \widetilde{Q}_h^\pi(s,a) &= Q_h^\pi(s,\pi_h(s)) - Q_h^\pi(s,a) + \Big(\widetilde{V}_h^\pi(s) - V_h^\pi(s)\Big) + \Big(Q_h^\pi(s,a) - \widetilde{Q}_h^\pi(s,a)\Big) \\
&\geqslant \delta' + \Big(\widetilde{V}_h^\pi(s) - V_h^\pi(s)\Big) + \Big(Q_h^\pi(s,a) - \widetilde{Q}_h^\pi(s,a)\Big) \\
&\geqslant \delta' - \delta - |\widetilde{V}_h^\pi(s) - V_h^\pi(s)| - |\widetilde{V}_{h+1}^\pi(s) - V_{h+1}^\pi(s)| \\
&\geqslant \delta' - (2H - 2h + 2)\delta,
\end{aligned}$$

where the last inequality follows by value difference lemma as $\|M - M'\|_1 \leqslant \delta$. Thus, it suffices to take

$$\delta' \triangleq 2(H+1)\delta$$

to guarantee that π is the unique globally optimal policy in the MDP M'. Indeed, for any other policy ρ, by induction, we have at step h,

$$\begin{aligned}
V_h^\rho(s) &= \mu(s,\rho_h(s),h) + p_h V_{h+1}^\rho(s,\rho_h(s)) \\
&< \mu(s,\rho_h(s),h) + p_h V_{h+1}^\pi(s,\rho_h(s)) \\
&< \mu(s,\pi_h(s),h) + p_h V_{h+1}^\pi(s,\pi_h(s)) = V_h^\pi(s),
\end{aligned}$$

which concludes the proof.

□

The goal of the next lemma is to show that from an instance M, one can pick a policy $\pi \in \Pi^\star_{\mathrm{det}}(M)$ and build an easier instance M' (with larger value gaps) close to M and such that π is still optimal in M'. This will be used to show that if an instance M lies on the boundary of $\mathrm{Alt}(\widetilde{M})$ (so a difficult instance with nearly zero gap) of another instance $\widetilde{M}$, one can build an easier instance M' close to M, which is still in $\mathrm{Alt}(\widetilde{M})$, showing that there is enough volume around any point in the interior of the alternative set of an MDP.

Lemma 43. *Let $M \in \mathbf{M}$ and let $\pi \in \Pi_{\mathrm{det}}$ be a globally optimal policy for M. For any $\varepsilon \leqslant \frac{1}{3(1+2^{H-1})}$, there exists an MDP $\tilde{M} \triangleq (\tilde{\mu}, \tilde{p})$ and coefficients $(\gamma_{s,a,h})_{s,a,h} \in [0,1]$ such that for all (s,a,h),*

$$\tilde{p}_h(\cdot \mid s,a) \triangleq (1-\gamma_{s,a,h})p_h(\cdot \mid s,a) + \gamma_{s,a,h}\, p_h(\cdot \mid s,\pi_h(s)), \qquad |\mu(s,a,h) - \tilde{\mu}(s,a,h)| \leqslant d_{s,a,h} \triangleq 2^{H-h+1}\varepsilon, \tag{K.1}$$

with

$$\gamma_{s,a,h} \in \left[0,\ \frac{(1+2^{H-h})\varepsilon}{1-(2^{H-h+1}+3)\varepsilon}\right].$$

Moreover, π is strongly globally optimal in $\tilde{M}$: for all (s,h),

$$\tilde{Q}_h^\pi(s,\pi_h(s)) - \max_{a \neq \pi_h(s)} \tilde{Q}_h^\pi(s,a) \geqslant \varepsilon.$$

Proof. We construct $\tilde{M}$ bottom-up from stage H to stage 1. Throughout, Q, V (resp. $\tilde{Q}, \tilde{V}$) denote value functions in M (resp. $\tilde{M}$), and we maintain the induction hypothesis

$$0 \leqslant \tilde{V}_h^\pi(s) - V_h^\pi(s) \leqslant \Omega_h(\varepsilon) \triangleq (2^{H-h+1} - 1)\varepsilon, \qquad \forall\, s. \tag{K.2}$$

Base case $h = H$. If $\mu(s, \pi_H(s), H) - \mu(s, a, H) \leqslant \varepsilon$, set

$$\tilde{\mu}(s, a, H) \triangleq \max(0, \mu(s, a, H) - \varepsilon), \qquad \tilde{\mu}(s, \pi_H(s), H) \triangleq \min(1, \mu(s, \pi_H(s), H) + \varepsilon), \tag{K.3}$$

and $\gamma_{s,a,H} = 0$. This ensures $\tilde{Q}^\pi_H(s, \pi_H(s)) - \tilde{Q}^\pi_H(s, a) \geqslant \varepsilon$ and $0 \leqslant \tilde{V}^\pi_H(s) - V^\pi_H(s) \leqslant \varepsilon = \Omega_H(\varepsilon)$, so (K.2) holds at $h = H$.

Induction step. Assume (K.2) holds at stage $h + 1$. At stage h, for each (s, a) we consider three cases.

Case 1. If

$$\mu(s, \pi_h(s), h) - \mu(s, a, h) + p_h \tilde{V}^\pi_{h+1}(s, \pi_h(s)) - p_h \tilde{V}^\pi_{h+1}(s, a) \geqslant \varepsilon, \tag{K.4}$$

set $\tilde{p}_h(\cdot \mid s, a) \triangleq p_h(\cdot \mid s, a)$ and $\tilde{\mu}(s, a, h) \triangleq \mu(s, a, h)$, i.e., $\gamma_{s,a,h} = 0$. The gap is already at least ε without any perturbation.

Case 2. If (K.4) does not hold but

$$\tilde{\mu}(s, \pi_h(s), h) - \tilde{\mu}(s, a, h) \geqslant \mu(s, \pi_h(s), h) - \mu(s, a, h) + \Omega_{h+1}(\varepsilon) + \varepsilon, \tag{K.5}$$

set $\gamma_{s,a,h} = 0$. Shifting rewards alone is sufficient to guarantee the gap.

Case 3. Otherwise, neither (K.4) nor (K.5) holds. Since (K.5) does not hold, we have

$$\mu(s, \pi_h(s), h) + \Omega_{h+1}(\varepsilon) + \varepsilon \geqslant 1 \quad \text{and} \quad \mu(s, a, h) - \Omega_{h+1}(\varepsilon) - \varepsilon \leqslant 0, \tag{K.6}$$

Indeed, if any of the statements above above didn't hold, we would have Eq.(K.5). This further implies

$$\mu_h(s, \pi_h(s)) - \mu_h(s, a) \geqslant 1 - 2\left(\Omega_{h+1}(\varepsilon) + \varepsilon\right), \tag{K.7}$$

and since (K.4) does not hold either, we have

$$p_h \tilde{V}^\pi_{h+1}(s, \pi_h(s)) - p_h \tilde{V}^\pi_{h+1}(s, a) \leqslant \varepsilon + 2(\Omega_{h+1}(\varepsilon) + \varepsilon) - 1. \tag{K.8}$$

By the mixture definition of $\tilde{p}_h$ and the induction hypothesis,

$$\begin{aligned}
\tilde{Q}^\pi_h(s, \pi_h(s)) - \tilde{Q}^\pi_h(s, a) &\geqslant Q^\pi_h(s, \pi_h(s)) - Q^\pi_h(s, a) - \Omega_{h+1}(\varepsilon) \\
&\quad + \gamma\big(p_h \tilde{V}^\pi_{h+1}(s, a) - p_h \tilde{V}^\pi_{h+1}(s, \pi_h(s))\big) \\
&\geqslant -\Omega_{h+1}(\varepsilon) + \gamma\big(1 - \varepsilon - 2(\Omega_{h+1}(\varepsilon) + \varepsilon)\big),
\end{aligned}$$

where the first inequality uses $\pi \in \Pi^\star_{\mathrm{det}}(M)$ (so $Q^\pi_h(s, \pi_h(s)) - Q^\pi_h(s, a) \geqslant 0$) and (K.2), and the second uses (K.8). Setting

$$\gamma_{s,a,h} \triangleq \frac{\varepsilon + \Omega_{h+1}(\varepsilon)}{1 - \varepsilon - 2(\Omega_{h+1}(\varepsilon) + \varepsilon)} = \frac{(1 + 2^{H-h})\varepsilon}{1 - (2^{H-h+1} + 3)\varepsilon} \tag{K.9}$$

gives $\tilde{Q}^\pi_h(s, \pi_h(s)) - \tilde{Q}^\pi_h(s, a) \geqslant \varepsilon$. The condition $\gamma_{s,a,h} \in [0, 1]$ requires $1 - (2^{H-h+1} + 3)\varepsilon \geqslant (1 + 2^{H-h})\varepsilon$, which holds whenever $\varepsilon \leqslant \frac{1}{3+2^{H-h}+2^{H-h+1}}$. This is satisfied for all $h \in [H]$ if $\varepsilon \leqslant \frac{1}{3(1+2^{H-1})}$.

Verifying the induction hypothesis at stage h. Since $\tilde{p}_h(\cdot \mid s, \pi_h(s)) = p_h(\cdot \mid s, \pi_h(s))$ and $\tilde{\mu}(s, \pi_h(s), h) \leqslant \mu(s, \pi_h(s), h) + \varepsilon$,

$$\begin{aligned}
\tilde{V}^\pi_h(s) &= \tilde{\mu}(s, \pi_h(s), h) + p_h \tilde{V}^\pi_{h+1}(s, \pi_h(s)) \\
&\leqslant \mu(s, \pi_h(s),) + \varepsilon + \Omega_{h+1}(\varepsilon) + p_h V^\pi_{h+1}(s, \pi_h(s)) \\
&= V^\pi_h(s) + \underbrace{2\Omega_{h+1}(\varepsilon) + \varepsilon}_{= \Omega_h(\varepsilon)},
\end{aligned}$$

and, it is simple to verify that $\tilde{V}^\pi_h(s) \geqslant V^\pi_h(s)$ since both rewards and transitions are translated in favour of π. This confirms (K.2) at stage h, with the closed form $\Omega_h(\varepsilon) = (2^{H-h+1} - 1)\varepsilon$.

From the update rule (K.3) and the definition of Ω_{h+1},

$$|\tilde{\mu}(s, a, h) - \mu(s, a, h)| \leqslant \Omega_{h+1}(\varepsilon) + \varepsilon = 2^{H-h+1}\varepsilon = d_{s,a,h},$$

which holds for all (s, a, h), completing the proof. □

Lemma 44. *Let an MDP $M \in \mathbf{M}$ with a unique stage-optimal policy $\pi \in \Pi_{det}$. If π is a global optimal policy in M and $\Delta^\pi(M) > 0$, then for any $\gamma \in \left(0, \frac{\Delta}{6SAH(H+1)}\right)$, π is a global optimal policy for any MDP in the set $(1-\gamma)M + \gamma\mathbf{M}$.*

Proof. Let $\pi \in \Pi_{\text{det}}$ be a global optimal policy in M. From Lemma 42 we have for $\varepsilon \leqslant \Delta$ and for any $M' \in \mathbb{B}_1(M, \frac{\varepsilon}{2(H+1)})$, π is an optimal policy in M'. Now it suffices to take γ such that $(1-\gamma)M+\gamma\mathbf{M} \subset \mathbb{B}_1(M, \gamma')$. For this, observe that for any $M' \in \mathbf{M}$ we have

$$\begin{aligned} \|M - ((1-\gamma)M + \gamma M')\|_1 &= \gamma \|M - M'\|_1 \\ &\leqslant 2\gamma SAH. \end{aligned}$$

Thus we have $(1-\gamma)M + \gamma\mathbf{M} \subset \mathbb{B}_1(M, 3\gamma SAH)$ and taking γ such that $3\gamma SAH \leqslant \frac{\Delta}{2(H+1)}$ and using Lemma 42 concludes the proof. □

The goal of the following lemma is to, given an MDP M, generate $M' \in \mathbf{M}$ such that the transition probabilities in M' are larger than some threshold while M and M' still have the same optimal policy.

Lemma 45. *Let an MDP $M \triangleq (\mu, p)$ with optimal policy π such that its minimum policy gap $\Delta^\pi(M) > 0$. Let u denote the uniform transition kernel and define for any $\gamma \in (0,1)$ the MDP $M_\gamma = (\mu, (1-\gamma)p + \gamma u)$. For $\gamma \in (0, \Delta_M(\pi)/(3H))$, π is state-wise striclty optimal in M_γ.*

Proof. Let γ be fixed and simplify notation by letting $\widetilde{M} = M_\gamma$, $\Delta_M(\pi) = \Delta$, and define $\widetilde{Q}, \widetilde{V}$ the Q-function and value function in $\widetilde{M}$. We have for any $a \neq \pi_h(s)$

$$\begin{aligned} \widetilde{Q}_h^\pi(s, \pi_h(s)) - \widetilde{Q}_h^\pi(s,a) &\geqslant \Delta + \widetilde{Q}_h^\pi(s, \pi_h(s)) - Q_h^\pi(s, \pi_h(s)) + Q_h^\pi(s,a) - \widetilde{Q}_h^\pi(s,a) \\ &= \Delta + \widetilde{V}_h^\pi(s) - V_h^\pi(s) + p_h V_{h+1}^\pi(s,a) - \tilde{p}_h \widetilde{V}_{h+1}^\pi(s,a) \\ &= \Delta + \widetilde{V}_h^\pi(s) - V_h^\pi(s) + p_h V_{h+1}^\pi(s,a) - p_h \widetilde{V}_{h+1}^\pi(s,a) + \gamma(p_h - u)\widetilde{V}_{h+1}^\pi(s,a) \\ &\geqslant \Delta - 3H\gamma \end{aligned}$$

so it suffices for γ to be smaller than $\Delta/(3H)$ to guarantee the claimed property. □

The next lemma is proven in Menard et al. 2021. We restate it for completeness.

Lemma 46. *For a sequence $(u_t)_{t\geqslant 1} \in (0,1)^{\mathbb{N}}$ and $U_t \triangleq \sum_{l=1}^t u_l$ we have*

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

L BEYOND EXACT POLICY IDENTIFICATION

In this section, we discuss the extension of our algorithm to the setting with (ε, δ)-PAC policy identification with $\varepsilon > 0$. In this case, we let $\Pi_\varepsilon^\star(M)$ be the set of (Markov, deterministic and stochastic) policies π such that

$$\max_{\pi'} V_0^{\pi'} \leqslant V_0^\pi + \varepsilon\,.$$

The goal of the agent is then to identify a policy $\pi^\star \in \Pi_\varepsilon^\star(M)$ with high probability, i.e., given an error rate δ, her final recommendation $\hat{\pi}^\tau$ at stopping time τ should satisfy

$$\Pr_M(\tau < \infty, \hat{\pi}_\tau \notin \Pi_\varepsilon^\star(M)) \leqslant \delta.$$

Al-Marjani et al. 2023b derived a lower bound for this setting, which we state below. To introduce this lower bound, we introduce the following notation. For a policy π and given $\varepsilon \geqslant 0$ let $\text{Alt}_\varepsilon^\pi(M) \triangleq \left\{M' : M \ll M' \text{ and } \pi \notin \Pi_\varepsilon^\star(M')\right\}$. In words, it is the set of M', for which $M \ll M'$ and where the π is not a near-optimal policy.

Theorem 47. *Fix an $M \in \mathbf{M}$. For any adaptive stopping time τ for (ε, δ)-PAC policy identification*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_M[\tau]}{\log \frac{1}{\delta}} \geqslant \Gamma_\varepsilon(M) \,,$$

where

$$\Gamma_\varepsilon^{-1}(M) \triangleq \max_{\pi \in \Pi_\varepsilon^\star(M)} \max_{w \in \Omega_M} \inf_{M' \in \mathrm{Alt}_\varepsilon^\pi(M)} \left[\sum_{s,a,h} w_h(s,a) \, \mathrm{KL}(M, M')_{s,a,h} \right] .$$

Intuitively, the lower bound states that the optimal sample complexity should scale with the easiest to identify policy among the ε-optimal set of policies. From previous works in bandits, asymptotically matching this bound often required "sticking procedures" to stick to an answer to verify. Indeed, for $\varepsilon = 0$ and $|\Pi_\varepsilon^\star(M)| = 1$, sticking at each episode, the algorithm's sampling rule is meant to collect evidence for assessing the true optimality of its current best policy guess $\hat{\pi}_t$. For $\varepsilon > 0$, it is likely that at episode t, $|\widehat{\Pi}_{t,\varepsilon}| > 1$. The challenge is then to pick which answer in this set the sampling rule should aim to verify.

Following the approach developed in Appendix H it's simple to prove that

Theorem 48. *Fix an MDP $M \in \mathbf{M}$ and let ν_t denote the posterior distribution introduced in Eq.(1). For any adaptive algorithm for best policy identification, with probability one,*

$$\limsup_{t \to \infty} -\frac{1}{t} \log \left(\min_{\pi \in \Pi_\varepsilon^\star(M)} \Pr_{\widetilde{M} \sim \nu_t | \mathcal{H}_{t-1}} \big(\pi \notin \Pi_\varepsilon^\star(\widetilde{M}) \big) \right) \leqslant \Gamma_\varepsilon(M).$$

To state the modification of the Algorithm 1, let a generic sticking procedure take as input the empirical $\widehat{M}$ and output a policy in $\widehat{\Pi}_{t,\varepsilon}$. We require the procedure to be deterministic and stable, that is if the sticking recommends an answer at time t, then as long as this answer belongs to $\widehat{\Pi}_{t',\varepsilon}$ for $t' \geqslant t$, it should keep recommending it. In case of ties, the sticking break rules with consistency, so that this defines a strict total order on $\Pi^\star(M)$.

Because of the smooth forced exploration, we should have that on a good event $\widehat{\Pi}_{t,\varepsilon} \to \Pi_\varepsilon^\star(\widetilde{M})$ so that the sticking rule is simply a rule to select an answer to collect samples to verify it. We restate below the slight modification of Algorithm 1 to account for the $\varepsilon > 0$ case. Remark that compared to the initial algorithm, only Line ?? has changed, to pick an answer to verify among the set of empirically admissible answers.

Intuitively, the sticking procedure transforms the problem into verifying a single answer, as opposed to verifying multiple correct answers.

As the algorithm sticks to an answer π, the conditional posterior sampling is now a no-regret learner on $\mathrm{Alt}_\varepsilon^\pi(M)$. Following the same lines as in previous sections, one could derive the no-regret property of the conditional posterior sampling on $\mathrm{Alt}_\varepsilon^\pi(M)$. Combining these observations, the following will hold with probability one

A posterior-based stopping time for the multiple correct answers setting. We propose the following stopping rule for the (ε, δ)-PAC policy identifcation with $\varepsilon > 0$.

$$\tau \triangleq \inf \left\{ t \geqslant 1 : \exists \pi \in \widehat{\Pi}_{t,\varepsilon} \text{ such that } \forall \, k \, \leqslant B(t,\delta), \pi \in \Pi_\varepsilon^\star(\widetilde{M}_{t,k}) \right\}. \tag{L.1}$$

In words, the stopping rule triggers when there exists an answer for which finding a challenger by conditional resampling takes more than $B(t,\delta)$ samples. Observe that for this stopping rule, we have

$$\begin{aligned}
\Pr(\tau > t \mid \mathcal{H}_{t-1}) &\leqslant \Pr(\forall \, \pi \in \widehat{\Pi}_{t,\varepsilon}, \exists \, k \leqslant B(t,\delta) : \pi \notin \Pi_\varepsilon^\star(\widetilde{M}_{t,k}) \mid \mathcal{H}_{t-1}) \\
&\leqslant \min_{\pi \in \widehat{\Pi}_{t,\varepsilon}} \Pr(\exists \, k \leqslant B(t,\delta) : \pi \notin \Pi_\varepsilon^\star(\widetilde{M}_{t,k}) \mid \mathcal{H}_{t-1}) \\
&\leqslant B(t,\delta) \min_{\pi \in \widehat{\Pi}_{t,\varepsilon}} \Pr(\widetilde{M}_t \in \mathrm{Alt}_\varepsilon^\pi(\hat{M}_t) \mid \mathcal{H}_{t-1}) \,,
\end{aligned}$$

where $\widetilde{M}_t, \widetilde{M}_{t,1}, \ldots, \widetilde{M}_{t,k}$ are *i.i.d* .

By the concentration of the posterior distribution, we would have with probability one,

$$\begin{aligned}\Pr(\tau > t \mid \mathcal{H}_{t-1}) &\lessapprox \widetilde{O}\left(B(t,\delta) \min_{\pi \in \widehat{\Pi}_{t,\varepsilon}} \exp\left\{ - \inf_{\widetilde{M} \in \mathrm{Alt}^{\pi}_{\varepsilon}(\hat{M}_t)} \left[\sum_{s,a,h} n_t(s,a,h) \, \mathrm{KL}(\hat{M}_t \| \widetilde{M})_{s,a,h} \right] \right\}\right) \\ &\lessapprox \widetilde{O}\left(B(t,\delta) \exp\left\{ - \max_{\pi \in \widehat{\Pi}_{t,\varepsilon}} \inf_{\widetilde{M} \in \mathrm{Alt}^{\pi}_{\varepsilon}(\hat{M}_t)} \left[\sum_{s,a,h} n_t(s,a,h) \, \mathrm{KL}(\hat{M}_t \| \widetilde{M})_{s,a,h} \right] \right\}\right)\end{aligned}$$

where $\tilde{\mathcal{O}}$ hides constant and terms that are at most polynomial in t. As the empirical estimates converge, the term in the exponent will be equivalent (up to logarithmic terms in t)

$$\max_{\pi \in \Pi^{\star}_{\varepsilon}(M)} \inf_{\widetilde{M} \in \mathrm{Alt}^{\pi}_{\varepsilon}(M)} \left[\sum_{s,a,h} n_t(s,a,h) \, \mathrm{KL}(M \| \widetilde{M})_{s,a,h} \right]$$

which is larger than evaluating it at any response computed by the sticking procedure. Combining the convergence of the algorithm with "sticking" with the stopping time τ, one can prove results similar to the previous section in terms of sample complexity. It is also possible to ensure that if the sampling rule guarantees saddle-point convergence, then plugging in this stopping rule will ensure optimality. We leave as an open question the design of a converging, computationally efficient sampling rule for the setting with $\varepsilon > 0$.

Given m posterior samples, checking if the stopping triggers require solving the feasibility problem

$$\max_{\pi \in \widehat{\Pi}_{t,\varepsilon}} \min_{k \leqslant m} (\widetilde{V}^{k,\pi}_{t,0} + \varepsilon - \widetilde{V}^{k,\star}_{t,0}) > 0$$

and solving this might ultimately require enumerating $\widehat{\Pi}_{t,\varepsilon}$, which is a major caveat. For the computation of the stopping rule.

M IMPLEMENTATION DETAILS AND ADDITIONAL EXPERIMENTS

This section contains implementation details and additional experimental results comparing PIPS to PSRL.

M.1 Implementation Details

In this section, we provide details on the implementation and experimental setup for reproducibility.

Experimental setup. For each environment shown in Figure 2 and Figure 8, rewards are modeled as Gaussian random variables with means specified in the figures. In the RiverSwim environment, the reward variance is set to 1/20, while in the MOCA instance it is set to 1/4. Although both environments are stationary, we did not include this information in our algorithm, which allows us to estimate the model and maintain distributions over non-stationary MDPs.

Online learner. For the policy improvement step of the online learner, we use AdaHedge, implementing the pseudo-code described in De Rooij et al. 2014, which automatically adjusts its learning rate at each round. We recall that the initialization of Algorithm 3 with a generic online learner is described and analyzed in Appendix D, with its pseudo-code in Algorithm 3.

In practice, we observe little difference in performance compared to the natural Hedge update. As argued in Appendix D, both algorithms ensure sublinear regret and thus preserve the theoretical guarantees of Algorithm 1.

Posterior inflation and thresholds calibration. The posterior is inflated using $\eta_t = 1/(t+1)^4$, which provides a good trade-off between empirical performance and runtime. Running without inflation (i.e., $\eta_t = 1$) tends to yield slower experiments without noticeable benefits. However, for simplicity, we recommend using $\eta_t \leftarrow 1$ with $B(t,\delta) \approx \frac{SAH}{\delta} \log(tSAH/\delta)$ which experimentally ensure δ-correctness.

Algorithm 5: ε-PIPS: Policy Identification via Posterior Sampling in Tabular MDPs

Require: Initial exploration $\pi_1^{\text{exp}} = \{\pi_1^{\text{exp}}(\cdot|s,h)\}_{s,h}$; initial forced exploration c_1; mixing parameter $\gamma \in (0,1)$; clipping parameter $\alpha \in (0, \frac{1}{2})$; posterior inflation $(\eta_t)_{t\geqslant 1}$; initial MDP $\hat{M}_0 \in \mathbf{M}$ (arbitrary); $\varepsilon \geqslant 0$

for *episode* $t = 1, 2, \ldots$ **do**

```
// Stick to an answer among current empirical ε-best guesses
```

Compute $\hat{\pi}_t \leftarrow \text{STICKING}(\hat{M}_t, \varepsilon)$

```
// Computing a challenger MDP
```

for $k = 1, 2, \ldots$ **do**

Sample candidate challenger $\widetilde{M}_{t,k} \sim \nu_t^{\eta_t}$

Compute the optimal Q-function $\widetilde{Q}_t^{k,\star}$ of $\widetilde{M}_{t,k}$ by backward induction

if $\hat{\pi}_t$ *is* ε*-suboptimal for* $\widetilde{Q}_t^{k,\star}$ **then**

Set challenger $\widetilde{M}_t \leftarrow \widetilde{M}_{t,k}$ and **break**

end

end

Compute c_t as in Algorithm 4

Draw $Z_t \sim \text{Bernoulli}(t^{-\gamma})$ and define $\forall (s,h) \in \mathcal{S} \times [H]$

$$\tilde{\pi}_t^{\text{exp}}(\cdot|s,h) \leftarrow \begin{cases} c_t(\cdot|s,h), & \text{if } Z_t = 1, \\ \pi_t^{\text{exp}}(\cdot|s,h), & \text{if } Z_t = 0, \end{cases}$$

```
// Rollout and data collection
```

Set $s_1^t \leftarrow s_{\text{init}}$

for $h = 1$ **to** H **do**

Sample action $a_h^t \sim \tilde{\pi}_t^{\text{exp}}(\cdot|s_h^t, h)$

Observe reward r_h^t and transition to s_{h+1}^t

end

```
// KL-based bonuses for the online learner
```

Construct bonus kernel $r_t^{\text{exp}}(s,a,h)$ from $\hat{M}_t$ and $\widetilde{M}_t$ as in Eq.(3)

```
// Behaviour policy improvement
```

Update main exploration policy $\pi_{t+1}^{\text{exp}} \leftarrow \text{ComputeExplorationPolicy}\big(t, \pi_t^{\text{exp}}, r_t^{\text{exp}}, \hat{p}_t, n_t\big)$

end

Other algorithms implementation. For each algorithm, we refer to the pseudo-code in the original paper for the implementation. For the UCBVI algorithm (Azar, Osband, et al. 2017) and SSR (Xiong et al. 2022) we implement the following idealized Bernstein confidence bonuses

$$\text{bonus}_t(s,a,h) = \min\left(\sqrt{\frac{\text{Var}_{\hat{p}_t}[\bar{V}_{t,h+1}](s,a)\log(t+1)}{\max(1, n_t(s,a,h))}} + (H-h+1)\frac{\log(t+1)}{\max\left(1, n_t(s,a,h)\right)}, H-h+1\right).$$

For PSRL, the prior distribution is the same as for Algorithm 1.

M.2 Additional Experiments

We present further experimental results on the MOCA instance. Specifically, we study the number of posterior samples required before identifying an alternative model. This experiment evaluates the behavior of the posterior stopping time. Since both PIPS and PSRL rely on posterior sampling, we compare their performance in this setting.

Each algorithm is run on the MOCA instance for $T = 10{,}000$ episodes. At every episode, we record the number of posterior resamples needed to obtain an alternative model. For PSRL, this quantity is immaterial, but for

PIPS it directly influences the sampling rule. We report the average (over 50 independent runs).

The results in Figure 5 show that for PIPS, the number of resamples grows exponentially with t, as further confirmed by the log-plot in Figure 6. This behavior is consistent with theory: conditionally on the history, the number of resamples follows a geometric law with parameter scaling with $e^{-t\Gamma}$, so its expectation grows exponentially. This indicates that evidence accumulates rapidly, enabling the algorithm to certify the optimal policy.

In contrast, PSRL exhibits only polynomial growth in the number of resamples (also evident from the log-log plot in Figure 7). This suggests that its posterior concentrates at a polynomial rate.

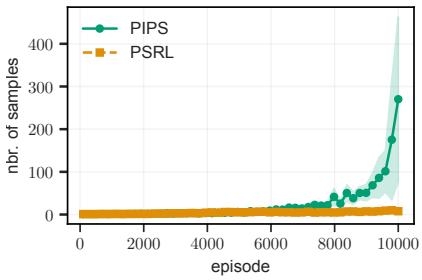

Figure 5: Number of samples for finding an alternative model averaged on 100 runs.

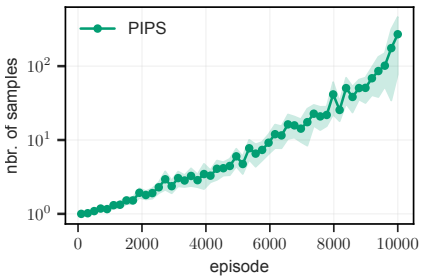

Figure 6: Log number of samples for finding an alternative model averaged on 100 runs.

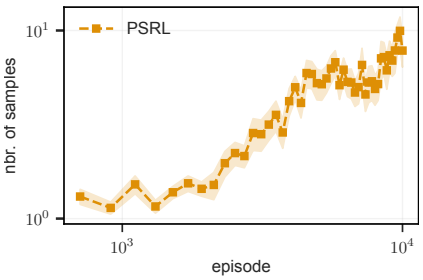

Figure 7: Log-log plot of number of samples for finding an alternative model averaged on 100 runs.

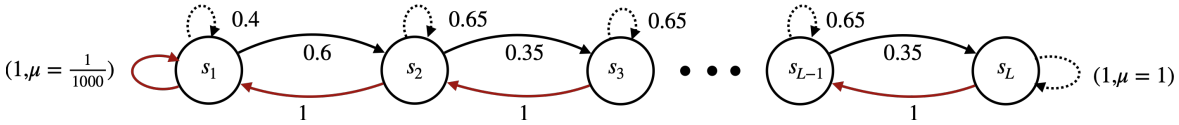

Figure 8: The L-state RiverSwim environment. Each state allows two actions: LEFT (red arrows) and RIGHT (black arrows). Rewards are placed at the two ends of the chain. Transition probabilities are annotated above the corresponding arrows.