Improved Bayesian Regret Bounds for Thompson Sampling in Reinforcement Learning

Ahmadreza Moradipari[*][†]   **Mohammad Pedramfar**[*][‡]   **Modjtaba Shokrian Zini**[*][§]

Vaneet Aggarwal[¶]

Abstract

In this paper, we prove the first Bayesian regret bounds for Thompson Sampling in reinforcement learning in a multitude of settings. We simplify the learning problem using a discrete set of surrogate environments, and present a refined analysis of the information ratio using posterior consistency. This leads to an upper bound of order $\widetilde{O}(H\sqrt{d_{l_1}T})$ in the time inhomogeneous reinforcement learning problem where H is the episode length and d_{l_1} is the Kolmogorov l_1-dimension of the space of environments. We then find concrete bounds of d_{l_1} in a variety of settings, such as tabular, linear and finite mixtures, and discuss how how our results are either the first of their kind or improve the state-of-the-art.

1 Introduction

Reinforcement Learning (RL) is a sequential decision-making problem in which an agent interacts with an unknown environment typically modeled as a Markov Decision Process (MDP) [38, 8]. The goal of the agent is to maximize its expected cumulative reward. This problem has a variety of applications, including robotics, game playing, resource management, and medical treatments. The key challenge in RL is to balance the so-called exploration-exploitation trade-off efficiently: exploring unseen state-action pairs to gain more knowledge about the unknown environment or exploiting the current knowledge to maximize the expected cumulative reward. Two efficient approaches have been developed to control this trade-off: *optimism in the face of uncertainty* (OFU) and *Thompson Sampling* (TS) (or Posterior Sampling (PS)). OFU constructs a confidence set of statistically plausible MDPs that includes the true MDP with high probability and plays an optimistic policy according to the MDP with maximum gain from this set [5, 40]. TS samples a statistically plausible MDP from a posterior distribution and plays the optimistic policy of the sampled MDP [29, 31]. In this work, we focus on the latter, and by combining an information theoretical approach first introduced by [36] with analysis based on posterior consistency tools, we prove state-of-the-art Bayesian regret bounds in a variety of settings.

In this paper, we start by defining the Bayesian RL problem, where transition and reward functions are Bayesian and time inhomogeneous. The Bayesian RL problem we consider is more comprehensive than in previous works, as we allow for both Bayesian transition and Bayesian rewards, and do not make any assumption on their individual prior. To simplify the learning problem, we utilize the notion of surrogate environments, which is a discretization of the environments space, and its learning task and TS regret is a proxy to that of the main problem. The construction of the surrogate environments was first introduced by [18] with an incorrect proof, which is fixed in our work by

---

[*] Authors have equal contribution.

[†] Toyota Motor North America, InfoTech Labs, Mountain View, CA, USA, `ahmadreza.moradipari@toyota.com`

[‡] Purdue University, West Lafayette, IN, USA, `mpedramf@purdue.edu`

[§] `modjtaba.shokrianzini@gmail.com`

[¶] Purdue University, West Lafayette, IN, USA, `vaneet@purdue.edu`

defining the surrogate environments through an optimization. Of main importance is the size of this new environment space. The Bayesian regret decomposes to the product of two terms, one being the cumulative mutual information of the environment and history traversed by the policy. By the well-known entropy estimation of the mutual information, this significant factor in the regret is connected to the l_1-dimensions (d_{l_1}) of the transition and reward functions space, which can be more succinctly interpreted as the l_1-dimension d_{l_1} of the environment space. The latter is in turn estimated by the size of the space of surrogate environments.

The information ratio, representing a trade-off of exploration/exploitation, is the other significant term in the decomposition of the TS Bayesian regret. In an improvement to [18], our novel analysis of this ratio based on posterior consistency tools, shows that this trade-off is bounded by $H^{3/2}$, where H is the episode length. This bound is general and independent of the dimension of transition/reward function space at each step, which is is a key factor behind the advantage of our regret bound, such as the $\sqrt{SA}$ advantage in the tabular case compared to [18], or the lack of any restriction on the prior (e.g., Dirichlet prior) compared to [31]. Following a further refined approach, we finally estimate the TS Bayesian regret to be $\widetilde{O}(\lambda\sqrt{d_{l_1}T})$ for large enough T in the time inhomogeneous setting. Here, a new term 'value diameter' λ, which is the average difference of the optimal value functions at different states, is used in bounding the information ratio, where instead of $H^{3/2}$, we have the smaller term $\lambda H^{1/2}$. Bounding the information ratio with λ is a conceptual contribution of our work, which shows that the ratio is bounded by a *value-dependent* term, which is in nature different from H but always $\leq H + 1$. Further, there exists another bound for λ; in environments where states are reachable from one another in D steps, we have $\lambda \leq D + 1$. In 'well-connected' MDPs, one could have $D \ll H$, implying an improvement over the $H^{3/2}$ information ratio bound.

Our generic bound is abstract in terms of d_{l_1}, so we estimate it in more explicit terms for useful applications. [18] have bounded d_{l_1} in the tabular and linear case without formalizing this notion, and while for tabular MDPs, d_{l_1} was bounded by SAH, for linear MDPs with feature space dimension d_f, we investigate their claim of the bound $d_f H$. Detailed in Appendix G, we show a counterexample to their analysis, and we manage to find a correct estimate in this setting. We also introduce finite mixtures MDPs and are the first to prove a TS Bayesian regret of order $\widetilde{O}(\lambda\sqrt{HmT})$, where m is the number of mixtures.

Lastly, we note that our regret bound of order $\widetilde{O}(\lambda\sqrt{d_{l_1}T})$ is the first in the general nonlinear time inhomogeneous Bayesian RL setting for TS, and generalizing [31, Conj. 1], we conjecture it to be optimal if λ can be replaced by $\widetilde{O}(\sqrt{H})$.

Related work. Since the introduction of information ratio by [35, 36], a new line of research has emerged to provide tighter regret bounds for TS. The general approach involves factoring the Bayesian regret into two components: an information ratio that captures the trade-off between optimal action selection and information gain, and a cumulative information gain term that depends on the target environment and the history of previous observations. Then, both components are bounded separately using information theoretic tools.

In the bandit setting, this analysis has been used to bound Bayesian regret for TS [14, 9], as well as that of a new algorithm called information-directed sampling (IDS) [35, 26, 23, 19, 20]. This analysis has also been used in partial monitoring [25, 24] and RL with a specific Dirichlet prior and additional assumptions [28, 27] or when the true environment is too complicated to learn [4]. More recently, [18] studied the Bayesian regret of TS in RL without any prior assumptions for tabular MDP. This is the closest work to our paper and we discuss our generalization in detail in Section 5.

The Bayesian tabular MDP case has also been studied with the additional Dirichlet prior assumption in [31], where they achieve a regret bound matching ours. In an independent approach, the first non-linear Bayesian RL model was considered by [16] with a regret bound of $dH^{3/2}T^{1/2}$ where d is a notion of dimension of their model, but their results were limited to Gaussian process settings with linear kernels. Finally, [11] considered general non-linear Bayesian RL models and introduced an algorithm that obtains $dH^{1+\alpha/2}T^{1-\alpha/2}$ where α is a tuning parameter and d is the dimension of $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

It is worth noting that there is another line of work that incorporates confidence regions into TS to achieve Bayesian regret bounds that can match the best possible frequentist regret bounds by

UCB in both bandit settings [35] and RL [31, 30, 32, 12]. However, this technique often results in a sub-optimal Bayesian regret, as the best bound known for UCB itself is not optimal.

Table 1: Bayesian regret bounds for TS (i.e. PSRL)

Reference	Tabular	Linear	General	Comments
[29]	$\sqrt{H^3 S^2 A L}$	-	-	-
[30]	-	-	$L^*\sqrt{d_K d_E H L}$	Uses Eluder dimension Lipschitz assumption
[31]	$\sqrt{H^3 S A L}$	-	-	Dirichlet prior
[28]	$\sqrt{H^3 S A L}$	-	-	Assumptions on prior
[12]	$L^*\sqrt{H^3 S^2 A^2 L}$	-	$L^*\gamma\sqrt{HL}$	Assumptions on regularity & noise Lipschitz assumption
[18]	$\sqrt{H^4 S^2 A^2 L}$	-	-	-
This paper	$\lambda\sqrt{H^2 S A L}$	$\lambda\sqrt{d_{l_1} H L}$	$\lambda\sqrt{d_{l_1} H L}$	Assumptions 1 & 2 Holds in the limit $L \to \infty$

As discussed in Section 4.3 of [16], the Lipschitz term L^*, which is used in the grayed papers in the table, may grow exponentially in episode length. Note that [18] claims a regret bound for the linear setting. However, as discussed in Appendix G.1, their proof is incorrect.

While our work's emphasis is on theoretical guarantees for TS, we discuss here the experiments using this algorithm. Previous works on PSRL [35, 26, 23, 20, 31] come with extensive experiments on TS (and/or its variants), and discussions on computational efficiency of PSRL. In particular, experiments in [31] support the assertion that "PSRL dramatically outperforms existing algorithms based on OFU". In addition, PSRL with oracle access has been shown to be the most performant, esp. when compared to recent OFU based UCBVI/UCBVI-B, or even variants of PSRL such as Optimistic PSRL [39, Fig. 1.3]. However, an important limitation in experiments is the need for oracle access to an optimal policy, and that can not be always satisfied efficiently. Nevertheless, clever engineering can make TS work even in large scale Deep RL. Indeed, for general RL settings, the recent work [37] shows how to implement TS in Deep RL on the Atari benchmark and concludes that "Posterior Sampling Deep RL (PSDRL) significantly outperforms previous state-of-the-art randomized value function approaches, its natural model-free counterparts, while being competitive with a state-of-the-art (model-based) reinforcement learning method in both sample efficiency and computational efficiency". In summary, experiments in the literature provide enough support for the empirical performance of TS.

2   Preliminaries

2.1   Finite-horizon MDP

We follow the literature's conventions in our notation and terminology to avoid confusion when comparing results. The environment is a tuple $\mathcal{E} = (\mathcal{S}, \mu_{\mathcal{S}}, \mathcal{A}, \mu_{\mathcal{A}}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$, where $\mathcal{S}$ is the topological measurable state space, $\mathcal{A}$ is the topological measurable action space, $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{A}}$ are base probability measures on $\mathcal{S}$ and $\mathcal{A}$ respectively, H is the episode length, $P_h : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}, \mu_{\mathcal{S}}}$ is the transition probability kernel, and $r_h : \mathcal{S} \times \mathcal{A} \to \Delta_{[0,1], \text{Lebesgue}}$ is the reward function, where we fix the convention $r(s, a) := \mathbb{E}_x[r(x|s, a)] = \int_0^1 x r(x|s, a)\, \mathrm{d}x$ as we mostly deal with its mean value. Notice that $\Delta_{X, \mu}$ is the set of probability distributions over X that are absolutely continuous with respect to μ. We will use Δ_X when the base measure is clear from the context. We assume $\mathcal{S}$, $\mathcal{A}$ are known and deterministic while the transition probability kernel and reward are unknown and random. Throughout the paper, the implicit dependence of P_h and r_h on $\mathcal{E}$ should be clear from the context.

Let Θ_h^P be the topological function space of P_h and $\Theta^P = \Theta_1^P \times \cdots \times \Theta_H^P$ be the full function space. The space Θ_h^P is assumed to be separable and equipped with prior probability measure ρ_h^P yielding the product prior probability measure $\rho^P = \rho_1^P \otimes \cdots \otimes \rho_H^P$ for Θ^P. The exact same definition with similar notations $\Theta_h^R, \rho_h^R, \rho^R, \Theta^R$ applies for the reward function. Notice the explicit assumption of time inhomogeneity in these definitions, with all 'layers' h being independent. The two sets define the set of all environments parametrized by $\Theta = \Theta_1 \times \cdots \times \Theta_H$ where $\Theta_h = \Theta_h^P \times \Theta_h^R$. Note that the prior is assumed to be known to the learner. This setting implies that an environment $\mathcal{E}$ sampled according to the prior $\rho = \rho^P \otimes \rho^R$ is essentially determined by its transition and reward functions pair $\{(P_h, r_h)\}_{h=1}^H$. We simplify the notation to view Θ as the set of all environments, i.e., saying

3

$\mathcal{E} \in \Theta$ should be viewed as $\{(P_h, r_h)\}_{h=1}^H \in \Theta$. The space of all possible real-valued functions $\{(P_h, r_h)\}_{h=1}^H$ has a natural vector space structure. Therefore it is meaningful to discuss the notion of the convex combination of environments. We assume that Θ is a convex subspace of the space of all possible environments. This assumption is not restrictive, since we may replace any environment space with its convex hull. Note that we do not assume that the support of the prior is convex.

Remark 1. The case of joint prior may be of interest, but to our knowledge all prior works also take ρ^P, ρ^R to be independent.

Agent, policy and history. An agent starts at an initial state s_1^ℓ, which is fixed for all episodes ℓ. It observes a state s_h^ℓ at layer h episode ℓ, takes action a_h^ℓ, and receives reward r_h^ℓ. The environment changes to the next random state s_{h+1}^ℓ with probability $P_h(s_{h+1}^\ell | s_h^\ell, a_h^\ell)$. The agent stops acting at s_{H+1} and the environment is reset to its initial state.

We define $\mathcal{H}_{\ell,h}$ as the history $(s_1^\ell, a_1^\ell, r_1^\ell, \ldots, s_h^\ell, a_h^\ell, r_h^\ell)$. Denote by $\mathcal{D}_\ell = (\mathcal{H}_{1,H}, \ldots, \mathcal{H}_{\ell-1,H})$ the history up to episode ℓ, where $\mathcal{D}_1 := \emptyset$. Finally, let $\Omega_h = \prod_{i=1}^h (\mathcal{S} \times \mathcal{A} \times [0,1])$ be the set of all possible histories up to layer h.

A policy π is represented by stochastic maps $(\pi_1, \ldots, \pi_H)$ where each $\pi_h : \Omega_{h-1} \times \mathcal{S} \to \Delta_{\mathcal{A}, \mu_\mathcal{A}}$. Let Π_S denote the entire stationary policy class, stationary meaning a dependence only on the current state and layer and let $\Pi \subseteq \Pi_S$.

Value and state occupancy functions. Define the value function $V_{h,\pi}^\mathcal{E}$ as the value of the policy π interacting with $\mathcal{E}$ at layer h:

$$V_{h,\pi}^\mathcal{E}(s) := \mathbb{E}_\pi^\mathcal{E} \left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s \right], \tag{1}$$

where $\mathbb{E}_\pi^\mathcal{E}$ denotes the expectation over the trajectory under policy, transition, and reward functions π, P_h, r_h. The value function at step $H+1$ is set to null, $V_{H+1,\pi}^\mathcal{E}(\cdot) := 0$. We assume there is a measurable function $\pi_\mathcal{E}^* : \Theta \to \Pi$ such that $V_{h,\pi_\mathcal{E}^*}^\mathcal{E}(s) = \max_{\pi \in \Pi} V_{h,\pi}^\mathcal{E}(s)$, $\forall s \in \mathcal{S}, h \in [H]$. The optimal policy π^* is a function of $\mathcal{E}$, making it a random variable in the Bayesian setting. Lastly, let the *state-action occupancy probability measure* be $\mathbb{P}_\pi^\mathcal{E}(s_h = s, a_h = a)$, also known as the state occupancy measure under policy π and environment $\mathcal{E}$. It follows from the definitions that this measure is absolutely continuous with respect to $\mu_{\mathcal{S} \times \mathcal{A}} := \mu_\mathcal{S} \times \mu_\mathcal{A}$. Let $d_{h,\pi}^\mathcal{E}(s, a)$ denote the Radon–Nikodym derivative so that we have $d_{h,\pi}^\mathcal{E}(s,a) \, d\mu_{\mathcal{S} \times \mathcal{A}} = d\mathbb{P}_\pi^\mathcal{E}(s_h = s, a_h = a)$. We will assume throughout the paper that this density $d_{h,\pi}^\mathcal{E}(s,a)$ is measurable and upper bounded for all $\pi, \mathcal{E}, s, a, h$. The upper bound is a reasonable assumption, and it happens trivially in the tabular case ($d_{h,\pi}^\mathcal{E}(s,a) \leq SA$). This also happens, e.g., when one assumes that the maps $(\mathcal{E}, s, a, s', h) \mapsto P_h^\mathcal{E}(s'|s,a)$ and $(\pi, s, a, h) \mapsto \pi_h(a|s)$ are continuous and $\Theta, \mathcal{S}, \mathcal{A}$ and the set of all optimal policies (as a subset of Π) are compact.

2.2 Bayesian regret

We formulate the expected regret over L episodes and $T = LH$ total steps in an environment $\mathcal{E}$ as

$$\mathfrak{R}_L(\mathcal{E}, \pi) = \mathbb{E} \left[\sum_{\ell=1}^L \left(V_{1,\pi_\mathcal{E}^*}^\mathcal{E}(s_1^\ell) - V_{1,\pi^\ell}^\mathcal{E}(s_1^\ell) \right) \right], \tag{2}$$

where the expectation is over the randomness of $\pi = \{\pi^\ell\}_\ell$. The Bayesian regret is $\mathfrak{BR}_L(\pi) = \mathbb{E}[\mathfrak{R}_L(\mathcal{E}, \pi)]$. For Thompson Sampling (TS), the algorithm selects the optimal policy of a given sample $\mathcal{E}_\ell$ picked from the posterior $\mathcal{E}_\ell \sim \mathbb{P}(\mathcal{E} \in \cdot | \mathcal{D}_\ell)$:

$$\pi_{\text{TS}}^\ell = \text{argmax}_{\pi \in \Pi} V_{1,\pi}^{\mathcal{E}_\ell}(s_1^\ell). \tag{3}$$

Importantly, the law of TS aligns with the posterior, i.e., $\mathbb{P}(\mathcal{E}|\mathcal{D}_\ell) = \mathbb{P}(\pi_{\text{TS}}^\ell = \pi_\mathcal{E}^* | \mathcal{D}_\ell)$.

Remark 2. Note that $\mathbb{P}(\pi_{\text{TS}}^\ell = \pi_\mathcal{E}^* | \mathcal{D}_\ell)$ is a probability for a specific measure on the space of optimal policies. To ensure that $\int_{\Pi^*} \mathbb{P}(\pi^* | \mathcal{D}_\ell) d\rho_{\Pi^*} = 1$, we need an appropriate measure ρ_{Π^*} on Π^*. Given the law of TS, the natural choice for this measure is the push-forward of the prior measure ρ under the map $star : \Theta \to \Pi^*$, where $star(\mathcal{E}) = \pi_\mathcal{E}^*$.

4

2.3 Notations

For Bayesian RL, conditional expressions involving a given history $\mathcal{D}_\ell$ are widely used. We adopt the notation in [18] to refer to such conditionals; let $\mathbb{P}_\ell(\cdot) := \mathbb{P}(\cdot|\mathcal{D}_\ell)$, $\mathbb{E}_\ell[\cdot] := \mathbb{E}[\cdot|\mathcal{D}_\ell]$. We can rewrite the Bayesian regret as

$$\mathfrak{BR}_L(\pi) = \sum_{\ell=1}^{L} \mathbb{E}\left[\mathbb{E}_\ell\left[V_{1,\pi_\mathcal{E}^*}^\mathcal{E}(s_1^\ell) - V_{1,\pi}^\mathcal{E}(s_1^\ell)\right]\right] \tag{4}$$

and define the conditional mutual information $\mathbb{I}_\ell(X;Y) := D_{\mathrm{KL}}(\mathbb{P}((X,Y) \in \cdot|\mathcal{D}_\ell)||\mathbb{P}(X \in \cdot|\mathcal{D}_\ell) \otimes \mathbb{P}(Y \in \cdot|\mathcal{D}_\ell))$. For a random variable χ and random policy π, the following will be involved in the information ratio:

$$\mathbb{I}_\ell^\pi(\chi;\mathcal{H}_{\ell,h}) := \mathbb{I}_\ell(\chi;\mathcal{H}_{\ell,h}|\pi) = \mathbb{E}_\pi[D_{\mathrm{KL}}(\mathbb{P}_\ell((\chi,\mathcal{H}_{\ell,h}) \in \cdot|\pi)||\mathbb{P}_\ell(\chi \in \cdot|\pi) \otimes \mathbb{P}_\ell(\mathcal{H}_{\ell,h} \in \cdot|\pi))], \tag{5}$$

Note that $\mathbb{E}[\mathbb{I}_\ell(X;Y)] = \mathbb{I}(X;Y|\mathcal{D}_\ell)$. To clarify, $\mathbb{P}_\ell(\mathcal{H}_{\ell,h} \in \cdot|\pi)$ is the probability of $\mathcal{H}_{\ell,h}$ being generated under π within some environment. Given that the histories under consideration are generated by the TS algorithm, they are always generated in the true environment $\mathcal{E}$ under an optimal policy $\pi_{\mathcal{E}'}^*$. For $\pi = \pi_{\mathrm{TS}}^\ell$, this can be computed as $\mathbb{P}_\ell(\mathcal{H}_{\ell,h}|\pi) = \int_\mathcal{E} P(\mathcal{H}_{\ell,h}|\pi,\mathcal{E})\,d\mathbb{P}_\ell(\mathcal{E})$, where $P(\mathcal{H}_{\ell,h}|\pi,\mathcal{E})$ is an expression in terms of transition and reward functions of $\mathcal{E}$ and π.

Finally, we define $\bar{\mathcal{E}}_\ell$ as the mean MDP where $P_h^{\bar{\mathcal{E}}_\ell}(\cdot|s,a) = \mathbb{E}_\ell[P_h^\mathcal{E}(\cdot|s,a)]$ is the mean of posterior measure, and similarly for $r_h^{\bar{\mathcal{E}}_\ell}(\cdot|s,a) = \mathbb{E}_\ell[r_h^\mathcal{E}(\cdot|s,a)]$. We note that under the independence assumption across layers, the same is given for the state-occupancy density $d_{h,\pi}^{\bar{\mathcal{E}}_\ell} = \mathbb{E}_\ell[d_{h,\pi}^\mathcal{E}]$.

3 Bayesian RL problems

Definition 1. A Bayesian RL in this paper refers to the time-inhomogeneous finite-horizon MDP with independent priors on transition and reward functions, as described in Section 2.1.

The Bayesian RL *problem* is the task of finding an algorithm π with optimal Bayesian regret as defined in Eq. (4). Below we list the variations of this problem. A setting considered by most related works such as [31, 16] is the following:

Definition 2. The **time (reward) homogeneous** Bayesian RL refers to the Bayesian RL setting where the prior ρ^P (ρ^R) is over the space Θ^P (Θ^R) containing the single transition (reward) function P (r) defining $\mathcal{E}$, i.e., all layers have the same transition (reward) functions.

Definition 3. The **tabular** Bayesian RL is a Bayesian RL where $\mathcal{S}, \mathcal{A}$ are finite sets.

Definition 4 (Linear MDP [41, 22]). Let $\phi^P : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_f^P}, \phi^R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_f^R}$ be feature maps with bounded norm $\|\phi^P(s,a)\|_2, \|\phi^R(s,a)\|_2 \leq 1$. The **linear** Bayesian RL is a Bayesian RL where for any $\mathcal{E} = \{(P_h^\mathcal{E}, r_h^\mathcal{E})\}_{h=1}^H \in \Theta$, there exists vector-valued maps $\psi_h^{P,\mathcal{E}}(s), \psi_h^{R,\mathcal{E}}(s)$ with bounded l_2-norm such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$P_h^\mathcal{E}(\cdot|s,a) = \langle \phi^P(s,a), \psi_h^{P,\mathcal{E}}(\cdot) \rangle, \ \ r_h^\mathcal{E}(\cdot|s,a) = \langle \phi^R(s,a), \psi_h^{R,\mathcal{E}}(\cdot) \rangle \tag{6}$$

A restricted version of the finite mixtures called linear mixture was first considered in [6] in the frequentist setting. Here, we consider the general setting.

Definition 5. The **finite mixtures** Bayesian RL is a Bayesian RL where for any $h \in [H]$ there exists fixed conditional distributions $\{Z_{h,i}^P : \mathcal{S} \times \mathcal{A} \to \Delta_\mathcal{S}\}_{i=1}^{m_h^P}$ and $\{Z_{h,i}^R : \mathcal{S} \times \mathcal{A} \to \Delta_{[0,1]}\}_{i=1}^{m_h^R}$, such that for any environment $\mathcal{E}$ given by $\{(P_h^\mathcal{E}, r_h^\mathcal{E})\}_{h=1}^H$, there exists parametrized probability distributions $a_h^{P,\mathcal{E}} : \mathcal{S} \times \mathcal{A} \to \Delta_{m_h^P}, a_h^{R,\mathcal{E}} : \mathcal{S} \times \mathcal{A} \to \Delta_{m_h^R}$ such that

$$P_h^\mathcal{E}(\cdot|s,a) = \sum_{i=1}^{m_h^P} a_{h,i}^{P,\mathcal{E}}(s,a)Z_{h,i}^P(\cdot|s,a), \ \ r_h^\mathcal{E}(\cdot|s,a) = \sum_{i=1}^{m_h^R} a_{h,i}^{R,\mathcal{E}}(s,a)Z_{h,i}^R(\cdot|s,a) \tag{7}$$

4 Surrogate learning

Next, we define the discretized surrogate learning problem, and bound the size of the surrogate environments space, a significant term in the regret. To do so, we need to first define the Kolmogorov dimension of a set of parametrized distributions, esp. working out the case of l_1−distance. In the definitions below, we implicitly assume any required minimal measurability assumptions on the involved sets.

Definition 6. Given a set $\mathcal{F}$ of $\mathcal{O}$−parametrized distributions $P : \mathcal{O} \to \Delta(\mathcal{S})$ over a set $\mathcal{S}$ where both $\mathcal{O}, \mathcal{S}$ are measurable. Let $\mathcal{M}(\cdot, \cdot) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^{\geq 0}$ be a *distance*, i.e., $\mathcal{M}(P, Q) \geq 0 \overset{=}{\leftrightarrow} P = Q$. Then its right ε−covering number is the size $K_{\mathcal{M}}(\varepsilon)$ of the smallest set $\mathcal{C}_{\mathcal{M}}(\varepsilon) = \{P_1, \ldots, P_{K_{\mathcal{M}}(\varepsilon)}\} \subset \mathcal{F}$ such that

$$\forall P \in \mathcal{F}, \ \exists P_j \in \mathcal{C}_{\mathcal{M}}(\varepsilon) : \ \mathcal{M}(P, P_j) \leq \varepsilon. \tag{8}$$

The potential asymmetry of $\mathcal{M}$ (e.g., KL-divergence) requires the notion of left/right covering number. The right covering number will be the default, so covering number will always refer to that.

Definition 7. Let $d_{\mathcal{M}}(\varepsilon) = \log(K_{\mathcal{M}}(\varepsilon))$. Define the Kolmogorov $\mathcal{M}$−dimension $d_{\mathcal{M}}$ of $\mathcal{F}$ as

$$d_{\mathcal{M}} = \limsup_{\varepsilon \to 0} \frac{d_{\mathcal{M}}(\varepsilon)}{\log(\frac{1}{\varepsilon})}. \tag{9}$$

For $l_1(P, Q) := \sup_{o \in \mathcal{O}} ||P(\cdot|o) - Q(\cdot|o)||_1$, applying Definition 6 to the sets Θ_h^P, Θ_h^R with $\mathcal{O} = \mathcal{S} \times \mathcal{A}$, and denote the respective covering numbers by $L_h^P(\varepsilon), L_h^R(\varepsilon)$ corresponding to covering sets $\mathcal{C}_h^P(\varepsilon), \mathcal{C}_h^R(\varepsilon)$. Similarly applying Eq. (9) and denote the corresponding l_1−dimensions by $d_{l_1,h}^P(\varepsilon), d_{l_1,h}^R(\varepsilon), d_{l_1,h}^P, d_{l_1,h}^R$ and $d_{l_1}^P := \sum_h d_{l_1,h}^P, d_{l_1}^R := \sum_h d_{l_1,h}^R$. The sums $d_{l_1,h} := d_{l_1,h}^P + d_{l_1,h}^R, d_{l_1} := d_{l_1}^P + d_{l_1}^R$ can be interpreted as the l_1−dimension of Θ_h and Θ, i.e., the environment space.

Remark 3. We can also apply this framework to the KL-divergence, by $\mathcal{M}_{\mathrm{KL}}(P, Q) := \sup_{o \in \mathcal{O}} D_{\mathrm{KL}}(P(\cdot|o)||Q(\cdot||o))$. This was implicitly used by [18] to prove their regret bound in the tabular case. Note that Pinsker's lemma (Lemma 9) implies that the KL-divergence is larger than the squared total variance, and the latter is trivially larger than the l_1 distance. Therefore, l_1−dimension is smaller than $d_{\mathcal{M}_{\mathrm{KL}}}$, allowing for tighter regret bounds.

We now revisit the definition of ε−value partitions and show their existence is guaranteed by finite l_1−covering numbers. These partitions are the origins of surrogate environments.

Definition 8. Given $\varepsilon > 0$, an ε−value partition for a Bayesian RL problem is a partition $\{\Theta_k\}_{k=1}^K$ over Θ such that for any $k \in [K]$ and $\mathcal{E}, \mathcal{E}' \in \Theta_k$,

$$V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s_1^\ell) \leq \varepsilon. \tag{10}$$

A *layered* ε−value partition is one where the transition functions are independent over layers after conditioning on k. Throughout this paper, we will only consider layered ε−value partition. We define $K_{\mathrm{surr}}(\varepsilon)$ as the minimum K for which there exists a layered ε−value partition.

Inspired by Eq. (9), we define the surrogate dimension as $d_{\mathrm{surr}} = \limsup_{\varepsilon \to 0} \frac{K_{\mathrm{surr}}(\varepsilon)}{\log(1/\varepsilon)}$.

Lemma 1. *Given a Bayesian RL, we have $K_{\mathrm{surr}}(\varepsilon) \leq \prod_h L_h^P(\varepsilon/(2H)^2) \times L_h^R(\varepsilon/(4H))$. This implies $d_{\mathrm{surr}} \leq d_{l_1}$.*

The above is proved in Appendix B. It is hard to find d_{surr}, but one can estimate d_{l_1}, and according to the above, this acts as a proxy for K_{surr}. This is useful as the regret relates to K_{surr}. But to show this, we need to construct *surrogate environments* inside each partition, and show that learning those is almost equivalent to the original problem. Let ζ be a discrete random variable taking values in $\{1, \cdots, K_{\mathrm{surr}}(\varepsilon)\}$ that indicates the partition $\mathcal{E}$ lies in, such that $\zeta = k$ if and only if $\mathcal{E} \in \Theta_k$.

Lemma 2. *For any ε−value partition and any $\ell \in [L]$, there are random environments $\tilde{\mathcal{E}}_\ell^* \in \Theta$ with their laws only depending on $\zeta, \mathcal{D}_\ell$, such that*

$$\mathbb{E}_\ell \left[V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi_{TS}^\ell}^{\mathcal{E}}(s_1^\ell) \right] - \mathbb{E}_\ell \left[V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) - V_{1,\pi_{TS}^\ell}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) \right] \leq \varepsilon. \tag{11}$$

The expectation in both equations is over $\mathcal{E}$ and $\pi_{TS}^\ell \in \{\pi_{\mathcal{E}'}^*\}_{\mathcal{E}' \in \Theta}$, with both sampled independently $\sim \mathbb{P}_\ell(\cdot)$, and the K different values of $\tilde{\mathcal{E}}_\ell^*$. The second expectation over $(\tilde{\mathcal{E}}_\ell^*, \mathcal{E})$ is over pairs that are in the same partition, i.e., $\tilde{\mathcal{E}}_\ell^*, \mathcal{E}$ are independent only after conditioning on ζ.

We note that the proof in [18, App. B.1] contains the use of a lemma that does not apply to construct the law of the environment $\tilde{\mathcal{E}}_\ell^*$. More details is provided in Appendix C, where we find $\tilde{\mathcal{E}}_\ell^*$ by minimizing an expected value of π_{TS}^ℓ.

5 Bayesian regret bounds for Thompson Sampling

5.1 General Bayesian regret bound

We start by introducing the notion of value diameter.

Definition 9. Given the environment $\mathcal{E}$, its value diameter is defined as

$$\lambda_\mathcal{E} := \max_{1 \leq h \leq H} (\sup_s V_{h,\pi_\mathcal{E}^*}^\mathcal{E}(s) - \inf_s V_{h,\pi_\mathcal{E}^*}^\mathcal{E}(s)) + \max_{1 \leq h \leq H, s \in \mathcal{S}, a \in \mathcal{A}} (r_h^{\sup}(s,a) - r_h^{\inf}(s,a)),$$

where $r_h^{\sup}(s,a)$ (and $r_h^{\inf}(s,a)$) is the supremum (and infimum) of the set of rewards that are attainable under the distribution $r_h(s,a)$ with non-zero probability. As a special case, if rewards are deterministic, then we have $r_h^{\sup}(s,a) = r_h^{\inf}(s,a)$ for all s,a. The (average) value diameter over Θ is denoted by $\lambda := \mathbb{E}_{\mathcal{E} \sim \rho}[\lambda_\mathcal{E}^2]^{1/2}$.

As the value function is between 0 and H, we have $\lambda_\mathcal{E} \leq H + 1$ implying $\lambda \leq H + 1$. Note that value diameter is closely related to the notion of diameter commonly defined in finite RL problems. Strictly speaking, for a time-homogeneous RL, it is straightforward to see that the value diameter is bounded from above by one plus the diameter [33].

We now discuss the assumptions surrounding our results. The main technical assumption of this paper is the existence of consistent estimators, which as we will see in Appendix K, is closely related to the notion of posterior consistency:

Assumption 1. *There exists a strongly consistent estimator of the true environment given the history.*

Roughly speaking, we assume that with unlimited observations under TS, it is possible to find the true environment. For this assumption to fail, we need to have two environments that produce the same distribution over histories under TS and are therefore indistinguishable from the point of view of TS. The precise description of this assumption is detailed in Appendix K.

Another necessary technical assumption is that almost all optimal policies visit almost all state action pairs in their respective environment.

Assumption 2. *For almost every environment $\mathcal{E} \in \Theta$ and almost every $(s,a) \in \mathcal{S} \times \mathcal{A}$ and every $h \in [H]$, we have*

$$d_{h,\pi_\mathcal{E}^*}^\mathcal{E}(s,a) \neq 0.$$

Recall that, for any environment $\mathcal{E} \in \Theta$, the policy $\pi_\mathcal{E}^*$ is the optimal policy of $\mathcal{E}$ within the policy class Π. Therefore, one example of how the above assumption holds is when Π is the set of ε-greedy algorithms and transition functions of environments assign non-zero probability to every state. Under these assumptions, we discuss our main result and its corollaries.

Theorem 3. *Given a Bayesian RL problem, for all $\varepsilon > 0$, we have*

$$\mathfrak{BR}_L(\pi_{TS}) \leq 2\lambda \sqrt{\log(K_{surr}(\varepsilon))T} + L\varepsilon + T_0 \tag{12}$$

where T_0 does not depend on T. This can be further upper bounded by

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda \sqrt{d_{l_1}T}). \tag{13}$$

for large enough T. Given a homogeneous l_1 dimension $d_{hom} = d_{l_1,h}, \forall h$, this simplifies to

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda \sqrt{H d_{hom} T}). \tag{14}$$

Remark 4. For all regret bounds, we will replace $\lambda \leq H + 1$ to compare our result. For the case of homogeneous dimensions, we obtain $\widetilde{O}(H^{3/2}\sqrt{d_{hom}T})$. Crucially, our main result shows a new conceptual understanding of the information ratio by bounding it by two terms of different nature: H and λ, where the latter can be bounded by either the largest diameter of the environments or H.

Remark 5. Despite not impacting the asymptotics, the impact of T_0 can be large depending on the structure of the RL problem, and could be dominant even for large Ts in practice.

Remark 6. Considering time as a part of the state observation, one could apply this regret analysis to particular time-homogeneous settings. However, this mapping of time-inhomogeneous RLs to homogeneous ones is not surjective, hence the result above does not readily extend to time-homogeneous settings.

While [16] were the first to consider a nonlinear Bayesian RL model, their bound is limited to the Gaussian process (with linear kernel) setting, while ours in the nonlinear time inhomogeneous setting makes no assumptions on the prior and is the first such bound. Our novel analysis allow us to upper bound the information ratio by $\lambda\sqrt{H}$ instead of, for example $H^{3/2}\sqrt{SA}$ ([18]) in the tabular case, improving the regret bound by a square root relevant to the dimension d of the problem.

The detailed proof is given in Appendix D. Following [18], the regret (4) is rewritten using Lemma 2 to reduce the problem into its surrogate, and we use the well-known information-ratio trick by multiplying and dividing by the mutual information. We follow that with a Cauchy-Schwarz, summarized below

$$\mathfrak{BR}_L(\pi_{\mathrm{TS}}) \leq \mathbb{E}\left[\sum_{\ell=1}^{L} \frac{\mathbb{E}_\ell\left[V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) - V_{1,\pi_{\mathrm{TS}}^\ell}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell)\right]}{\sqrt{\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*;\mathcal{H}_{\ell,H})}}\sqrt{\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*;\mathcal{H}_{\ell,H})}\right] + L\varepsilon \tag{15}$$

$$\leq \sqrt{\mathbb{E}\left[\sum_{\ell=1}^{L} \frac{\left(\mathbb{E}_\ell\left[V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) - V_{1,\pi_{\mathrm{TS}}^\ell}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell)\right]\right)^2}{\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*;\mathcal{H}_{\ell,H})}\right]\mathbb{E}\left[\sum_{\ell=1}^{L}\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*;\mathcal{H}_{\ell,H})\right]} + L\varepsilon \tag{16}$$

Note the cost ε at each episode (Lemma 2) in the first inequality, yielding the overall error $L\varepsilon$. Then, we can bound the mutual information appearing in the regret term by $\mathbb{E}\left[\sum_{\ell=1}^{L}\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*;\mathcal{H}_{\ell,H})\right] = I_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*;\mathcal{D}_\ell) \leq I_\ell^{\pi_{\mathrm{TS}}^\ell}(\zeta;\mathcal{D}_\ell) \leq H(\zeta) \leq \log(K_{\mathrm{surr}}(\varepsilon))$, where we used the mutual information chain rule, followed by data processing inequality to substitute $\tilde{\mathcal{E}}_\ell^* \to \zeta$, and finally used the trivial bound by the entropy. But the main novelty of our approach lies in our control of the first term

$$\Gamma_\ell(\pi_{\mathrm{TS}}^\ell) := \frac{\left(\mathbb{E}_\ell\left[V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) - V_{1,\pi_{\mathrm{TS}}^\ell}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell)\right]\right)^2}{\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*;\mathcal{H}_{\ell,H})} \tag{17}$$

called the information ratio. In our analysis, we have the following bound on its expectation.

$$\mathbb{E}[\Gamma_\ell(\pi_{\mathrm{TS}}^\ell) \mid \mathcal{E}_0] \leq \mathbb{E}\left[\sum_h \int \frac{\mathbb{E}_\ell\left[(\lambda_{\mathcal{E}} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a))^2\right]}{\mathbb{E}_\ell\left[d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\right]}\mu_{\mathcal{S}\times\mathcal{A}} \mid \mathcal{E}_0\right],$$

where the average is taken over all histories $\mathcal{D}_\ell$ that are generated from running TS on the true environment $\mathcal{E}_0$, and we have introduced the smaller term $\lambda_{\mathcal{E}}$ instead of H in [18]. While [18] essentially bound the above only in the tabular setting with SAH^3, we manage to generally bound the above with a more precise bound using Doob's consistency theorem. Assumption 1 allows us to use Doob's consistency theorem to conclude that for almost every environment $\mathcal{E}_0$, almost every infinite sequence of histories $(\mathcal{D}_\ell)_{\ell=1}^\infty$ sampled from $\mathcal{E}_0$, and every integrable function f, the posterior mean $\mathbb{E}_\ell[f(\mathcal{E})] = \mathbb{E}[f(\mathcal{E}) \mid \mathcal{D}_\ell]$ converges to $f(\mathcal{E}_0)$. In particular, we conclude that $\mathbb{E}[\Gamma_\ell(\pi_{\mathrm{TS}}^\ell) \mid \mathcal{E}_0]$ tends to $\lambda_{\mathcal{E}_0}^2 H$ in the limit, allowing us to claim that for large enough ℓ, the expected information ratio $\mathbb{E}[\Gamma_\ell(\pi_{\mathrm{TS}}^\ell)]$ is uniformly bounded by $2\mathbb{E}[\lambda_{\mathcal{E}}^2]H = 2\lambda^2 H$. As there are L many such ratios, the two bounds together yield $2\sqrt{\lambda^2 HL} \cdot \sqrt{\log(K_{\mathrm{surr}}(\varepsilon))} + L\varepsilon$. This bound is true for large enough ℓ, giving the additional additive term T_0 in the theorem. Since this term is additive, applying Lemma 1 to bound $\log(K_{\mathrm{surr}}(\varepsilon))$, we have successfully shown the asymptotic behavior of the regret, independent of the prior, is of order $\widetilde{O}(H\sqrt{d_{l_1}T})$.

5.2 Applications

In each application below, the challenge is to bound d_{l_1} using the specifics of the model, and except for the case of tabular Bayesian RL, such analysis has not been carried out rigorously. We formalize the corollaries and show they are state-of-the-art compared to the literature.

Tabular RL. The result below follows from Theorem 3; the main contribution comes from our new information ratio bound, followed by the estimate $\widetilde{O}((\frac{1}{\varepsilon})^{SAH})$ of $K_{\text{surr}}(\varepsilon)$ ([18]).

Corollary 4. *Given a tabular Bayesian RL problem, for large enough T,*

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{HSAT}), \tag{18}$$

where the polylogarithmic terms are explicitly in terms of H, S, A, L.

We observe that our result matches [31] when their result in the time homogeneous setting (Definition 2) is extended to time inhomogeneous. However, in that paper, the authors assume a Dirichlet based prior which we do not.

Linear RL. A previous state-of-the-art $\widetilde{O}(d_f H^{3/2}\sqrt{T})$ was claimed by [18] to hold for linear Bayesian RLs with deterministic reward. We note:

- As in the previous cases, their proof in bounding their information ratio includes a factor of d_f, which ours avoids.

- We show that the proof bounding $K_{\text{surr}}(\varepsilon)$ in [18, App. B.4] is incorrect, starting with a wrong application of Cauchy-Schwarz and a wrong mutual information in their definition of information ratio. We provide counterexamples for the estimates found therein to substantiate our claim (see Appendix G.1).

To state our own corollary in this case, we need to define a few notions. Let $d_{l_1}^f = d_{l_1}^{P,f} + d_{l_1}^{R,f}$ be the sum of the l_1-dimensions of the feature map space $\{\psi_h^{P,\mathcal{E}}\}_{\mathcal{E}\in\Theta}, \{\psi_h^{R,\mathcal{E}}\}_{\mathcal{E}\in\Theta}$ where the l_1-distance between feature maps is defined as $l_1(\psi_h^{\mathcal{E}}, \psi_h^{\mathcal{E}'}) = \int_s \|\psi_h^{\mathcal{E}} - \psi_h^{\mathcal{E}'}\|_1 \mu_{\mathcal{S}}$. Our corollary also provides a concrete bound in the case of *mixture* linear Bayesian RL where the feature maps are themselves a sum of finitely many **fixed** feature maps. This means for all $\mathcal{E}\in\Theta$, we have

$$\psi_h^{P,\mathcal{E}} = \sum_{i=1}^{m_h^P} a_{h,i}^{P,\mathcal{E}} \Psi_{h,i}^P(s), \quad \psi_h^{R,\mathcal{E}} = \sum_{i=1}^{m_h^R} a_{h,i}^{R,\mathcal{E}} \Psi_{h,i}^R(s) \tag{19}$$

where $\{\Psi_{h,i}^P(s)\}_{i=1}^{m_h^P}, \{\Psi_{h,i}^R(s)\}_{i=1}^{m_h^R}$ are finitely many fixed feature maps and $\forall \mathcal{E}, h : \sum_i |a_{h,i}^{P,\mathcal{E}}|^2, \sum_i |a_{h,i}^{R,\mathcal{E}}|^2 \leq C_a$ for some constant $C_a > 0$. Let $M = M^P + M^R = \sum_h m_h^P + \sum_h m_h^R$.

Corollary 5. *For a linear Bayesian RL, for large enough T,*

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{d_{l_1}^f T}). \tag{20}$$

Given a linear Bayesian RL with finitely many states and total feature space dimension $d_f = d_f^P + d_f^R$, we have $d_{l_1} \leq 2d_f HS$, yielding for large enough T,

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{Hd_f ST}). \tag{21}$$

Given a mixture linear Bayesian RL, for large enough T,

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{MT}), \tag{22}$$

The proof is given in Appendix G. The fact that d_{l_1} appears instead of d_f in the general bound is not counter-intuitive, as we should expect the complexity of the feature map space $\{\psi_h^{P,\mathcal{E}}(s)\}_{\mathcal{E}\in\Theta,h\in[H]}, \{\psi_h^{R,\mathcal{E}}(s)\}_{\mathcal{E}\in\Theta,h\in[H]}$ to play a role in the regret, especially as this space can be very complex, and model very different environments that can not be grouped in the same $\varepsilon-$value partition.

Therefore, opposite to the claim made by [18], this complexity can not be captured by simply d_f except maybe in degenerate cases, such as when $\mathcal{S}$ is finite, which is our second statement. More generally, if each feature map $\psi_h^{P,\mathcal{E}}(s), \psi_h^{R,\mathcal{E}}(s)$ can be characterized with a vector of uniformly bounded norm $\boldsymbol{a}_h^{P,\mathcal{E}} \in \mathbb{R}^{m_h^P}, \boldsymbol{a}_h^{R,\mathcal{E}} \in \mathbb{R}^{m_h^R}$, then we can bound the regret in terms of m_h^P, m_h^R's, as is done in Eq. (22) (the finite state case corresponds to $m_h^P = d_f^P S, m_h^R = d_f^R S$).

Finite mixtures RL. To state our finite mixtures model result, we need to set the following notations. Let $d^m_{l_1} = d^{m,P}_{l_1} + d^{m,R}_{l_1} = \sum_h d^{m,P}_{l_1,h} + \sum_h d^{m,R}_{l_1,h}$ correspond to the total l_1-dimension of the space of mixtures coefficient maps $\{\boldsymbol{a}^{P,\mathcal{E}}_h(s,a)\}_{\mathcal{E}\in\Theta}, \{\boldsymbol{a}^{R,\mathcal{E}}_h(s,a)\}_{\mathcal{E}\in\Theta}$ with l_1- distance defined as $l_1(\boldsymbol{a}^{\mathcal{E}}_h, \boldsymbol{a}^{\mathcal{E}'}_h) = \sup_{s,a} \|\boldsymbol{a}^{\mathcal{E}}_h(s,a) - \boldsymbol{a}^{\mathcal{E}'}_h(s,a)\|_1$. Define also the restricted finite mixtures model where $\boldsymbol{a}^{P,\mathcal{E}}_h, \boldsymbol{a}^{R,\mathcal{E}}_h$ are vectors in $\mathbb{R}^{m^P_h}, \mathbb{R}^{m^R_h}$ independent of (s,a) and let $M = M^P + M^R = \sum_h m^P_h + \sum_h m^R_h$.

Corollary 6. *Given a finite mixtures Bayesian RL problem, for large enough T,*

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{d^m_{l_1}T}) . \tag{23}$$

Assuming the restricted finite mixtures model, for large enough T,

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}\left(\lambda\sqrt{MT}\right) . \tag{24}$$

which, given a uniform dimension $m = m^P_h = m^R_h$, yields $\widetilde{O}(\lambda\sqrt{HmT})$.

We prove the above in Appendix H, deriving it from our generic bound, after relating the l_1-dimension d_{l_1} of the environment space to that of the mixtures coefficients. To the best of our knowledge, this is the first bound for finite mixtures Bayesian RL problems. We note that in a previous work ([6]), a restricted version of finite mixtures, like in Eq. (24), was considered in the frequentist setting.

We finish this section by proposing the following conjecture, in line with [31, Conj. 1].

Conjecture 7. *For the Bayesian RL, the following is true and optimal for **all** T:*

$$\mathfrak{BR}_L(\pi_{TS}) \leq O\left(\inf_{\varepsilon>0}(\sqrt{H\log(K_{\mathrm{surr}}(\varepsilon))T} + L\varepsilon)\right) . \tag{25}$$

where the constant factor is independent of the prior. This means there exists a Bayesian RL problem such that $\mathfrak{BR}_L(\pi_{TS}) = \widetilde{\Omega}(\sqrt{Hd_{\mathrm{surr}}T})$. All polylogarithmic terms are in terms of H, d_{surr}, T.

Note that the above coincides with the lower bound for the (model-based) time inhomogeneous frequentist setting; see e.g., [21] for the proven lower bound for the tabular case. This is also $\sqrt{H}$ higher (this factor being baked in d_{surr}) than that of the time homogeneous frequentist setting, which is expected, according to [21, App. D]. Note that in this conjecture, the λ in our bound is replaced by $\sqrt{H}$, and the conjecture is not for T large enough, but for all T. Supporting this conjecture requires experiments where TS can be exactly implemented assuming access to an oracle which provides the optimal policy for a query environment. Simulations have been performed for the similar [31, Conj. 1] in the time homogeneous case. Our conjecture is similar but with the additional expected factor of $\sqrt{H}$ due to time inhomogeneity, thus their simulation also supports the above.

6 Conclusions

In this paper, we have addressed the Bayesian Reinforcement Learning (RL) problem in the context of time inhomogeneous transition and reward functions. By considering both Bayesian transition and Bayesian rewards without prior assumptions, we have extended the scope of previous works, making our formulation more comprehensive. To simplify the learning problem, we have introduced surrogate environments, which discretize the environment space. We have established a connection between the size of this new environment space and the l_1-dimensions of the transition and reward functions space, providing insights into the l_1-dimension of the environment space denoted by d_{l_1}. We have employed posterior consistency tools to analyze the information ratio, which captures the trade-off between exploration and exploitation. We conjecture that (at least a weakened version of) our posterior consistency assumption should hold in general, which is left for future work. Our analysis has resulted in a refined approach to estimate the Bayesian regret in Thompson Sampling (TS), yielding a regret bound of $\widetilde{O}(\lambda\sqrt{d_{l_1}T})$ for large enough time steps T. The result is specialized to linear, tabular, and finite mixtures MDPs.

Limitations: While the paper provides asymptotic generic regret bound for TS in a generalized setup which improve the state of the art results, finding lower bounds, esp. one dependent on λ, are left open. In addition, the issue of prior misspecificity is not discussed and left for future studies.

References

[1] Amirhesam Abedsoltan, Mikhail Belkin, and Parthe Pandit. Toward large kernel models. *arXiv preprint arXiv:2302.02605*, 2023.

[2] Alekh Agarwal, Yujia Jin, and Tong Zhang. VOQL: Towards optimal regret in model-free rl with nonlinear function approximation. *arXiv preprint arXiv:2212.06069*, 2022.

[3] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

[4] Dilip Arumugam and Benjamin Van Roy. Deciding what to model: Value-equivalent sampling for reinforcement learning. 35:9024–9044, 2022.

[5] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[6] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[7] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[8] Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[9] Sébastien Bubeck and Mark Sellke. First-order bayesian regret analysis of thompson sampling. In *Algorithmic Learning Theory*, pages 196–233. PMLR, 2020.

[10] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[11] Souradip Chakraborty, Amrit Bedi, Alec Koppel, Pratap Tokekar, Furong Huang, and Dinesh Manocha. Posterior coreset construction with kernelized stein discrepancy for model-based reinforcement learning. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[12] Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 3197–3205. PMLR, 2019.

[13] Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12040–12051, 2021.

[14] Shi Dong and Benjamin Van Roy. An information-theoretic analysis for thompson sampling with many actions. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] Yaqi Duan, Mengdi Wang, and Martin J Wainwright. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*, 2021.

[16] Ying Fan and Yifei Ming. Model-based reinforcement learning for continuous control with posterior sampling. In *International Conference on Machine Learning*, pages 3078–3087. PMLR, 2021.

[17] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[18] Botao Hao and Tor Lattimore. Regret bounds for information-directed reinforcement learning. *arXiv preprint arXiv:2206.04640*, 2022.

[19] Botao Hao, Tor Lattimore, and Wei Deng. Information directed sampling for sparse linear bandits. *Advances in Neural Information Processing Systems*, 34:16738–16750, 2021.

[20] Botao Hao, Tor Lattimore, and Chao Qin. Contextual information-directed sampling. In *International Conference on Machine Learning*, pages 8446–8464. PMLR, 2022.

[21] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

[22] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[23] Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.

[24] Tor Lattimore and Andras Gyorgy. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.

[25] Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pages 2111–2139. PMLR, 2019.

[26] Fang Liu, Swapna Buccapatnam, and Ness Shroff. Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[27] Xiuyuan Lu. *Information-directed sampling for reinforcement learning*. Stanford University, 2020.

[28] Xiuyuan Lu and Benjamin Van Roy. Information-theoretic confidence bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

[30] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

[31] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.

[32] Ian Osband, Benjamin Van Roy, Daniel J Russo, Zheng Wen, et al. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20(124):1–62, 2019.

[33] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[34] Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32, 2019.

[35] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27, 2014.

[36] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[37] Remo Sasso, Michelangelo Conserva, and Paulo Rauber. Posterior sampling for deep reinforcement learning. *arXiv preprint arXiv:2305.00477*, 2023.

[38] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[39] Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Éric Moulines, Remi Munos, Alexey Naumov, Mark Rowland, Michal Valko, and Pierre Ménard. Optimistic posterior sampling for reinforcement learning with few samples and tight guarantees. *Advances in Neural Information Processing Systems*, 35:10737–10751, 2022.

[40] Aristide Tossou, Debabrota Basu, and Christos Dimitrakakis. Near-optimal optimistic rein-forcement learning using empirical bernstein inequalities. *arXiv preprint arXiv:1905.12425*, 2019.

[41] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[42] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[43] Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learn-ingvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020.

[44] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

A   Related works

In the related works section of the main text, we mostly focused on Bayesian regret. Here we include a brief paragraph on bounds on frequentist regret.

For the frequentist setting, various algorithms with provable regret guarantees have been proposed for model-free tabular MDPs. These include UCBVI [7], optimistic Q-learning [21], RLSVI [34, 42], and UCB-Advantage [43]. These algorithms were further generalized to linear or linear mixture MDPs, such as LSVI-UCB [22], OPPO [10], and UCRL-VTR [6, 44]. Slightly more related to our work, model-based frequentist bounds have also been shown for a variant of posterior sampling (PS) in the tabular setting [3]. For the specific variant of *optimistic* PSRL, the optimal bound in the tabular setting with a Dirichlet prior was shown in [39]. To our knowledge, a frequentist bound for PS is still an open problem for general RLs. Minimax regret bounds have also been studied for variants of TS, as in [13]. Most recently, [2] presented VOQL, an algorithm that achieves the optimal bound of $\widetilde{O}(d\sqrt{HT})$ in the general model-free nonlinear setting, where d represents the generalized Eluder dimension of the value function space. Note that the notion of dimension used in our regret bounds is different, and unrelated, to the Eluder dimension used in model-free estimations. For frequentist model-based, the optimal bound was achieved in the tabular setting by [7]. As another research direction, [15] utilized kernel-Hilbert spaces to estimate the value of infinite horizon Markov reward process (MRP) for RL problem, and [1] pave the way for scalability challenges in kernel models.

B   Proof of Lemma 1

To avoid conflict with the environment space notation $\Theta = \Theta_1 \times \cdots \times \Theta_H$, we adopt the notation $\Theta_{\vec{k}}^{\varepsilon}$ to refer to $\varepsilon-$value partitions.

Proof of covering number estimate.
Let $\{B_{h,i}^P(\varepsilon/(2H)^2)\}_{i=1}^{L_h^P(\varepsilon/(2H)^2)}, \{B_{h,i}^R(\varepsilon/(4H))\}_{i=1}^{L_h^R(\varepsilon/(4H))}$ be the $\varepsilon-$balls giving an $\varepsilon-$covering for Θ_h^P, Θ_h^R. Then define the $\varepsilon-$value partition $\cup_{k=1}^K \Theta_{\vec{k}}^{\varepsilon} = \Theta$ where $K = \prod_h L_h^P(\varepsilon/(2H)^2) \times L_h^R(\varepsilon/(4H))$ as follows. Each $k \in [K]$ can be enumerated as a $2H-$tuple $(i_1, j_1, \ldots, i_H, j_H)$ where $i_h \in [L_h^P(\varepsilon/(2H)^2)], j_h \in [L_h^R(\varepsilon/(4H))]$. Define $\Theta_{\vec{k}}^{\varepsilon} = \{\mathcal{E}|P_h^{\mathcal{E}} \in B_{h,i_h}^P(\varepsilon/(2H)^2), r_h^{\mathcal{E}} \in B_{h,j_h}^R(\varepsilon/(4H))\}$. It is straightforward to check that $\cup_k \Theta_{\vec{k}}^{\varepsilon} = \Theta$. Any environment appearing redundantly can be removed from all but one of the $\Theta_{\vec{k}}^{\varepsilon}$'s it lives in, so that we have a true partition of $\Theta_{\vec{k}}^{\varepsilon}$.

Next, we will need to use the following lemma.

Proving that our partition is an $\varepsilon-$value partition requires us to show that for any $\mathcal{E}, \mathcal{E}' \in \Theta_{\vec{k}}^{\varepsilon}$: $V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^{\ell}) - V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s_1^{\ell}) \leq \varepsilon$. We have

$$V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^{\ell}) - V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s_1^{\ell}) = \sum_{h=1}^{H} \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\mathcal{E}'} \left[\mathbb{E}_{s' \sim P_h^{\mathcal{E}}(\cdot|s_h,a_h)}[V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s')] - \mathbb{E}_{s' \sim P_h^{\mathcal{E}'}(\cdot|s_h,a_h)}[V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s')] \right]$$

$$+ \sum_{h=1}^{H} \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\mathcal{E}'}[r_h^{\mathcal{E}}(s_h,a_h) - r_h^{\mathcal{E}'}(s_h,a_h)], \tag{26}$$

Rewrite the first term and bound it as follows:

$$\sum_{h=1}^{H} \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\mathcal{E}'} \left[\int_{\mathcal{S}} P_h^{\mathcal{E}}(s'|s_h^{\ell},a_h^{\ell})V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s') - \int_{\mathcal{S}} P_h^{\mathcal{E}'}(s'|s_h^{\ell},a_h^{\ell})V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s') \right]$$

$$\leq \sum_{h=1}^{H} \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\mathcal{E}'} \left[\left(\int_{\mathcal{S}} \left| P_h^{\mathcal{E}}(s'|s_h^{\ell},a_h^{\ell}) - P_h^{\mathcal{E}'}(s'|s_h^{\ell},a_h^{\ell}) \right| V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s') \right) \right]. \tag{27}$$

where integrals are with respect to the measure on $\mathcal{S}$. Then, we can bound probability transitions terms by

$$\sum_{h=1}^{H}\mathbb{E}_{\pi_{\mathcal{E}}^{*}}^{\mathcal{E}'}\left[\left(\int_{\mathcal{S}}\left|P_{h}^{\mathcal{E}}(s'|s_{h}^{\ell},a_{h}^{\ell})-P_{h}^{\mathcal{E}'}(s'|s_{h}^{\ell},a_{h}^{\ell})\right|V_{h+1,\pi_{\mathcal{E}}^{*}}^{\mathcal{E}}(s')\right)\right]$$

$$\leq H\sum_{h=1}^{H}\mathbb{E}_{\pi_{\mathcal{E}}^{*}}^{\mathcal{E}'}\left[\int_{\mathcal{S}}\left|P_{h}^{\mathcal{E}}(s'|s_{h}^{\ell},a_{h}^{\ell})-P_{h}^{\mathcal{E}'}(s'|s_{h}^{\ell},a_{h}^{\ell})\right|\right]$$

$$\leq H\sum_{h=1}^{H}\sup_{s,a}\left(\int_{\mathcal{S}}\left|P_{h}^{\mathcal{E}}(s'|s,a)-P_{h}^{\mathcal{E}'}(s'|s,a)\right|\right) \tag{28}$$

$$=H\sum_{h=1}^{H}l_{1}(P_{h}^{\mathcal{E}},P_{h}^{\mathcal{E}'})\leq H(2\frac{\varepsilon}{4H^{2}}\cdot H)$$

and similarly, reward terms by

$$\sum_{h=1}^{H}E_{\pi_{\mathcal{E}}^{*}}^{\mathcal{E}'}[r_{h}^{\mathcal{E}}(s_{h}^{\ell},a_{h}^{\ell})-r_{h}^{\mathcal{E}'}(s_{h}^{\ell},a_{h}^{\ell})]=\sum_{h=1}^{H}E_{\pi_{\mathcal{E}}^{*}}^{\mathcal{E}'}\left[\int_{0}^{1}x\left(r_{h}^{\mathcal{E}}(x|s_{h}^{\ell},a_{h}^{\ell})-r_{h}^{\mathcal{E}'}(x|s_{h}^{\ell},a_{h}^{\ell})\right)\mathrm{d}x\right]$$

$$\leq\sum_{h=1}^{H}E_{\pi_{\mathcal{E}}^{*}}^{\mathcal{E}'}\left[\int_{0}^{1}\left|x\left(r_{h}^{\mathcal{E}}(x|s_{h}^{\ell},a_{h}^{\ell})-r_{h}^{\mathcal{E}'}(x|s_{h}^{\ell},a_{h}^{\ell})\right)\right|\mathrm{d}x\right]$$

$$\leq\sum_{h=1}^{H}\sup_{s,a}\int_{0}^{1}\left|x\left(r_{h}^{\mathcal{E}}(x|s,a)-r_{h}^{\mathcal{E}'}(x|s,a)\right)\right|\mathrm{d}x$$

$$=\sum_{h=1}^{H}l_{1}(r_{h}^{\mathcal{E}},r_{h}^{\mathcal{E}'})\leq H(2\frac{\varepsilon}{4H}) \tag{29}$$

See also Remark 7 for the reward term bound. In the first inequality, we used that $V_{h+1,\pi_{\mathcal{E}}^{*}}^{\mathcal{E}}(s')$ is always bounded by H, and in both cases we used the fact that the transition and reward functions of $\mathcal{E},\mathcal{E}'$ live inside the same balls, with their l_{1} distance being at most twice the radii $\varepsilon/(2H)^{2}$ and $\varepsilon/(4H)$, respectively. Adding up the above two estimates equals ε, as desired. This shows that our Θ_{k}^{ε} partition is an $\varepsilon-$value partition, hence $K_{\mathrm{surr}}(\varepsilon)\leq K=\prod_{h}L_{h}^{P}(\varepsilon/(2H)^{2})\times L_{h}^{R}(\varepsilon/(4H))$.

Remark 7. Notice that the $l_{1}-$distance of two reward functions is over their probability distributions, and is larger than their expected norm difference, i.e.,

$$l_{1}(r,r')=\sup_{s,a\in\mathcal{S}\times\mathcal{A}}||r(\cdot|s,a)-r'(\cdot|s,a)||_{1}=\sup_{s,a\in\mathcal{S}\times\mathcal{A}}\int_{0}^{1}|r(x|s,a)-r'(x|s,a)|$$

$$\geq\sup_{s,a\in\mathcal{S}\times\mathcal{A}}\int_{0}^{1}x|r(x|s,a)-r'(x|s,a)|=\sup_{s,a\in\mathcal{S}\times\mathcal{A}}\mathbb{E}[|r(s,a)-r'(s,a)|]. \tag{30}$$

Proof of $d_{\mathrm{surr}}\leq d_{l_{1}}$**.**
By taking the log, dividing by $\log(1/\varepsilon)$, and taking the lim sup of both sides of this inequality, we can infer the second statement of the lemma:

$$d_{\mathrm{surr}}=\limsup_{\varepsilon\to0}\frac{\log(K_{\mathrm{surr}}(\varepsilon))}{\log(1/\varepsilon)}$$

$$\leq\sum_{h=1}^{H}\limsup_{\varepsilon\to0}\frac{\log(L_{h}^{P}(\varepsilon/(2H)^{2}))}{\log(\frac{1}{\varepsilon/(2H)^{2}})}\cdot\frac{1}{1+\log((2H)^{2})/\log(\varepsilon/(2H)^{2})}+$$

$$\sum_{h=1}^{H}\limsup_{\varepsilon\to0}\frac{\log(L_{h}^{R}(\varepsilon/(4H)))}{\log(\frac{1}{\varepsilon/(2H)^{2}})}\cdot\frac{1}{1+\log(4H)/\log(\varepsilon/(4H))} \tag{31}$$

$$=\sum_{h}d_{l_{1},h}^{P}+\sum_{h}d_{l_{1},h}^{R}=d_{l_{1}}.$$

Fact 1. We separate the statement proved in Eqs. (28) and (29) as fact, useful for future use: for all $\mathcal{E}, \mathcal{E}' \in \Theta$,

$$V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s_1^\ell) \leq H \sum_{h=1}^{H} l_1(P_h^{\mathcal{E}}, P_h^{\mathcal{E}'}) + \sum_{h=1}^{H} l_1(r_h^{\mathcal{E}}, r_h^{\mathcal{E}'}). \tag{32}$$

C Proof of Lemma 2

While we follow the proof of the same lemma in [18, App. B.1], we will need to correct some mistakes. Let us restate the equation of the statement:

Lemma. *For any $\varepsilon-$value partition and any $\ell \in [L]$, there are random environments $\tilde{\mathcal{E}}_\ell^* \in \Theta$ with their laws only depending on $\zeta, \mathcal{D}_\ell$, such that*

$$\mathbb{E}_\ell \left[V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi_{TS}^\ell}^{\mathcal{E}}(s_1^\ell) \right] - \mathbb{E}_\ell \left[V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) - V_{1,\pi_{TS}^\ell}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) \right] \leq \varepsilon. \tag{33}$$

Proof. Assume a partition Θ_k^ε satisfying Definition 8 with error ε exists. Let $\mathcal{E}_\ell \sim P(\cdot|\mathcal{D}_\ell)$. We want to decompose $\mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\mathcal{E}_\ell}(s_1^\ell) \big| \mathcal{E}_\ell \in \Theta_k^\varepsilon \right]$, where the expectation is over all $\mathcal{E}_\ell \in \Theta_k^\varepsilon$ and all $\pi_{TS}^\ell = \pi_{\mathcal{E}_\ell'}^*$, where $\mathcal{E}_\ell' \in \Theta$, with $\mathcal{E}_\ell'$ independent of $\mathcal{E}_\ell$. We decompose this by writing the expectation only over the former.

$$\begin{aligned}
\mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\mathcal{E}_\ell}(s_1^\ell) \big| \mathcal{E}_\ell \in \Theta_k^\varepsilon \right] &= \int_{\mathcal{E}_0 \in \Theta_k^\varepsilon} \mathbb{P}\left(\mathcal{E}_\ell = \mathcal{E}_0 | \mathcal{E}_\ell \in \Theta_k^\varepsilon\right) \mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\mathcal{E}_0}(s_1^\ell) \big| \mathcal{E}_\ell \in \Theta_k^\varepsilon \right] \mathrm{d}\rho(\mathcal{E}_0) \\
&= \int_{\mathcal{E}_0 \in \Theta_k^\varepsilon} \mathbb{P}\left(\mathcal{E}_\ell = \mathcal{E}_0 | \mathcal{E}_\ell \in \Theta_k^\varepsilon\right) \mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\mathcal{E}_0}(s_1^\ell) \right] \mathrm{d}\rho(\mathcal{E}_0),
\end{aligned} \tag{34}$$

where the last equation is due to the independence between $\mathcal{E}_\ell$ and $\mathcal{E}_\ell'$.

We would like to find some $\tilde{\mathcal{E}}_{k,\ell}^*$ such that its corresponding expected value is smaller than the integral above, i.e.

$$\mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\tilde{\mathcal{E}}_{k,\ell}^*}(s_1^\ell) \right] \leq \int_{\mathcal{E}_0 \in \Theta_k^\varepsilon} \mathbb{P}\left(\mathcal{E}_\ell = \mathcal{E}_0 | \mathcal{E}_\ell \in \Theta_k^\varepsilon\right) \mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\mathcal{E}_0}(s_1^\ell) \right] \mathrm{d}\rho(\mathcal{E}_0) \tag{35}$$

We set values of this random variable as $\tilde{\mathcal{E}}_{k,\ell}^* = \mathbb{E}_\ell[\mathcal{E}|\mathcal{E} \in \Theta_k^\varepsilon], \forall k \in [K_{\mathrm{surr}}(\varepsilon)]$. In other words, the posterior mean over Θ_k^ε. Now, we can define a new random environment $\tilde{\mathcal{E}}_\ell^*$ which prior can be easily computed as $\mathbb{P}_\ell \left(\tilde{\mathcal{E}}_\ell^* = \tilde{\mathcal{E}}_{k,\ell}^* \right) = \mathbb{P}_\ell(\mathcal{E} \in \Theta_k^\varepsilon)$ and $\zeta(\tilde{\mathcal{E}}_{k,\ell}^*) = k$. Note the law of $\tilde{\mathcal{E}}_\ell^*$ only depends on $\zeta \in [K]$ and $\mathcal{D}_\ell$, and conditional on ζ, $\tilde{\mathcal{E}}_\ell^*$ is independent of $\mathcal{E}$, as desired. While this definition of $\tilde{\mathcal{E}}_\ell^*$ may not be in Θ, that is fine as we do not use this condition in our proof (see also Remark 8).

Notice that as a result, the overall posterior mean of $\bar{\mathcal{E}}_\ell^* := \mathbb{E}_\ell[\tilde{\mathcal{E}}_\ell^*]$ coincides with $\bar{\mathcal{E}}_\ell$. Also note that by this definition of $\tilde{\mathcal{E}}_{k,\ell}^*$, and the independence over layers even after conditioning on Θ_k^ε, the mean $\mathbb{E}_\ell[V_{1,\pi_{TS}^\ell}^{\tilde{\mathcal{E}}_{k,\ell}^*}(s_1^\ell)]$ is in fact equal to the mean on the right hand side of Eq. (35), thus our construction is sharp in that it satisfies this inequality with an equality.

Now, we are ready to make the connection with Eq. (11), showing that solving the surrogate environment problem is 'almost the same' as solving the original problem, up to some ε. Integrating over the different values of k with prior $\mathbb{P}_\ell(\mathcal{E} \in \Theta_k^\varepsilon)$,

$$\mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) \right] \leq \mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\mathcal{E}}(s_1^\ell) \right] \implies \mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) \right] - \mathbb{E}_\ell \left[V_{1,\pi_{TS}^\ell}^{\mathcal{E}}(s_1^\ell) \right] \leq 0 \tag{36}$$

Lastly, by the partition property,

$$\mathbb{E}_\ell \left[V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) \right] \leq \varepsilon. \tag{37}$$

where in the expectation above, $\zeta(\tilde{\mathcal{E}}_\ell^*) = \zeta(\mathcal{E})$, i.e., $\mathcal{E}$ and $\tilde{\mathcal{E}}_\ell^*$ are clearly independent only after conditioning on $\zeta = k$, as also required in Eq. (11). Adding the above two inequalities gives

$$\mathbb{E}_\ell \left[V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi_{TS}^\ell}^{\mathcal{E}}(s_1^\ell) \right] - \mathbb{E}_\ell \left[V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) - V_{1,\pi_{TS}^\ell}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) \right] \leq \varepsilon \tag{38}$$

finishing the proof. Notice that Eq. (37) is the only place where the property of $\varepsilon-$value partitioning is used. $\qquad\square$

Remark 8. We note that if $\tilde{\mathcal{E}}_\ell^*$ is any convex combination of environments in $\Theta_{\zeta(\mathcal{E})}$, then by the ε−value partition property $\mathbb{E}_\ell[V_{1,\pi_{\tilde{\mathcal{E}}}^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi_{\tilde{\mathcal{E}}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell)] \leq \varepsilon$. The distinct property that surrogate environments satisfy is

$$\mathbb{E}_\ell[V_{1,\pi_{\mathrm{TS}}^\ell}^{\tilde{\mathcal{E}}_{k,\ell}^*}(s_1^\ell)] \leq \mathbb{E}_\ell[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}}(s_1^\ell)|\mathcal{E} \in \Theta_k^{\bar\varepsilon}], \ \forall k \in [K]. \tag{39}$$

and therefore, satisfy Eq. (33).

Remark 9. For the purpose of fully addressing the issue present in [18] in proving the above lemma, we show an alternative construction which does not assume a layered ε−value partition. One can always find a decreasing sequence $\mathbb{E}_\ell\left[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}^1}(s_1^\ell)\right] \geq \mathbb{E}_\ell\left[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}^2}(s_1^\ell)\right] \geq \ldots$ with limit $\inf_{\mathcal{E}_0 \in \Theta_k^\varepsilon} \mathbb{E}_\ell\left[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}_0}(s_1^\ell)\right]$. Then we claim there exists some J such that the above inequality is true for $\tilde{\mathcal{E}}_{k,\ell}^* = \mathcal{E}^J$. Otherwise, we have

$$\mathbb{E}_\ell\left[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}^i}(s_1^\ell)\right] > \int_{\mathcal{E}_0 \in \Theta_k^\varepsilon} \mathbb{P}\left(\mathcal{E}_\ell = \mathcal{E}_0|\mathcal{E}_\ell \in \Theta_k^\varepsilon\right) \mathbb{E}_\ell\left[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}_0}(s_1^\ell)\right] \mathrm{d}\rho(\mathcal{E}_0) \tag{40}$$

for all i. Taking the limit $i \to \infty$, we get

$$\inf_{\mathcal{E}_0 \in \Theta_k^\varepsilon} \mathbb{E}_\ell\left[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}_0}(s_1^\ell)\right] \geq \int_{\mathcal{E}_0 \in \Theta_k^\varepsilon} \mathbb{P}\left(\mathcal{E}_\ell = \mathcal{E}_0|\mathcal{E}_\ell \in \Theta_k^\varepsilon\right) \mathbb{E}_\ell\left[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}_0}(s_1^\ell)\right] \mathrm{d}\rho(\mathcal{E}_0) \tag{41}$$

which can only be true if $\mathbb{E}_\ell\left[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}_0}(s_1^\ell)\right] = \inf_{\mathcal{E}_0 \in \Theta_k^\varepsilon} \mathbb{E}_\ell\left[V_{1,\pi_{\mathrm{TS}}^\ell}^{\mathcal{E}_0}(s_1^\ell)\right]$ almost everywhere. Any $\mathcal{E}_0$ satisfying this equality would also satisfy our requirement for $\tilde{\mathcal{E}}_{k,\ell}^*$. It is important to note that $\tilde{\mathcal{E}}_{k,\ell}^*$ depends on both k and $\mathcal{D}_\ell$, as mentioned in the lemma's statement. The finishes the construction of $\tilde{\mathcal{E}}_\ell^*$ and the rest follows similar to the above proof.

Remark 10 (Incorrect proof by [18] of the lemma). Eq. (34) is different from what appears in [18, App. B.1], where we have corrected for the abuse of notation of $\mathcal{E}$ occurring in e.g. "$\mathbb{P}\left(\mathcal{E}_\ell = \mathcal{E}|\mathcal{E}_\ell \in \Theta_k^\varepsilon\right) \mathbb{E}_\ell\left[V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^\ell)\right]$". We further note the use of summation over $\mathcal{E} \in \Theta_k^\varepsilon$ in their equation, instead of an integral which is required even in the tabular case. This is not easily fixed by just replacing sum with integral, since the application of [18, Lemma D.1] used afterwards in their proof depends on having a finite sum. We mention this lemma below:

(Lemma 1 in [14]) Let $\{a_i\}_{i=1}^N$ and $\{b_i\}_{i=1}^N$ be two sequences of real numbers, where $N < \infty$. Let $\{p_i\}_{i=1}^N$ be such that $p_i \geq 0$ for all i and $\sum_{i=1}^N p_i = 1$. Then there exists indices $j, k \in [N]$ and $r \in [0, 1]$ such that

$$ra_j + (1 - r)a_k \leq \sum_{i=1}^N a_i p_i, rb_j + (1 - r)b_k \leq \sum_{i=1}^L b_i p_i \,.$$

The application of this lemma is key to the proof as the authors cite it to find the right surrogate environments that satisfy Eq. (39). However, even in this application, it is not clear what exactly is the 'second' set of numbers (b_j above), as is required for that Lemma to be a nontrivial result, which makes its usage even more questionable.

D   Proof of Theorem 3

We restate the theorem for ease of reference as it contains multiple statements.

Theorem. *Given a Bayesian RL problem, we have*

$$\mathfrak{BR}_L(\pi_{TS}) \leq \inf_{\varepsilon > 0} \left(\lambda\sqrt{\log(K_{\mathrm{surr}}(\varepsilon))T} + L\varepsilon\right) + T_0 \tag{42}$$

where T_0 does not depend on T. This can be further upper bounded by

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{d_{l_1}T}) \,. \tag{43}$$

for large enough T. Given a homogeneous l_1 dimension $d_{\mathrm{hom}} = d_{l_1,h}, \forall h$, this simplifies to

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{Hd_{\mathrm{hom}}T}) \,. \tag{44}$$

Proof. The proof starts by employing the surrogate environment learning bound from Appendix C:

$$\mathbb{E}_\ell \left[V^{\mathcal{E}}_{1,\pi^*_{\mathcal{E}}}(s^\ell_1) - V^{\mathcal{E}}_{1,\pi^\ell_{\mathrm{TS}}}(s^\ell_1) \right] - \varepsilon \le \mathbb{E}_\ell \left[V^{\tilde{\mathcal{E}}^*_\ell}_{1,\pi^*_{\mathcal{E}}}(s^\ell_1) - V^{\tilde{\mathcal{E}}^*_\ell}_{1,\pi^\ell_{\mathrm{TS}}}(s^\ell_1) \right] . \tag{45}$$

We have

$$\begin{aligned}
\mathfrak{BR}_L(\pi_{\mathrm{TS}}) &= \sum_{\ell=1}^L \mathbb{E}\left[\mathbb{E}_\ell \left[V^{\mathcal{E}}_{1,\pi^*_{\mathcal{E}}}(s^\ell_1) - V^{\mathcal{E}}_{1,\pi^\ell_{\mathrm{TS}}}(s^\ell_1) \right] \right] \\
&= \sum_{\ell=1}^L \mathbb{E}\left[\mathbb{E}_\ell \left[V^{\mathcal{E}}_{1,\pi^*_{\mathcal{E}}}(s^\ell_1) - V^{\mathcal{E}}_{1,\pi^\ell_{\mathrm{TS}}}(s^\ell_1) \right] - \varepsilon \right] + L\varepsilon \\
&= \sum_{\ell=1}^L \mathbb{E}\left[\mathbb{E}_\ell \left[V^{\tilde{\mathcal{E}}^*_\ell}_{1,\pi^*_{\mathcal{E}}}(s^\ell_1) - V^{\tilde{\mathcal{E}}^*_\ell}_{1,\pi^\ell_{\mathrm{TS}}}(s^\ell_1) \right] \right] + L\varepsilon
\end{aligned} \tag{46}$$

Given that π^ℓ_{TS} is independent from $\tilde{\mathcal{E}}^*_\ell$ and the independence of the latter's prior over different layers (due to the layered $\varepsilon-$value partition), we observe that $\mathbb{E}_\ell[V^{\tilde{\mathcal{E}}^*_\ell}_{1,\pi^\ell_{\mathrm{TS}}}(s^\ell_1)] = \mathbb{E}_\ell[V^{\bar{\mathcal{E}}^*_\ell}_{1,\pi^\ell_{\mathrm{TS}}}(s^\ell_1)]$. Again, since π^ℓ_{TS} is also independent from $\bar{\mathcal{E}}^*_\ell$, and that π^ℓ_{TS} and $\mathcal{E}$ have the same laws conditional on $\mathcal{D}_\ell$, we can rewrite the latter as $\mathbb{E}_\ell[V^{\bar{\mathcal{E}}^*_\ell}_{1,\pi^*_{\mathcal{E}}}(s^\ell_1)]$. This comes with the obvious note that $\bar{\mathcal{E}}^*_\ell$ and $\mathcal{E}$ are independent, in contrast to $\tilde{\mathcal{E}}^*_\ell$ and $\mathcal{E}$ that are dependent through ζ. This allows us to rewrite

$$\mathbb{E}_\ell \left[V^{\tilde{\mathcal{E}}^*_\ell}_{1,\pi^*_{\mathcal{E}}}(s^\ell_1) - V^{\tilde{\mathcal{E}}^*_\ell}_{1,\pi^\ell_{\mathrm{TS}}}(s^\ell_1) \right] = \mathbb{E}_\ell \left[V^{\tilde{\mathcal{E}}^*_\ell}_{1,\pi^*_{\mathcal{E}}}(s^\ell_1) - V^{\bar{\mathcal{E}}^*_\ell}_{1,\pi^*_{\mathcal{E}}}(s^\ell_1) \right] \tag{47}$$

Due to our construction of $\bar{\mathcal{E}}^*_\ell$ in Appendix C, we can substitute $\bar{\mathcal{E}}^*_\ell = \bar{\mathcal{E}}_\ell = \mathbb{E}_\ell[\mathcal{E}]$. Using Lemma 8, we can rewrite the above mean as

$$\begin{aligned}
= \sum_{h=1}^H \mathbb{E}_\ell \Bigg[\mathbb{E}^{\bar{\mathcal{E}}_\ell}_{\pi^*_{\mathcal{E}}} \Bigg[& \mathbb{E}_{(s',r')\sim(P^{\tilde{\mathcal{E}}^*_\ell}_h \otimes r^{\tilde{\mathcal{E}}^*_\ell}_h)(\cdot|s_h,a_h)} [r' + V^{\tilde{\mathcal{E}}^*_\ell}_{h+1,\pi^*_{\mathcal{E}}}(s')] \\
& - \mathbb{E}_{(s',r')\sim(P^{\bar{\mathcal{E}}_\ell}_h \otimes r^{\bar{\mathcal{E}}_\ell}_h)(\cdot|s_h,a_h)} [r' + V^{\tilde{\mathcal{E}}^*_\ell}_{h+1,\pi^*_{\mathcal{E}}}(s')] \Bigg] \Bigg] .
\end{aligned} \tag{48}$$

Denoting

$$\begin{aligned}
\Delta^{\tilde{\mathcal{E}}^*_\ell}_h(s_h,a_h) := & \mathbb{E}_{(s',r')\sim(P^{\tilde{\mathcal{E}}^*_\ell}_h \otimes r^{\tilde{\mathcal{E}}^*_\ell}_h)(\cdot|s_h,a_h)} [r' + V^{\tilde{\mathcal{E}}^*_\ell}_{h+1,\pi^*_{\mathcal{E}}}(s')] \\
& - \mathbb{E}_{(s',r')\sim(P^{\bar{\mathcal{E}}_\ell}_h \otimes r^{\bar{\mathcal{E}}_\ell}_h)(\cdot|s_h,a_h)} [r' + V^{\tilde{\mathcal{E}}^*_\ell}_{h+1,\pi^*_{\mathcal{E}}}(s')],
\end{aligned} \tag{49}$$

we have

$$= \sum_{h=1}^H \mathbb{E}_\ell \left[\int_{s,a} d^{\bar{\mathcal{E}}_\ell}_{h,\pi^*_{\mathcal{E}}}(s,a) \Delta^{\tilde{\mathcal{E}}^*_\ell}_h(s,a) \, \mathrm{d}\mu_{\mathcal{S}\times\mathcal{A}} \right] . \tag{50}$$

Let $\mathcal{B}_\ell := \{(s,a,h) \mid \mathbb{E}_\ell\left[d^{\bar{\mathcal{E}}_\ell}_{h,\pi^*_{\mathcal{E}}}(s,a) \right] \ne 0\}$ and let $\int_{(s,a,h)} := \sum_h \int_{(s,a)}$ denote the integral over the space $[H] \times \mathcal{S} \times \mathcal{A}$ where we use the product of counting measure on $[H]$ and $\mu_{\mathcal{S}\times\mathcal{A}}$. We apply Cauchy-Schwarz inequality using the similar technique in [18, App. A.2], which we modify to

include the value diameter (Definition 9). Since $\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a) \le 2H$, we have

$$
\mathbb{E}_\ell\left[\int_{(s,a,h)} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)\right]
$$

$$
= \mathbb{E}_\ell\left[\int_{(s,a,h)\notin\mathcal{B}_\ell} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)\right] + \mathbb{E}_\ell\left[\int_{(s,a,h)\in\mathcal{B}_\ell} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)\right]
$$

$$
\le 2H\mathbb{E}_\ell\left[\int_{(s,a,h)\notin\mathcal{B}_\ell} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\right] + \mathbb{E}_\ell\left[\int_{(s,a,h)\in\mathcal{B}_\ell} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)\right]
$$

$$
= \mathbb{E}_\ell\left[\int_{(s,a,h)\in\mathcal{B}_\ell} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)\right]
$$

$$
= \mathbb{E}_\ell\left[\int_{(s,a,h)\in\mathcal{B}_\ell} \frac{\lambda_\mathcal{E} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)}{\mathbb{E}_\ell\left[d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\right]^{1/2}}\mathbb{E}_\ell\left[d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\right]^{1/2}\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)}{\lambda_\mathcal{E}}\right]
$$

$$
\le \left(\mathbb{E}_\ell\left[\int_{(s,a,h)\in\mathcal{B}_\ell} \frac{(\lambda_\mathcal{E} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a))^2}{\mathbb{E}_\ell\left[d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\right]}\right]\right)^{1/2}
$$

$$
\left(\mathbb{E}_\ell\left[\int_{(s,a,h)\in\mathcal{B}_\ell} \mathbb{E}_\ell\left[d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\right]\left(\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)}{\lambda_\mathcal{E}}\right)^2\right]\right)^{1/2}
$$

$$
\le \left(\int_{(s,a,h)\in\mathcal{B}_\ell} \frac{\mathbb{E}_\ell\left[(\lambda_\mathcal{E} d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a))^2\right]}{\mathbb{E}_\ell\left[d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\right]}\right)^{1/2} \left(\mathbb{E}_\ell\left[\int_{(s,a,h)} \mathbb{E}_\ell\left[d_{h,\pi^*}^{\bar{\mathcal{E}}_\ell}(s,a)\right]\left(\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)}{\lambda_\mathcal{E}}\right)^2\right]\right)^{1/2}
$$

$$
= \sqrt{\mathcal{T}^\ell \cdot \mathcal{I}^\ell},
$$

where we used $\mathcal{T}^\ell$ and $\mathcal{I}^\ell$ to denote the first and the second term respectively. Note that the total regret of each episode is at most H. Therefore, going back to the Bayesian regret formulationa and using Cauchy-Schwarz, we get

$$
\mathfrak{BR}_L(\pi_{\mathrm{TS}}) \le \mathbb{E}\left[\sum_{\ell=L_0+1}^{L} \sqrt{\mathcal{T}^\ell \cdot \mathcal{I}^\ell}\right] + L_0 H + L\varepsilon \tag{51}
$$

$$
\le \mathbb{E}\left[\left(\sum_{\ell=L_0+1}^{L} \mathcal{T}^\ell\right)^{1/2} \cdot \left(\sum_{\ell=L_0+1}^{L} \mathcal{I}^\ell\right)^{1/2}\right] + L_0 H + L\varepsilon \tag{52}
$$

$$
\le \sqrt{\mathbb{E}\left[\sum_{\ell=L_0+1}^{L} \mathcal{T}^\ell\right] \cdot \mathbb{E}\left[\sum_{\ell=L_0+1}^{L} \mathcal{I}^\ell\right]} + L_0 H + L\varepsilon \tag{53}
$$

$$
\le \sqrt{L\left(\sup_{L_0+1\le\ell\le L} \mathbb{E}\left[\mathcal{T}^\ell\right]\right) \cdot \mathbb{E}\left[\sum_{\ell=1}^{L} \mathcal{I}^\ell\right]} + L_0 H + L\varepsilon, \tag{54}
$$

for every $0 \le L_0 < L$. We estimate each term separately in Appendix F and Appendix E.

Remark 11. While the spirit of the argument, in applying surrogate learning coupled with information ratio and Cauchy Schwarz is similar to [18], the technical aspects are different for estimating $\sum_\ell \mathcal{I}^\ell$, and more importantly, the entire analysis is different for $\sum_\ell \mathcal{T}^\ell$, as can be observed in what follows.

We gather the results to finish the proof. In Appendix E, we show that $\mathbb{E}[\sum_\ell \mathcal{I}^\ell] \le \frac{1}{2}\log(K_{\mathrm{surr}}(\varepsilon))$.

In Appendix F, we show that $\limsup \mathbb{E}\left[\mathcal{T}^\ell\right]$ is bounded by $\lambda^2 H$. Thus $\sup_{L_0+1 \le \ell \le L} \mathbb{E}\left[\mathcal{T}^\ell\right] \le 2\lambda^2 H$ for large enough $L_0 > 0$. Hence,

$$\mathfrak{BR}_L(\pi_{\mathrm{TS}}) \le \lambda\sqrt{\log(K_{\mathrm{surr}}(\varepsilon))T} + L\varepsilon + T_0 \tag{55}$$

for all $\varepsilon > 0$, where $T_0 = L_0 H$. Taking the infimum over ε gives the desired regret bound. The next statement of the theorem in Eq. (43) is an application of Lemma 1 followed by selecting $\varepsilon = 1/L$. Notice that due to nonzero ε effects, polylogarithmic terms in H, L are picked up when comparing d_{l_1} with $d_{l_1}(\varepsilon)$. More precisely, according to Lemma 1, one compares $\log(K_{\mathrm{surr}}(\varepsilon))$ with $\sum_h \log(L_h^P(\varepsilon/(2H)^2)) + \log(L_h^R(\varepsilon/(4H)))$. From the definition of $d_{l_1,h}^P$ we have $\log(L_h^P(\varepsilon/(2H)^2)) \sim O(d_{l_1,h}^P \log((2H)^2/\varepsilon))$, so this includes a logarithmic factor of H^2, and choosing $\varepsilon = 1/L$ means a logarithmic factor of $H^2 L$. A similar argument can be made for L_h^R.

Finally, the last statement of the theorem follows by definition: $d_{l_1} L = (\sum_h d_{l_1,h})L = d_{\mathrm{hom}} H L = d_{\mathrm{hom}} T$. This finishes the proof of the main theorem. $\square$

Remark 12. Equations similar to Eq. (46) can be found in [18, App. B.2]. However, a small correction must be made to their derivation. The authors first apply Cauchy-Schwarz and then take the square of Eq. (45) to replace the original regret with the surrogate regret. As seen later in the proof, we do it in the opposite order, because in (45), the left side may be negative, so we can not assume the square of that estimation to be also correct. While outside the focus of this paper, we note that a side-effect of this correction is another one to their definition of surrogate-IDS in [18, Eq. (4.3)], wherein minimization should be over the square root of their information ratio.

E   Bounding $\mathbb{E}[\sum_\ell \mathcal{I}^\ell]$

We start by proving $\mathcal{I}^\ell \le \frac{1}{2}\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H}), \forall \ell \in [L]$. We recall that TS property implies $\mathbb{E}_\ell[d_{h,\pi_{\mathcal{E}}^*}^{\bar{\mathcal{E}}_\ell}(s,a)] = \mathbb{E}_\ell[d_{h,\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}(s,a)]$. So,

$$\mathcal{I}^\ell = \sum_{h=1}^H \mathbb{E}_\ell\left[\int_{s,a} \mathbb{E}_\ell[d_{h,\pi_{\mathcal{E}}^*}^{\bar{\mathcal{E}}_\ell}(s,a)]\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)^2}{\lambda_{\mathcal{E}}^2}\,\mathrm{d}\mu_{\mathcal{S}\times\mathcal{A}}\right] \tag{56}$$

$$= \sum_{h=1}^H \mathbb{E}_\ell\left[\int_{s,a} \mathbb{E}_\ell[d_{h,\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}(s,a)]\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)^2}{\lambda_{\mathcal{E}}^2}\,\mathrm{d}\mu_{\mathcal{S}\times\mathcal{A}}\right]. \tag{57}$$

Next, we swap the two integrals, one represented by $\mathbb{E}_\ell$ and the one over s, a,

$$= \sum_{h=1}^H \int_{s,a} \mathbb{E}_\ell[d_{h,\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}(s,a)]\mathbb{E}_\ell\left[\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)^2}{\lambda_{\mathcal{E}}^2}\right]\,\mathrm{d}\mu_{\mathcal{S}\times\mathcal{A}} \tag{58}$$

Note that, given $\mathcal{D}_\ell$, $\Delta_h^{\tilde{\mathcal{E}}_\ell^*}$ is independent of $d_{h,\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}(s,a)$. Therefore, using the identity $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ for two independent random variables, and swapping back the two integrals,

$$= \sum_{h=1}^H \mathbb{E}_\ell\left[\int_{s,a} d_{h,\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}(s,a)\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)^2}{\lambda_{\mathcal{E}}^2}\,\mathrm{d}\mu_{\mathcal{S}\times\mathcal{A}}\right] = \sum_{h=1}^H \mathbb{E}_\ell\left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}\left[\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)^2}{\lambda_{\mathcal{E}}^2}\right]\right] \tag{59}$$

20

Finally, notice that $\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)^2}{\lambda_{\mathcal{E}}^2}$ can be estimated by Pinsker's inequality (Lemma 9) as

$$
\frac{\Delta_h^{\tilde{\mathcal{E}}_\ell^*}(s,a)^2}{\lambda_{\mathcal{E}}^2}
$$

$$
= \left(\mathbb{E}_{(s',r')\sim(P_h^{\tilde{\mathcal{E}}_\ell}\otimes r_h^{\tilde{\mathcal{E}}_\ell})(\cdot|s_h,a_h)} \left[\frac{r' + V_{h+1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s')}{\lambda_{\mathcal{E}}} \right] \right.
$$

$$
\left. - \mathbb{E}_{(s',r')\sim(P_h^{\bar{\mathcal{E}}_\ell}\otimes r_h^{\bar{\mathcal{E}}_\ell})(\cdot|s_h,a_h)} \left[\frac{r' + V_{h+1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s')}{\lambda_{\mathcal{E}}} \right] \right)^2
$$

$$
= \left(\mathbb{E}_{(s',r')\sim(P_h^{\tilde{\mathcal{E}}_\ell^*}\otimes r_h^{\tilde{\mathcal{E}}_\ell^*})(\cdot|s_h,a_h)} \left[\frac{r' + V_{h+1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s') - r_h^{\inf}(s_h,a_h) - \inf_s V_{h+1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s)}{\lambda_{\mathcal{E}}} \right] \right.
$$

$$
\left. - \mathbb{E}_{(s',r')\sim(P_h^{\bar{\mathcal{E}}_\ell}\otimes r_h^{\bar{\mathcal{E}}_\ell})(\cdot|s_h,a_h)} \left[\frac{r' + V_{h+1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s') - r_h^{\inf}(s_h,a_h) - \inf_s V_{h+1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s)}{\lambda_{\mathcal{E}}} \right] \right)^2
$$

$$
\leq \frac{1}{2} D_{\mathrm{KL}} \left((P_h^{\tilde{\mathcal{E}}_\ell^*}\otimes r_h^{\tilde{\mathcal{E}}_\ell^*})(\cdot|s_h,a_h) \| (P_h^{\bar{\mathcal{E}}_\ell}\otimes r_h^{\bar{\mathcal{E}}_\ell})(\cdot|s_h,a_h) \right)
$$

where we note the trick of adding and subtracting the constant term $r_h^{\inf}(s_h,a_h) + \inf_s V_{h+1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s)$ to the expected values in order to make the expression inside between zero and $\lambda_{\mathcal{E}}$. This enables the application of Pinsker's inequality which requires the random variable X in $(\mathbb{E}_P[X] - \mathbb{E}_Q[X])^2 \leq \frac{1}{2}D_{\mathrm{KL}}(P\|Q)$ to be smaller than one. Therefore

$$
\mathcal{I}^\ell \leq \frac{1}{2} \sum_{h=1}^{H} \mathbb{E}_\ell \left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell} \left[D_{\mathrm{KL}} \left((P_h^{\tilde{\mathcal{E}}_\ell^*}\otimes r_h^{\tilde{\mathcal{E}}_\ell^*})(\cdot|s_h,a_h) \| (P_h^{\bar{\mathcal{E}}_\ell}\otimes r_h^{\bar{\mathcal{E}}_\ell})(\cdot|s_h,a_h) \right) \right] \right]. \tag{60}
$$

Lastly, we use Lemma 10, wherein we show a fact similar to [18, App. C.1.] but for $\tilde{\mathcal{E}}_\ell^*$ instead of $\mathcal{E}$, proving the above equals $\frac{1}{2} \mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H})$.

Next, observe that $\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H}) \leq \mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\zeta; \mathcal{H}_{\ell,H})$. Indeed, recall that the definition of $\mathbb{I}_\ell$ conditions on a $\mathcal{D}_\ell$; and we know that conditional on a $\mathcal{D}_\ell$, the surrogate environment $\tilde{\mathcal{E}}_\ell^*$ is only dependent on ζ by construction, hence the data processing inequality applies and we have

$$
\mathcal{I}^\ell \leq \frac{1}{2} \mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\zeta; \mathcal{H}_{\ell,H}). \tag{61}
$$

Next, we use the mutual information chain rule, observing that

$$
\mathbb{E} \left[\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\zeta; \mathcal{H}_{\ell,H}) \right] = \mathbb{I}(\zeta; \mathcal{H}_{\ell,H} | \mathcal{H}_{\ell-1,H}, \ldots, \mathcal{H}_{1,H}), \tag{62}
$$

and therefore

$$
\mathbb{I}(\zeta; \mathcal{D}_{L+1}) = \mathbb{I}(\zeta; (\mathcal{H}_{1,H}, \ldots, \mathcal{H}_{L,H})) = \sum_{\ell=1}^{L} \mathbb{E} \left[\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\zeta; \mathcal{H}_{\ell,H}) \right]. \tag{63}
$$

Applied to the above, and noting that $\mathbb{I}(\zeta; \mathcal{D}_{L+1}) \leq H(\zeta) \leq \log(K_{\mathrm{surr}}(\varepsilon))$, this finishes our estimation of $\mathbb{E}[\sum_\ell \mathcal{I}^\ell] \leq \frac{1}{2}\log(K_{\mathrm{surr}}(\varepsilon))$.

F Bounding $\mathbb{E}[\mathcal{T}^\ell]$

This is where we use analysis tools from posterior consistency. We focus on bounding

$$
\mathcal{T}^\ell = \int_{(s,a,h)\in\mathcal{B}_\ell} \frac{\mathbb{E}_\ell \left[(\lambda_{\mathcal{E}} d_{h,\pi_{\mathcal{E}}^*}^{\bar{\mathcal{E}}_\ell}(s,a))^2 \right]}{\mathbb{E}_\ell \left[d_{h,\pi_{\mathcal{E}}^*}^{\bar{\mathcal{E}}_\ell}(s,a) \right]}. \tag{64}
$$

21

Note that we have

$$\mathbb{E}_\ell \left[d_{h,\pi_{\mathcal{E}}^*}^{\bar{\mathcal{E}}_\ell}(s,a) \right] = \mathbb{E}_\ell \left[d_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s,a) \right],$$

where $\mathcal{E}'$ is sampled from the posterior $\mathbb{P}_\ell$ independent of $\mathcal{E}$. Similarly

$$
\begin{aligned}
\mathbb{E}_\ell \left[(\lambda_{\mathcal{E}} d_{h,\pi_{\mathcal{E}}^*}^{\bar{\mathcal{E}}_\ell}(s,a))^2 \right] &= (\mathbb{E}_\ell)_{\mathcal{E}\sim\mathbb{P}_\ell} \left[\lambda_{\mathcal{E}}^2 \left((\mathbb{E}_\ell)_{\mathcal{E}'\sim\mathbb{P}_\ell} \left[d_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s,a) \right] \right)^2 \right] \\
&\leq (\mathbb{E}_\ell)_{\mathcal{E}\sim\mathbb{P}_\ell} \left[\lambda_{\mathcal{E}}^2 (\mathbb{E}_\ell)_{\mathcal{E}'\sim\mathbb{P}_\ell} \left[d_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s,a)^2 \right] \right] \\
&= \mathbb{E}_\ell \left[(\lambda_{\mathcal{E}} d_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s,a))^2 \right].
\end{aligned}
$$

For any ℓ, s, a, h and $\mathcal{D}_\ell$, we define

$$g_\ell(s,a,h,\mathcal{D}_\ell) := \frac{\mathbb{E}_\ell \left[(\lambda_{\mathcal{E}} d_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s,a))^2 \right]}{\mathbb{E}_\ell \left[d_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s,a) \right]}, \tag{65}$$

whenever $(s,a,h) \in \mathcal{B}_\ell$ and $g_\ell(s,a,h,\mathcal{D}_\ell) := 0$ otherwise. Clearly we have

$$\mathcal{T}^\ell \leq \int_{(s,a,h)} g_\ell(s,a,h,\mathcal{D}_\ell). \tag{66}$$

Moreover, given our assumption on state action occupation density, we have $M_d :=$ $\sup_{s,a,h,\pi,\mathcal{E}} d_{h,\pi}^{\mathcal{E}}(s,a) < \infty$, which implies $g_\ell \leq M_d H^2 < \infty$ and $\mathcal{T}^\ell \leq M_d H^3 < \infty$. Let $\mathcal{E}_0$ be the true environment. This means $\mathbb{E}[\cdot|\mathcal{E}_0] = \mathbb{E}_{\mathcal{D}_\ell\sim\mathbb{P}(\cdot|\mathcal{E}_0)}[\cdot]$.

According to Corollary 12, we have

$$\lim_{\ell\to\infty} \mathbb{E}_\ell \left[d_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s,a) \right] = d_{h,\pi_{\mathcal{E}_0}^*}^{\mathcal{E}_0}(s,a), \tag{67}$$

$$\lim_{\ell\to\infty} \mathbb{E}_\ell \left[(\lambda_{\mathcal{E}} d_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s,a))^2 \right] = (\lambda_{\mathcal{E}_0} d_{h,\pi_{\mathcal{E}_0}^*}^{\mathcal{E}_0}(s,a))^2. \tag{68}$$

According to Assumption 2, $d_{h,\pi_{\mathcal{E}_0}^*}^{\mathcal{E}_0}(s,a) \neq 0$ for almost every $\mathcal{E}_0, s, a$ and h. For any such values of $(\mathcal{E}_0, s, a, h)$ and almost every $\mathcal{D}_\ell$ sampled from true environment $\mathcal{E}_0$, we conclude that

$$\lim_{\ell\to\infty} g_\ell(s,a,h,\mathcal{D}_\ell) = \lambda_{\mathcal{E}_0}^2 d_{h,\pi_{\mathcal{E}_0}^*}^{\mathcal{E}_0}(s,a). \tag{69}$$

Therefore, using dominated convergence theorem, for almost every $\mathcal{E}_0$ we have

$$\lim_{\ell\to\infty} \mathbb{E}[\mathcal{T}^\ell|\mathcal{E}_0] \leq \lim_{\ell\to\infty} \mathbb{E}_{\mathcal{D}_\ell\sim\mathbb{P}(\cdot|\mathcal{E}_0)} \left[\int_{(s,a,h)} g_\ell(s,a,h,\mathcal{D}_\ell) \right] \tag{70}$$

$$= \lim_{\ell\to\infty} \int_{(s,a,h)} \mathbb{E}_{\mathcal{D}_\ell\sim\mathbb{P}(\cdot|\mathcal{E}_0)} \left[g_\ell(s,a,h,\mathcal{D}_\ell) \right] \tag{71}$$

$$= \int_{(s,a,h)} \mathbb{E}_{\mathcal{D}_\ell\sim\mathbb{P}(\cdot|\mathcal{E}_0)} \left[\lim_{\ell\to\infty} g_\ell(s,a,h,\mathcal{D}_\ell) \right] \tag{72}$$

$$= \int_{(s,a,h)} \mathbb{E}_{\mathcal{D}_\ell\sim\mathbb{P}(\cdot|\mathcal{E}_0)} \left[\lambda_{\mathcal{E}_0}^2 d_{h,\pi_{\mathcal{E}_0}^*}^{\mathcal{E}_0}(s,a) \right] \tag{73}$$

$$= \int_{(s,a,h)} \lambda_{\mathcal{E}_0}^2 d_{h,\pi_{\mathcal{E}_0}^*}^{\mathcal{E}_0}(s,a) \tag{74}$$

$$= \lambda_{\mathcal{E}_0}^2 H. \tag{75}$$

Therefore we may use dominated convergence theorem again to see that

$$\lim_{\ell\to\infty} \mathbb{E}[\mathcal{T}^\ell] = \lim_{\ell\to\infty} \mathbb{E}[\mathbb{E}[\mathcal{T}^\ell|\mathcal{E}_0]] = \mathbb{E}[\lim_{\ell\to\infty} \mathbb{E}[\mathcal{T}^\ell|\mathcal{E}_0]] \leq \mathbb{E}[\lambda_{\mathcal{E}_0}^2 H] = \mathbb{E}[\lambda_{\mathcal{E}}^2] H = \lambda^2 H. \tag{76}$$

Therefore, there exists $L_0 > 0$ such that $\mathbb{E}[\mathcal{T}^\ell] \leq 2\lambda^2 H$ for $\ell > L_0$.

G   Proof of Corollary 5

We restate the corollary below. We prove it in a more general case where the maps ϕ^P, ϕ^R are time inhomogeneous, i.e. ϕ_h^P, ϕ_h^R. In addition, the dimension of their target space can also depend on h, i.e. we use $d_f^{P,h}, d_f^{R,h}$, instead of just d_f^P, d_f^R. For the case where the dimensions are homogeneous, we use d_f^{hom}. This new notation impacts the statement as follows:

Corollary. *For a linear Bayesian RL, for large enough T,*

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{d_{l_1}^f T}). \tag{77}$$

Given a linear Bayesian RL with finitely many states and homogeneous feature space dimension d_f^{hom}, we have $d_{l_1}^f \leq 2d_f^{\text{hom}} HS$, yielding for large enough T,

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{H d_f^{\text{hom}} ST}). \tag{78}$$

Given a mixture linear Bayesian RL, for large enough T,

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{MT}), \tag{79}$$

The first statement follows from the generic bound, but we need to relate the l_1-dimension of the environment space to that of the feature maps space, where we recall the definitions $d_{l_1}^f = d_{l_1}^{P,f} + d_{l_1}^{R,f}$ as the sum of the l_1-dimensions of the feature map space $\{\psi_h^{P,\mathcal{E}}\}_{\mathcal{E}\in\Theta}, \{\psi_h^{R,\mathcal{E}}\}_{\mathcal{E}\in\Theta}$ where the l_1-distance between feature maps is defined as $l_1(\psi_h^{\mathcal{E}}, \psi_h^{\mathcal{E}'}) = \int_s \|\psi_h^{\mathcal{E}} - \psi_h^{\mathcal{E}'}\|_1 \mu_{\mathcal{S}}$. We shall use Fact 1:

$$V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s_1^\ell) \leq H\sum_{h=1}^{H} l_1(P_h^{\mathcal{E}}, P_h^{\mathcal{E}'}) + \sum_{h=1}^{H} l_1(r_h^{\mathcal{E}}, r_h^{\mathcal{E}'}). \tag{80}$$

We estimate

$$l_1(P_h^{\mathcal{E}}, P_h^{\mathcal{E}'}) = \sup_{s,a}\int_{s'}|P_h^{\mathcal{E}}(s'|s,a) - P_h^{\mathcal{E}'}(s'|s,a)| = \sup_{s,a}\int_{s'}|\phi_h^P(s,a)\cdot(\psi_h^{P,\mathcal{E}}(s') - \psi_h^{P,\mathcal{E}}(s'))| \leq$$

$$\sup_{s,a}\int_{s'}\sum_{i=1}^{d_{f,h}^P}|\phi_h^P(s,a)_i(\psi_h^{P,\mathcal{E}}(s') - \psi_h^{P,\mathcal{E}}(s'))_i| \tag{81}$$

Since $\|\phi_h^P(s,a)\|_2 \leq 1 \implies |\phi_h^P(s,a)_i| \leq 1, \forall i \in [d]$. Therefore

$$l_1(P_h^{\mathcal{E}}, P_h^{\mathcal{E}'}) \leq \int_{s'}\|(\psi_h^{P,\mathcal{E}}(s') - \psi_h^{P,\mathcal{E}}(s'))\|_1 = l_1(\psi_h^{P,\mathcal{E}}, \psi_h^{P,\mathcal{E}}). \tag{82}$$

The similar bound can be achieved for $l_1(r_h^{\mathcal{E}}, r_h^{\mathcal{E}'})$. As a result $d_{l_1} \leq d_{l_1}^f$ and we get the first statement.

For Eq. (78), we note that if S is a finite, then $\psi_h^{P,\mathcal{E}}$ can be viewed as a $d_f^{P,h} \times S$ matrix, or simply a vector with dimension that size. Therefore, we can view our problem as asking for the asymptotics of the $\varepsilon-$covering number in $\mathbb{R}^{d_f^{P,h}S}$. As long as the collection $\{\psi_h^{P,\mathcal{E}}\}_{\mathcal{E}}$ is within a finite ball, which they are (Definition 4), the covering number is well-known to scale at most as $O((\frac{\log(C_\psi)}{\varepsilon})^{d_f^{P,h}S})$ where C_ψ is the radius of that ball. Applying the similar argument for the rewards, we have $d_{l_1} \leq d_{l_1}^f \leq \sum(d_f^{P,h} + d_f^{R,h})S$, which equals $2Hd_f^{\text{hom}}S$ given a homogeneous feature space dimension $d_f^{\text{hom}} = d_f^{P,h} = d_f^{R,h}$, for all $h \in [H]$.

The finite mixtures statement is a straightforward generalization of the above, where every $\psi_h^{P,\mathcal{E}}, \psi_h^{R,\mathcal{E}}$ is characterized with a finite m_h^P, m_h^R-dimensional vector instead of specifically being $d_f^{P,h}S, d_f^{R,h}S$-dimensional. We note

$$l_1(\psi_h^{P,\mathcal{E}}, \psi_h^{P,\mathcal{E}}) = \int_{s'}\|(\psi_h^{P,\mathcal{E}}(s') - \psi_h^{P,\mathcal{E}}(s'))\|_1 \leq C_\Psi\|\boldsymbol{a}_h^{P,\mathcal{E}} - \boldsymbol{a}_h^{P,\mathcal{E}'}\|_1 \tag{83}$$

where $C_\Psi = \max_{1\leq i\leq m_h^P}\|\Psi_{h,i}^P\|_1$. So the same argument above applies, where we consider the collection of finite dimensional vectors $\{\boldsymbol{a}_h^{P,\mathcal{E}}\}_{\mathcal{E}}$ and the $\varepsilon-$covering number, and similarly for $\{\boldsymbol{a}_h^{R,\mathcal{E}}\}_{\mathcal{E}}$.

G.1 Incorrect proof for the Bayesian regret bound of linear Bayesian RL with deterministic rewards

Here, we discuss the proof of [18] for linear RLs, and the mistakes in their argument. We start by citing the similar equations in [18, App. B.4] in bounding the value difference with the feature maps difference:

For any $\mathcal{E}_1, \mathcal{E}_2 \in \Theta_k$, [...]

$$V_{1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_1}(s_1) - V_{1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_2}(s_1)$$

$$= \sum_{h=1}^{H} \mathbb{E}_{\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_2} \left[P_h^{\mathcal{E}_1}(\cdot|s_h^\ell, a_h^\ell)^\top V_{h+1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_1}(\cdot) - P_h^{\mathcal{E}_2}(\cdot|s_h^\ell, a_h^\ell)^\top V_{h+1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_1}(\cdot) \right]$$

$$= \sum_{h=1}^{H} \mathbb{E}_{\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_2} \left[\phi(s_h^\ell, a_h^\ell)^\top \sum_{s'} V_{h+1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_1}(s') \psi_h^{\mathcal{E}_1}(s') \right.$$

$$\left. - \phi(s_h^\ell, a_h^\ell)^\top \sum_{s'} V_{h+1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_1}(s') \psi_h^{\mathcal{E}_2}(s') \right],$$

[...] Moreover, since the value function is always bounded by H, we have

$$V_{1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_1}(s_1) - V_{1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_2}(s_1)$$

$$= H \sum_{h=1}^{H} \mathbb{E}_{\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_2} \left[\phi(s_h^\ell, a_h^\ell)^\top \left(\sum_{s'} \psi_h^{\mathcal{E}_1}(s') - \sum_{s'} \psi_h^{\mathcal{E}_2}(s') \right) \right]$$

$$\leq H \sum_{h=1}^{H} \mathbb{E}_{\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_2} \left[\left\| \phi(s_h^\ell, a_h^\ell) \right\|_2 \right] \left\| \sum_{s'} \psi_h^{\mathcal{E}_1}(s') - \sum_{s'} \psi_h^{\mathcal{E}_2}(s') \right\|_2$$

$$\leq H \sum_{h=1}^{H} \left\| \sum_{s'} \psi_h^{\mathcal{E}_1}(s') - \sum_{s'} \psi_h^{\mathcal{E}_2}(s') \right\|_2. \tag{84}$$

Clearly, in the first equation, the equality must be replaced by $\leq$, and more importantly, given that the value function $V_{h+1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_1}(\cdot)$ has argument $\cdot = s'$, the l_2-norm should be taken on the **inside** of the integral $\int_{s'} \| \phi(s_h^\ell, a_h^\ell)^\top \left(\psi_h^{\mathcal{E}_1}(s') - \psi_h^{\mathcal{E}_2}(s') \right) \|_2$ (we note we also replaced the sum $\sum_{s'}$ with integral, as linear RLs could have infinitely many states). If our proposed correction were to be followed then the next equations would change to ones similar to ours except with an l_2-distance:

$$\leq H \sum_{h=1}^{H} \mathbb{E}_{\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_2} \left[\left\| \phi(s_h^\ell, a_h^\ell) \right\|_2 \right] \int_{s'} \left\| \psi_h^{\mathcal{E}_1}(s') - \psi_h^{\mathcal{E}_2}(s') \right\|_2 \tag{85}$$

$$\leq H \sum_{h=1}^{H} \int_{s'} \left\| \psi_h^{\mathcal{E}_1}(s') - \psi_h^{\mathcal{E}_2}(s') \right\|_2 \tag{86}$$

Otherwise, let us assume that the authors were correct, then we have managed to bound $V_{1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_1}(s_1) - V_{1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_2}(s_1) \leq H \sum_{h=1}^{H} \left\| \sum_{s'} \psi_h^{\mathcal{E}_1}(s') - \sum_{s'} \psi_h^{\mathcal{E}_2}(s') \right\|_2$. We show that this is a bound by zero for an important subclass of linear RLs, i.e. all tabular RLs.

It is a well-known fact that tabular RLs can be viewed as linear RLs. The mapping works as follows. First let us enumerate the set $\{(s,a) \in \mathcal{S} \times \mathcal{A}\}$ by $1, \ldots, SA$. Call this assignment $N(s,a) \in [SA]$. Then define $\phi(s,a) = e_{N(s,a)} \in \mathbb{R}^{SA}$, which is the Euclidean basis state on axis $N(s,a)$. Let $\psi_h^{\mathcal{E}}(s') = (P_h^{\mathcal{E}}(s'|s,a))_{s,a}$. Then clearly all the conditions $\|\phi\|_2 \leq 1, \|\sum_{s'} \psi(s')\|_2 \leq C_\psi$ and most importantly $P_h(\cdot|s,a) = \langle \phi(s,a), \psi_h^{\mathcal{E}}(s') \rangle$, are satisfied. However we note that for any $\mathcal{E}, h$ we have $\sum_{s'} \psi_h^{\mathcal{E}}(s') = (\sum_{s'} P_h^{\mathcal{E}}(s'|s,a))_{s,a} = (1)_{s,a}$ which is the all one vector in $\mathbb{R}^{SA}$. In that case, $\sum_{s'} \psi_h^{\mathcal{E}}(s') - \sum_{s'} \psi_h^{\mathcal{E}'}(s')$ in Eq. (84) is the zero vector, with zero norm.

Therefore, were the estimation in [18, App. B.4] correct, for all tabular Bayesian RLs with deterministic reward, the difference of all value functions of the form $V_{1,\pi_{\mathcal{E}_1}^*}^{\mathcal{E}_1} - V_{1,\pi_{\mathcal{E}_2}^*}^{\mathcal{E}_2}$ would be bounded above by zero, meaning we have estimated $K_{\mathrm{surr}}(\varepsilon) = 1$ for all ε, which since $\log(1) = 0$, implies a constant regret bound as well. This counterexample further demonstrates the mistake above.

Overall, this makes the proof for [18, Theorem 4.10] incorrect and invalidates their claim of a regret bound $\widetilde{O}(d_f^{\mathrm{hom}} H^{3/2}\sqrt{T})$.

Remark 13. Another important gap in the proof of [18, Theorem 4.10] can be found in [18, App. B.5], where the surrogate regret is claimed to be bounded by a conditional mutual information by π^* instead of π_{TS}^ℓ. This is explained in further details in Appendix J.1.

H   Proof of Corollary 6

We restate the corollary.

Corollary. *Given a finite mixtures Bayesian RL problem, for large enough T,*

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}(\lambda\sqrt{d_{l_1}^m T}) . \tag{87}$$

Assuming the restricted finite mixtures model, for large enough T,

$$\mathfrak{BR}_L(\pi_{TS}) \leq \widetilde{O}\left(\lambda\sqrt{MT}\right) . \tag{88}$$

which, given a uniform dimension $m = m_h^P = m_h^R$, yields $\widetilde{O}(\lambda\sqrt{HmT})$.

Given Theorem 3 and Fact 1, we need to estimate the d_{l_1} of Θ by that of $\{a_h^{P,\mathcal{E}}(s,a)\}_{\mathcal{E}\in\Theta}$ and $\{a_h^{R,\mathcal{E}}(s,a)\}_{\mathcal{E}\in\Theta}$. Indeed writing the l_1-distance of two transition functions:

$$\sup_{s,a}\|P_h^{\mathcal{E}}(\cdot|s,a) - P_h^{\mathcal{E}'}(\cdot|s,a)\|_1 = \sup_{s,a}\|\sum_{i=1}^{m_h}(a_{h,i}^{P,\mathcal{E}}(s,a) - a_{h,i}^{P,\mathcal{E}'}(s,a))Z_{h,i}^P(\cdot|s,a)\|_1 \leq$$
$$\sup_{s,a}\sum_{i=1}^{m_h}\|(a_{h,i}^{P,\mathcal{E}}(s,a) - a_{h,i}^{P,\mathcal{E}'}(s,a))Z_{h,i}^P(\cdot|s,a)\|_1 = \sup_{s,a}\sum_{i=1}^{m_h}|a_{h,i}^{P,\mathcal{E}}(s,a) - a_{h,i}^{P,\mathcal{E}'}(s,a)| = \tag{89}$$
$$\sup_{s,a}\|a_h^{P,\mathcal{E}}(s,a) - a_h^{P,\mathcal{E}'}(s,a)\|_1$$

where we used the triangle inequality and the fact that the density functions are positive and their integral equals one.

For the second statement, we are faced with the problem of finding an l_1-covering number for a collection of vectors on the m_h-dimensional simplex. It is a standard fact that the covering of the latter is of order $O\left((\frac{1}{\varepsilon})^{m_h}\right)$, implying Eq. (88).

I   Useful Lemmas

Lemma 8. *For any two environments $\mathcal{E}, \mathcal{E}'$ with potentially different transition and reward functions, and any policy π, we have*

$$V_{1,\pi}^{\mathcal{E}}(s_1) - V_{1,\pi}^{\mathcal{E}'}(s_1)$$
$$= \sum_{h=1}^{H}\mathbb{E}_\pi^{\mathcal{E}'}\left[\mathbb{E}_{s'\sim P_h^{\mathcal{E}}(\cdot|s_h,a_h)}[V_{h+1,\pi}^{\mathcal{E}}(s')] - \mathbb{E}_{s'\sim P_h^{\mathcal{E}'}(\cdot|s_h,a_h)}[V_{h+1,\pi}^{\mathcal{E}}(s')]\right]$$
$$+ \sum_{h=1}^{H}\mathbb{E}_\pi^{\mathcal{E}'}[r_h^{\mathcal{E}}(s_h,a_h) - r_h^{\mathcal{E}'}(s_h,a_h)]$$
$$= \sum_{h=1}^{H}\mathbb{E}_\pi^{\mathcal{E}'}\left[\mathbb{E}_{(s',r')\sim(P_h^{\mathcal{E}}\otimes r_h^{\mathcal{E}})(\cdot|s_h,a_h)}[r' + V_{h+1,\pi}^{\mathcal{E}}(s')]\right.$$
$$\left. - \mathbb{E}_{(s',r')\sim(P_h^{\mathcal{E}'}\otimes r_h^{\mathcal{E}'})(\cdot|s_h,a_h)}[r' + V_{h+1,\pi}^{\mathcal{E}}(s')]\right],$$

where $V_{H+1,\pi^*}^{\mathcal{E}}(\cdot) := 0$ and the expectation $\mathbb{E}_\pi^{\mathcal{E}'}$ is with respect to s_h, a_h.

Note that when rewards are deterministic, we have

$$V_{1,\pi}^{\mathcal{E}}(s_1) - V_{1,\pi}^{\mathcal{E}'}(s_1) = \sum_{h=1}^{H} \mathbb{E}_\pi^{\mathcal{E}'} \left[\mathbb{E}_{s' \sim P_h^{\mathcal{E}}(\cdot|s_h,a_h)}[V_{h+1,\pi}^{\mathcal{E}}(s')] - \mathbb{E}_{s' \sim P_h^{\mathcal{E}'}(\cdot|s_h,a_h)}[V_{h+1,\pi}^{\mathcal{E}}(s')] \right],$$

which is the statements of [18, Lemma D.3].

Proof. We have

$$V_{1,\pi}^{\mathcal{E}}(s_1) - V_{1,\pi}^{\mathcal{E}'}(s_1) = (V_{1,\pi}^{\mathcal{E}}(s_1) - V_{1,\pi}^{\mathcal{E}'_r\mathcal{E}}(s_1)) + (V_{1,\pi}^{\mathcal{E}'_r\mathcal{E}}(s_1) - V_{1,\pi}^{\mathcal{E}'}(s_1)) \tag{90}$$

where $\mathcal{E}'_{r\mathcal{E}}$ is an environment with the transition functions of $\mathcal{E}'$ but reward functions of $\mathcal{E}$. The first term above may be rewritten using [18, Lemma D.3]. The second term may be rewritten using the direct definition of value function as $V_{1,\pi}^{\mathcal{G}} = \mathbb{E}_\pi^{\mathcal{G}}[\sum_{h=1}^{H} r^{\mathcal{G}}(s_h, a_h)]$, which completes the proof. $\square$

Pinsker's lemma is at the center of relating the two concepts of regret and mutual information. We cite the following variant of the Pinsker's inequality from Fact 9 in [35].

Lemma 9. *For any distribution P and Q such that P is absolutely continuous with respect to Q, any random variable $X : \Omega \to \mathcal{X}$ and any $g : \mathcal{X} \to \mathbb{R}$ such that $\sup g - \inf g \leq 1$, we have*

$$\mathbb{E}_P[g(x)] - \mathbb{E}_Q[g(x)] \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P||Q)}. \tag{91}$$

J   Mutual information of surrogate environment and history

Recall that by performing the information ratio trick, Cauchy-Schwarz and Pinsker's inequality, we obtained the following term in our bound of the squared regret:

$$\frac{1}{2} \sum_{h=1}^{H} \mathbb{E}_\ell \left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell} \left[D_{\mathrm{KL}} \left((P_h^{\tilde{\mathcal{E}}_\ell^*} \otimes r_h^{\tilde{\mathcal{E}}_\ell^*})(\cdot|s_{h-1}^\ell, a_{h-1}^\ell) || (P_h^{\bar{\mathcal{E}}_\ell} \otimes r_h^{\bar{\mathcal{E}}_\ell})(\cdot|s_{h-1}^\ell, a_{h-1}^\ell) \right) \right] \right].$$

Now we would like to show that the above is $\frac{1}{2} \mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H})$. To be more careful in our arguments, we need to be reminded of what the random variable $\mathcal{H}_{\ell,H}$ is. We must view it as $\mathcal{H}_{\ell,H} = \mathcal{H}_{\ell,H}(\mathcal{E}, \pi_{\mathrm{TS}}^\ell)$ or $\mathcal{H}_{\ell,H}(\mathcal{E}, \pi_{\mathcal{E}_{\mathrm{TS}}}^*)$ where $\mathcal{E}, \mathcal{E}_{\mathrm{TS}}$ are two independent samples of $\mathbb{P}_\ell(\cdot)$, and $\mathcal{E}$ represents the same $\mathcal{E}$ in the regret above in $\pi_{\mathcal{E}}^*$, i.e. the true environment. Also note that $\mathcal{E}, \tilde{\mathcal{E}}_\ell^*$ are dependent, as we set $\tilde{\mathcal{E}}_\ell^*, \mathcal{E}$ to have the same ζ value.

Lemma 10. *With $\tilde{\mathcal{E}}_\ell^*$ defined according to the proof in Appendix C, in a Bayesian RL, we have*

$$\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H}) = \sum_{h=1}^{H} \mathbb{E}_\ell \left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell} \left[D_{\mathrm{KL}} \left((P_h^{\tilde{\mathcal{E}}_\ell^*} \otimes r_h^{\tilde{\mathcal{E}}_\ell^*})(\cdot|s_{h-1}^\ell, a_{h-1}^\ell) || (P_h^{\bar{\mathcal{E}}_\ell} \otimes r_h^{\bar{\mathcal{E}}_\ell})(\cdot|s_{h-1}^\ell, a_{h-1}^\ell) \right) \right] \right]$$

As a special case, when rewards are deterministic, we have

$$\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H}) = \sum_{h=1}^{H} \mathbb{E}_\ell \left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell} \left[D_{KL}(P_h^{\tilde{\mathcal{E}}_\ell^*}(\cdot|s_h, a_h) || P_h^{\bar{\mathcal{E}}_\ell}(\cdot|s_h, a_h)) \right] \right].$$

Proof. Using the chain rule of mutual information,

$$\begin{aligned}
\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H}) &= \sum_{h=1}^{H} \mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell} \left(\tilde{\mathcal{E}}_\ell^*; (s_h^\ell, a_h^\ell, r_h^\ell) | \mathcal{H}_{\ell,h-1} \right) \\
&= \sum_{h=1}^{H} \mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell} \left(\tilde{\mathcal{E}}_\ell^*; s_h^\ell | \mathcal{H}_{\ell,h-1} \right) + \sum_{h=1}^{H} \mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell} \left(\tilde{\mathcal{E}}_\ell^*; a_h^\ell | s_h^\ell, \mathcal{H}_{\ell,h-1} \right) \tag{92} \\
&\quad + \sum_{h=1}^{H} \mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell} \left(\tilde{\mathcal{E}}_\ell^*; r_h^\ell | s_h^\ell, a_h^\ell, \mathcal{H}_{\ell,h-1} \right).
\end{aligned}$$

Let us note what is meant by $\mathbb{I}_\ell^{\pi_{\text{TS}}^\ell}(\cdot)$ is $\mathbb{I}_\ell(\cdot|\pi_{\text{TS}}^\ell)$. In [18], the policy π used is fixed/independent from the random variables involved in the mutual information (given $\mathcal{D}_\ell$). Here, the same holds as π_{TS}^ℓ and $\tilde{\mathcal{E}}_\ell^*$ are independent.

- For the first term in Eq. (92), by using $\mathbb{I}(X;Y) = \int D_{\text{KL}}(P(Y|x)||P(Y))\,\mathrm{d}\mathbb{P}(x)$, and the definition of conditional mutual information, we have

$$\mathbb{I}_\ell\left(\tilde{\mathcal{E}}_\ell^*; s_h^\ell | \mathcal{H}_{\ell,h-1}, \pi_{\text{TS}}^\ell\right)$$

$$= \int \int D_{\text{KL}}\left(\mathbb{P}_\ell\left(s_h^\ell = \cdot | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*, \tilde{\mathcal{E}}_\ell^*\right) || \mathbb{P}_\ell\left(s_h^\ell = \cdot | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*\right)\right)$$

$$\mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}_\ell^* | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*)\,\mathrm{d}\mathbb{P}_\ell(\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*) \tag{93}$$

$$= \int \int D_{\text{KL}}\left(P_h^{\tilde{\mathcal{E}}_\ell^*}\left(\cdot | s_{h-1}^\ell, a_{h-1}^\ell\right) || \mathbb{P}_\ell\left(s_h^\ell = \cdot | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*\right)\right)$$

$$\mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}_\ell^* | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*)\,\mathrm{d}\mathbb{P}_\ell(\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*).$$

Where we substituted $\mathbb{P}_\ell\left(s_h^\ell = \cdot | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*, \tilde{\mathcal{E}}_\ell^*\right) = P_h^{\tilde{\mathcal{E}}_\ell^*}\left(\cdot | s_{h-1}^\ell, a_{h-1}^\ell\right)$. Let us see why this is the case. Let us analyze the meaning of the conditional on $(\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*, \tilde{\mathcal{E}}_\ell^*)$. Recall that $\mathcal{H}_{\ell,h-1} = \mathcal{H}_{\ell,h-1}(\mathcal{E}, \pi_{\mathcal{E}_{\text{TS}}}^*)$. Since $\tilde{\mathcal{E}}_\ell^*$ is given, the random variable $\mathcal{E}$ can only go over the partition $\Theta_{\zeta(\tilde{\mathcal{E}}_\ell^*)}^\varepsilon$. Of course, we can also drop all conditionals on previous state transitions, except for the last one $s_{h-1}^\ell(\mathcal{E}), a_{h-1}^\ell(\mathcal{E})$. Note that the policy $\pi_{\mathcal{E}_{\text{TS}}}^*$ is also irrelevant in this conditional, since the next state only depends on probability transitions and not on policy, hence why also we are not using the full notation $s_{h-1}^\ell(\mathcal{E}, \pi_{\mathcal{E}_{\text{TS}}}^*), a_{h-1}^\ell(\mathcal{E}, \pi_{\mathcal{E}_{\text{TS}}}^*)$. This implies

$$\mathbb{P}_\ell\left(s_h^\ell(\mathcal{E}) = \cdot | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*, \tilde{\mathcal{E}}_\ell^*\right) = \mathbb{P}_\ell\left(s_h^\ell(\mathcal{E}) = \cdot | s_{h-1}^\ell(\mathcal{E}), a_{h-1}^\ell(\mathcal{E}), \tilde{\mathcal{E}}_\ell^*\right) \tag{94}$$

$$= \int_\mathcal{E} P_h^\mathcal{E}\left(s_h^\ell = \cdot | s_{h-1}^\ell, a_{h-1}^\ell\right)\mathrm{d}\mathbb{P}_\ell(\mathcal{E}|\zeta(\mathcal{E}) \tag{95}$$

$$= \zeta(\tilde{\mathcal{E}}_\ell^*)) \tag{96}$$

$$= P_h^{\mathbb{E}_\ell[\mathcal{E}|\zeta(\mathcal{E})=\zeta(\tilde{\mathcal{E}}_\ell^*)]}\left(s_h^\ell = \cdot | s_{h-1}^\ell, a_{h-1}^\ell\right) \tag{97}$$

However, recall that we defined $\tilde{\mathcal{E}}_\ell^*$ to be the posterior mean of $\mathcal{E}$ over Θ_k^ε, i.e. $\mathbb{E}_\ell[\mathcal{E}|\zeta(\mathcal{E}) = \zeta(\tilde{\mathcal{E}}_\ell^*)] = \tilde{\mathcal{E}}_\ell^*$. Hence, the average above yields $P_h^{\tilde{\mathcal{E}}_\ell^*}\left(\cdot | s_{h-1}^\ell, a_{h-1}^\ell\right)$, as desired. Next, for the second term in the KL-divergence,

$$\mathbb{P}_\ell\left(s_h^\ell(\mathcal{E}) = \cdot | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*\right) = \int \mathbb{P}_\ell\left(s_h^\ell(\mathcal{E}) = \cdot | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*, \mathcal{E}\right)\mathrm{d}\mathbb{P}_\ell(\mathcal{E}|\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*)$$

$$= \int P_h^\mathcal{E}(s_h^\ell = \cdot | s_{h-1}^\ell, a_{h-1}^\ell)\mathrm{d}\mathbb{P}_\ell(\mathcal{E}|\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*)$$

$$= \int P_h^\mathcal{E}(s_h^\ell = \cdot | s_{h-1}^\ell, a_{h-1}^\ell)\mathrm{d}\mathbb{P}_\ell(\mathcal{E})$$

$$= P_h^{\bar{\mathcal{E}}_\ell}\left(s_h^\ell = \cdot | s_{h-1}^\ell, a_{h-1}^\ell\right). \tag{98}$$

In the above equations, $\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*$ are given in the conditional, and the true environment $\mathcal{E}$ is being integrated. The second equality was explained in the previous case. Let us explain why $\mathrm{d}\mathbb{P}_\ell(\mathcal{E}|\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*) = \mathrm{d}\mathbb{P}_\ell(\mathcal{E})$ in the third equality. Due to the independence of priors over different layers, the conditional on $\mathcal{H}_{\ell,h-1}$ impacts transition functions of prior layers (i.e. $P_1^\mathcal{E}, \ldots, P_{h-1}^\mathcal{E}$), while the transition function in question is the one at layer h. Therefore, this conditional can be dropped, as well as $\pi_{\mathcal{E}_{\text{TS}}}^*$ since $\mathcal{E}_{\text{TS}}, \mathcal{E}$ are two independent samples of $\mathbb{P}_\ell$. Finally, the last equation is by the definition of probability kernel $P_h^{\bar{\mathcal{E}}_\ell}$. Eqs. (93) and (98) imply $\mathbb{I}_\ell^{\pi_{\text{TS}}^\ell}\left(\tilde{\mathcal{E}}_\ell^*; s_h^\ell | \mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*\right) =$

$$\int \int D_{\text{KL}}\left(P_h^{\tilde{\mathcal{E}}_\ell^*}(\cdot | s_{h-1}^\ell, a_{h-1}^\ell) || P_h^{\bar{\mathcal{E}}_\ell}(\cdot | s_{h-1}^\ell, a_{h-1}^\ell)\right)\mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}_\ell^*)\,\mathrm{d}\mathbb{P}_\ell(\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}_{\text{TS}}}^*). \tag{99}$$

where we note we also dropped the conditionals on $\mathcal{H}_{\ell,h-1}, \pi^*_{\mathcal{E}_{\mathrm{TS}}}$ in $\mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}^*_\ell)$, by the similar argument in the previous case for $\mathrm{d}\mathbb{P}_\ell(\mathcal{E}|\mathcal{H}_{\ell,h-1}, \pi^*_{\mathcal{E}_{\mathrm{TS}}})$ as the integrand is transitions at the $h-$th step. We continue by focusing on the outer integral with respect to

$$\mathrm{d}\mathbb{P}_\ell(\mathcal{H}_{\ell,h-1}, \pi^*_{\mathcal{E}_{\mathrm{TS}}}) = P(\mathcal{H}_{\ell,h-1}(\mathcal{E}, \pi^*_{\mathcal{E}_{\mathrm{TS}}})|\mathcal{E}, \pi^*_{\mathcal{E}_{\mathrm{TS}}}) \, \mathrm{d}\mu^{\otimes(h-1)}_{\mathcal{S}\times\mathcal{A}} \, \mathrm{d}\mathbb{P}_\ell(\mathcal{E}) \, \mathrm{d}\mathbb{P}_\ell(\pi^*_{\mathcal{E}_{\mathrm{TS}}}) \quad (100)$$

and note that since only transitions at the $(h-1)$-th step are inside the inner integral, one can marginalize prior $(h-2,\ldots,1)$ state-action-reward tuples, yielding

$$= \int_{s^\ell_{h-1}, a^\ell_{h-1}, \mathcal{E}, \pi^*_{\mathcal{E}_{\mathrm{TS}}}} P(s^\ell_{h-1}, a^\ell_{h-1}|\mathcal{E}, \pi^*_{\mathcal{E}_{\mathrm{TS}}})$$

$$\left(\int_{\tilde{\mathcal{E}}^*_\ell} D_{\mathrm{KL}}\left(P^{\tilde{\mathcal{E}}^*_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})||P^{\bar{\mathcal{E}}_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})\right) \mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}^*_\ell)\right)$$

$$\mathrm{d}\mu_{\mathcal{S}\times\mathcal{A}} \, \mathrm{d}\mathbb{P}_\ell(\mathcal{E}) \, \mathrm{d}\mathbb{P}_\ell(\pi^*_{\mathcal{E}_{\mathrm{TS}}})$$

$$= \int_{s^\ell_{h-1}, a^\ell_{h-1}, \pi^*_{\mathcal{E}_{\mathrm{TS}}}, \tilde{\mathcal{E}}^*_\ell} \left(\int_{\mathcal{E}} P(s^\ell_{h-1}, a^\ell_{h-1}|\mathcal{E}, \pi^*_{\mathcal{E}_{\mathrm{TS}}}) \, \mathrm{d}\mathbb{P}_\ell(\mathcal{E})\right)$$

$$D_{\mathrm{KL}}\left(P^{\tilde{\mathcal{E}}^*_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})||P^{\bar{\mathcal{E}}_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})\right) \mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}^*_\ell, \pi^*_{\mathcal{E}_{\mathrm{TS}}}) \, \mathrm{d}\mu_{\mathcal{S}\times\mathcal{A}}$$

where we simply rearranged the measures and integrals, and note the independence $\mathbb{P}_\ell(\tilde{\mathcal{E}}^*_\ell)\mathbb{P}_\ell(\pi^*_{\mathcal{E}_{\mathrm{TS}}}) = \mathbb{P}_\ell(\tilde{\mathcal{E}}^*_\ell, \pi^*_{\mathcal{E}_{\mathrm{TS}}})$. For the outer integral, notice that $P(s^\ell_{h-1}, a^\ell_{h-1}|\mathcal{E}, \pi^*_{\mathcal{E}_{\mathrm{TS}}}) = d^{\mathcal{E}}_{h,\pi^*_{\mathcal{E}_{\mathrm{TS}}}}(s^\ell_{h-1}, a^\ell_{h-1})$ by definition. So using the linearity of expectation and independence of priors over different layers

$$\int_{\mathcal{E}} d^{\mathcal{E}}_{h,\pi^*_{\mathcal{E}_{\mathrm{TS}}}}(s^\ell_{h-1}, a^\ell_{h-1}) \, \mathrm{d}\mathbb{P}_\ell(\mathcal{E}) = d^{\bar{\mathcal{E}}_\ell}_{h,\pi^*_{\mathcal{E}_{\mathrm{TS}}}}(s^\ell_{h-1}, a^\ell_{h-1}). \quad (101)$$

Putting it all together, and going back to the notation $\pi^*_{\mathcal{E}_{\mathrm{TS}}} \to \pi^\ell_{\mathrm{TS}}$:

$$\mathbb{I}_\ell\left(\tilde{\mathcal{E}}^*_\ell; s^\ell_h|\mathcal{H}_{\ell,h-1}, \pi^\ell_{\mathrm{TS}}\right)$$

$$= \int_{s^\ell_{h-1}, a^\ell_{h-1}, \pi^\ell_{\mathrm{TS}}, \tilde{\mathcal{E}}^*_\ell} d^{\bar{\mathcal{E}}_\ell}_{h,\pi^\ell_{\mathrm{TS}}}(s^\ell_{h-1}, a^\ell_{h-1})$$

$$D_{\mathrm{KL}}\left(P^{\tilde{\mathcal{E}}^*_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})||P^{\bar{\mathcal{E}}_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})\right) \mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}^*_\ell, \pi^\ell_{\mathrm{TS}}) \, \mathrm{d}\mu_{\mathcal{S}\times\mathcal{A}}$$

$$= \int_{\pi^\ell_{\mathrm{TS}}, \tilde{\mathcal{E}}^*_\ell} \mathbb{E}^{\bar{\mathcal{E}}_\ell}_{\pi^\ell_{\mathrm{TS}}}\left[D_{\mathrm{KL}}\left(P^{\tilde{\mathcal{E}}^*_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})||P^{\bar{\mathcal{E}}_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})\right)\right] \mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}^*_\ell, \pi^\ell_{\mathrm{TS}})$$

$$= \mathbb{E}_\ell\left[\mathbb{E}^{\bar{\mathcal{E}}_\ell}_{\pi^\ell_{\mathrm{TS}}}\left[D_{\mathrm{KL}}\left(P^{\tilde{\mathcal{E}}^*_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})||P^{\bar{\mathcal{E}}_\ell}_h(\cdot|s^\ell_{h-1}, a^\ell_{h-1})\right)\right]\right] ,$$

where $\mathbb{E}^{\bar{\mathcal{E}}_\ell}_{\pi^\ell_{\mathrm{TS}}}$ is taken with respect to $s^\ell_{h-1}, a^\ell_{h-1}$ and $\mathbb{E}_\ell$ is taken with respect to $\pi^\ell_{\mathrm{TS}}, \tilde{\mathcal{E}}^*_\ell$.

- For the second term in Eq. (92),

$$\mathbb{I}_\ell\left(\tilde{\mathcal{E}}^*_\ell; a^\ell_h|s^\ell_h, \mathcal{H}_{\ell,h-1}, \pi^\ell_{\mathrm{TS}}\right)$$

$$= \int\int D_{\mathrm{KL}}\left(\mathbb{P}_\ell\left(a^\ell_h = \cdot|s^\ell_h, \mathcal{H}_{\ell,h-1}, \pi^\ell_{\mathrm{TS}}, \tilde{\mathcal{E}}^*_\ell\right)||\mathbb{P}_\ell\left(a^\ell_h = \cdot|s^\ell_h, \mathcal{H}_{\ell,h-1}, \pi^\ell_{\mathrm{TS}}\right)\right) .$$

where the integrals are with respect to $\mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}^*_\ell|s^\ell_h, \mathcal{H}_{\ell,h-1}, \pi^\ell_{\mathrm{TS}}) \, \mathrm{d}\mathbb{P}_\ell(s^\ell_h, \mathcal{H}_{\ell,h-1}, \pi^\ell_{\mathrm{TS}})$. When $s^\ell_h, \pi^\ell_{\mathrm{TS}}$ are given, both sides of the KL term are equal to $\pi^\ell_{\mathrm{TS}}(\cdot|s^\ell_h)$ and thus the above is zero.

- For the third term, we use an argument similar to the first term to see that

$$\mathbb{I}_\ell\left(\tilde{\mathcal{E}}^*_\ell; r^\ell_h|s^\ell_h, a^\ell_h, \mathcal{H}_{\ell,h-1}, \pi^\ell_{\mathrm{TS}}\right) = \mathbb{E}_\ell\left[\mathbb{E}^{\bar{\mathcal{E}}_\ell}_{\pi^\ell_{\mathrm{TS}}}\left[D_{\mathrm{KL}}\left(r^{\tilde{\mathcal{E}}^*_\ell}_h(\cdot|s^\ell_h, a^\ell_h)||r^{\bar{\mathcal{E}}_\ell}_h(\cdot|s^\ell_h, a^\ell_h)\right)\right]\right] .$$

$$(102)$$

Put together, it follows that

$$
\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}\left(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H}\right)
$$

$$
= \sum_{h=1}^{H} \mathbb{E}_\ell\left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}\left[D_{\mathrm{KL}}\left(P_h^{\tilde{\mathcal{E}}_\ell^*}(\cdot|s_{h-1}^\ell, a_{h-1}^\ell)||P_h^{\bar{\mathcal{E}}_\ell}(\cdot|s_{h-1}^\ell, a_{h-1}^\ell)\right)\right]\right]
$$

$$
+ \mathbb{E}_\ell\left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}\left[D_{\mathrm{KL}}\left(r_h^{\tilde{\mathcal{E}}_\ell^*}(\cdot|s_h^\ell, a_h^\ell)||r_h^{\bar{\mathcal{E}}_\ell}(\cdot|s_h^\ell, a_h^\ell)\right)\right]\right]
$$

$$
= \sum_{h=1}^{H} \mathbb{E}_\ell\left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}\left[D_{\mathrm{KL}}\left((P_h^{\tilde{\mathcal{E}}_\ell^*} \otimes r_h^{\tilde{\mathcal{E}}_\ell^*})(\cdot|s_{h-1}^\ell, a_{h-1}^\ell)||(P_h^{\bar{\mathcal{E}}_\ell} \otimes r_h^{\bar{\mathcal{E}}_\ell})(\cdot|s_{h-1}^\ell, a_{h-1}^\ell)\right)\right]\right]. \quad \square
$$

J.1 On the rewrite of mutual information in [18, App. B.5]

We start by citing the relevant equations involved in [18, App. B.5]. For $\Sigma_h = \mathbb{E}_\ell\left[\mathbb{E}_{\pi^*}^{\bar{\mathcal{E}}_\ell^*}\left[\phi(s_h^\ell, a_h^\ell)\right]\mathbb{E}_{\pi^*}^{\bar{\mathcal{E}}_\ell^*}\left[\phi(s_h^\ell, a_h^\ell)^\top\right]\right]$, the authors claim

$$
\sum_{h=1}^{H} \mathbb{E}_\ell\left[\left\|\Sigma_h^{1/2}\sum_{s'}(\psi_h^{\tilde{\mathcal{E}}_\ell^*}(s') - \psi_h^{\bar{\mathcal{E}}_\ell^*}(s'))V_{h+1,\pi^*}^{\tilde{\mathcal{E}}_\ell^*}(s')\right\|_2^2\right]
\tag{103}
$$

$$
= \mathbb{E}_\ell\left[\sum_{h=1}^{H}\mathbb{E}_{\pi^*}^{\bar{\mathcal{E}}_\ell}\left[\left(P_h^{\tilde{\mathcal{E}}_\ell^*}(\cdot|s_h^\ell, a_h^\ell)^\top V_{h+1,\pi^*}^{\tilde{\mathcal{E}}_\ell^*}(\cdot) - P_h^{\bar{\mathcal{E}}_\ell^*}(\cdot|s_h^\ell, a_h^\ell)^\top V_{h+1,\pi^*}^{\tilde{\mathcal{E}}_\ell^*}(\cdot)\right)^2\right]\right]
\tag{104}
$$

$$
\leq \frac{1}{2}\sum_{h=1}^{H} \mathbb{E}_\ell\left[\mathbb{E}_{\pi_{\mathcal{E}}^*}^{\bar{\mathcal{E}}_\ell}\left[D_{\mathrm{KL}}\left(P_h^{\tilde{\mathcal{E}}_\ell^*}(\cdot|s_{h-1}^\ell, a_{h-1}^\ell)||P_h^{\bar{\mathcal{E}}_\ell}(\cdot|s_{h-1}^\ell, a_{h-1}^\ell)\right)\right]\right]
\tag{105}
$$

$$
= \frac{1}{2}\mathbb{I}_\ell^{\pi^*}\left(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H}\right).
\tag{106}
$$

For the last part in Eq. (105), the authors do not provide a proof, and cite their own [18, Lemma A.1] as support. However, that lemma is for $\mathbb{I}_\ell^\pi(\mathcal{E}; \mathcal{H}_{\ell,H})$, where π is the **algorithm** and not the optimal policy of the true environment.

We need a rewrite of $\mathbb{I}_\ell^{\pi_{\mathcal{E}}^*}\left(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H}\right)$. We emphasize that in the former mutual information expression, the policy involved is the algorithm π, which clearly is not dependent on the true environment $\mathcal{E}$, unlike $\pi_{\mathcal{E}}^*$. Furthermore, the environment involved is also independent from the policy, but that is not the case here since $\tilde{\mathcal{E}}_\ell^*, \pi_{\mathcal{E}}^*$ are dependent through ζ. As we shall see, it is crucial for the policy in the mutual information expression to be independent from the true environment, in order for the argument in [18, Lemma A.1].

Since our own lemma above for $\mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell}\left(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H}\right)$ naturally extends [18, Lemma A.1] and takes the first step for the substitution of $\mathcal{E}$ by $\tilde{\mathcal{E}}_\ell^*$, we can analyze what happens in our own equations in Lemma 10, assuming we were to take $\pi_{\mathcal{E}}^*$, the optimal policy of the true environment $\mathcal{E}$, instead of π_{TS}^ℓ. We can apply the mutual information chain rule as before, and focus on

$$
\mathbb{I}_\ell\left(\tilde{\mathcal{E}}_\ell^*; s_h^\ell|\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}}^*\right) = \int\int D_{\mathrm{KL}}\left(\mathbb{P}_\ell(s_h^\ell = \cdot|\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}}^*, \tilde{\mathcal{E}}_\ell^*)||\mathbb{P}_\ell\left(s_h^\ell = \cdot|\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}}^*\right)\right)
$$

$$
\mathrm{d}\mathbb{P}_\ell(\tilde{\mathcal{E}}_\ell^*|\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}}^*)\,\mathrm{d}\mathbb{P}_\ell(\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}}^*)
$$

Recall that the history is of the form $\mathcal{H}_{\ell,H} = \mathcal{H}_{\ell,H}(\mathcal{E}, \pi_{\mathcal{E}_{\mathrm{TS}}}^*)$. The first thing to prove above should be $\mathbb{P}_\ell\left(s_h^\ell = \cdot|\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}}^*, \tilde{\mathcal{E}}_\ell^*\right) = P_h^{\tilde{\mathcal{E}}_\ell^*}\left(\cdot|s_{h-1}^\ell, a_{h-1}^\ell\right)$. However, note that since $\pi_{\mathcal{E}}^*$ is given, this means $\mathcal{E}$ is given, at least in (realistic) scenarios where there is uniqueness of optimal policies, and as a result, the true environment $\mathcal{E}$ in $\mathcal{H}_{\ell,H}(\mathcal{E}, \pi_{\mathcal{E}_{\mathrm{TS}}}^*)$ is determined uniquely. This implies that in fact $\mathbb{P}_\ell\left(s_h^\ell = \cdot|\mathcal{H}_{\ell,h-1}, \pi_{\mathcal{E}}^*, \tilde{\mathcal{E}}_\ell^*\right) = P_h^{\mathcal{E}}\left(s_h^\ell = \cdot|s_{h-1}^\ell, a_{h-1}^\ell\right)$. In general it would be the average $P_h^{\mathbb{E}[\mathcal{E}'|\pi_{\mathcal{E}'}^* = \pi_{\mathcal{E}}^*, \zeta(\mathcal{E}') = \zeta(\tilde{\mathcal{E}}_\ell^*)]}\left(s_h^\ell = \cdot|s_{h-1}^\ell, a_{h-1}^\ell\right)$.

Either way, in the very first step, we have shown that the dependence of the policy with the true environment can alter significantly the rewrite of the mutual information by the argument in [18, Lemma A.1]. Clearly, this does not lead to the desired rewrite in Eq. (105) and makes this claimed bound of the Bayesian regret by that mutual information (at the very least) unproven.

A more direct way to note the gap in the argument is the following. Recall that the denominator in the (surrogate) information ratio is supposed to represent the information gain by the algorithm on the true (or surrogate) environment. The surrogate mutual information ratio that one should bound is:

$$\frac{(\mathbb{E}_\ell \left[V_{1,\pi_{\tilde{\mathcal{E}}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) - V_{1,\pi}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) \right])^2}{\mathbb{I}_\ell^\pi \left(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H} \right)}, \tag{107}$$

where π is the algorithm (and we select $\pi = \pi_{\text{TS}}$). Clearly the algorithm can not know about the true environment $\mathcal{E}$, which makes it questionable to try to bound the surrogate regret $\mathbb{E}_\ell \left[V_{1,\pi_{\tilde{\mathcal{E}}}^*}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) - V_{1,\pi}^{\tilde{\mathcal{E}}_\ell^*}(s_1^\ell) \right]$ by a mutual information such as $\mathbb{I}_\ell^{\pi_{\tilde{\mathcal{E}}}^*} \left(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H} \right) := \mathbb{I}_\ell \left(\tilde{\mathcal{E}}_\ell^*; \mathcal{H}_{\ell,H} | \pi_{\mathcal{E}}^* \right)$, where there is assumed knowledge of the true environment in the conditional, as opposed to conditioning on the algorithm itself like in the ratio above. Therefore, the information ratio that the authors in [18, App. B.5] are (implicitly) trying to bound is not the right one.

K Posterior consistency

In this section we define the notion of posterior consistency and state Doob's consistency theorem. We start by describing posterior consistency in a general setting.

Let $\mathscr{X}$ be a measure space and for every $n \in \mathbb{N}$, let $X^{(n)}$ be an observation in the sample space $\mathscr{X}^n$ with distribution $P_\theta^{(n)}$ indexed by a parameter θ belonging to a separable metric space Ω. For instance $X^{(n)}$ might be a sample of size n from a given distribution P_θ with $P_\theta^{(n)}$ the corresponding product measure. Given a prior Π on the Borel sets of Ω, let $\Pi_n(\cdot \mid X^{(n)})$ be the posterior distribution given the observation $X^{(n)}$. Moreover, we assume that there is a measure $P_\theta^{(\infty)}$ on $\mathscr{X}^\infty$ such that $P_\theta^{(n)}$ is equal to the the image $P_\theta^{(\infty)} \circ (X^{(n)})^{-1}$ of the probability measure $P_\theta^{(\infty)}$ when pushed forward onto $\mathscr{X}^{(n)}$. We say an estimator $T := (T_n)_{n=1}^\infty$, where $T_n : \mathscr{X}^{(n)} \to \Omega$ is a measureable function for all $n \geq 1$, is a strongly consistent estimator of θ if for every $\theta_0 \in \Omega$ and almost every $X^{(\infty)} = (X^{(n)})_{n=1}^\infty$, we have

$$\lim_{n \to \infty} T_n(X^{(n)}) = \theta_0.$$

We can now describe the content of Assumption 1. Let $\Omega := \Theta$ with the measure $\Pi := \rho$ as the prior, and let $\mathscr{X}$ be the space of all single-episode histories. Also let $\theta_0 := \mathcal{E}_0$ and $P_{\theta_0}^{(\ell)}(\mathcal{D}_\ell) := \mathbb{P}(\mathcal{D}_\ell \mid \mathcal{E}_0)$ be the probability of observing the history $\mathcal{D}_\ell$ in the true environment $\mathcal{E}_0$. Existence of the measure $P_\theta^{(\infty)}$ on $\mathscr{X}^\infty$ as described above follows from the fact that for any $l' > l$, we have $P_{\theta_0}^{(\ell')}(\mathcal{D}_\ell) = P_{\theta_0}^{(\ell)}(\mathcal{D}_\ell) := \mathbb{P}(\mathcal{D}_\ell \mid \mathcal{E}_0)$. Assumption 1 states that there exists a strongly consistent estimator of the true environment T such that for almost every environment $\mathcal{E}_0$ and almost every infinite history $\mathcal{D} = (\mathcal{D}_\ell)_{\ell=1}^\infty$ sampled from the environment $\mathcal{E}_0$, we have

$$\lim_{n \to \infty} T_\ell(\mathcal{D}_\ell) = \mathcal{E}_0.$$

The existence of consistent estimators is closely related to the notion of posterior consistency:

Definition 10. The posterior distribution $\Pi_n(\cdot \mid X^{(n)})$ is said to be strongly consistent at $\theta_0 \in \Omega$ if for every neighbourhood U of θ_0 and $P_{\theta_0}^{(\infty)}$-almost every $X^{(\infty)}$, we have $\Pi_n(U^c \mid X^{(n)}) \to 0$ where $X^{(n)}$ is the projection of $X^{(\infty)}$ into the space $\mathscr{X}^n$.

Here we state a version of Doob's consistency theorem that we need for our application. (Theorem 6.9 in [17])

Theorem 11 (Doob's consistency theorem). *If there is a strongly consistent estimator $T_n : \mathscr{X}^{(n)} \to \Omega$, then the posterior is strongly consistent at Π-almost every $\theta \in \Omega$. In fact, $\int f(\theta')d\Pi_n(\theta' \mid X^{(n)}) \to f(\theta)$, almost surely $[P_\theta^{(\infty)}]$, for Π-almost every θ and every Π-integrable function f.*

Note that while this statement is not the exact statement of Theorem 6.9 in [17], it is equivalent to it as discussed in the paragraph following the theorem.

Corollary 12. *Given Assumption 1, for any Π-integrable function $f : \Theta \to \mathbb{R}$ and almost every $\mathcal{D}_\infty$ sampled from true environment $\mathcal{E}_0$, we have*

$$\lim_{\ell \to \infty} \mathbb{E}_\ell[f(\mathcal{E})] = f(\mathcal{E}_0).$$

Similarly, if $f : \Theta \times \Theta \to \mathbb{R}$ is bounded and $(\Pi \times \Pi)$-integrable, for almost every $\mathcal{D}_\infty$ sampled from true environment $\mathcal{E}_0$, we have

$$\lim_{\ell \to \infty} \mathbb{E}_\ell[f(\mathcal{E}, \mathcal{E}')] = f(\mathcal{E}_0, \mathcal{E}_0),$$

where the expectation is taken over all values of $\mathcal{E}$ and $\mathcal{E}'$, sampled according to $\mathbb{P}_\ell$.

Proof. The first statement immediately follows from Assumption 1 and Theorem 11. To prove the second part, we use the first part to see that for any fixed value of $\mathcal{E}' \in \Theta$ and almost every $\mathcal{D}_\infty$, we have

$$\lim_{\ell \to \infty} (\mathbb{E}_\ell)_{\mathcal{E} \sim \mathbb{P}_\ell}[f(\mathcal{E}, \mathcal{E}')] = f(\mathcal{E}_0, \mathcal{E}').$$

Now we use dominated convergence theorem to see that

$$\lim_{\ell \to \infty} (\mathbb{E}_\ell)_{\mathcal{E}, \mathcal{E}' \sim \mathbb{P}_\ell}[f(\mathcal{E}, \mathcal{E}')] = \lim_{\ell \to \infty} (\mathbb{E}_\ell)_{\mathcal{E}' \sim \mathbb{P}_\ell}[f(\mathcal{E}_0, \mathcal{E}')] = f(\mathcal{E}_0, \mathcal{E}_0). \qquad \square$$