Prior-dependent analysis of posterior sampling reinforcement learning with function approximation

Yingru Li   **Zhi-Quan Luo**

The Chinese University of Hong Kong, Shenzhen, China

Shenzhen Research Institute of Big Data

Abstract

This work advances randomized exploration in reinforcement learning (RL) with function approximation modeled by linear mixture MDPs. We establish the first prior-dependent Bayesian regret bound for RL with function approximation; and refine the Bayesian regret analysis for posterior sampling reinforcement learning (PSRL), presenting an upper bound of $\tilde{\mathcal{O}}(d\sqrt{H^3T\log T})$, where d represents the dimensionality of the transition kernel, H the planning horizon, and T the total number of interactions. This signifies a methodological enhancement by optimizing the $\mathcal{O}(\sqrt{\log T})$ factor over the previous benchmark (Osband and Van Roy, 2014) specified to linear mixture MDPs. Our approach, leveraging a value-targeted model learning perspective, introduces a decoupling argument and a variance reduction technique, moving beyond traditional analyses reliant on confidence sets and concentration inequalities to formalize Bayesian regret bounds more effectively.

1 Introduction

Reinforcement learning (RL) has become a cornerstone in the development of intelligent systems (Sutton and Barto, 2018), propelling the forefront of artificial intelligence and decision sciences towards creating agents capable of making autonomous decisions in dynamically complex environments. The crux of advancing RL has been the strategic integration of function approximation techniques and the assimilation of prior knowledge. Function approximation methods

have transcended traditional RL boundaries (Bertsekas and Tsitsiklis, 1996), enabling scalable solutions across extensive or continuous state spaces, a necessity for practical real-world problem-solving. Concurrently, the infusion of prior knowledge—ranging from domain expertise to insights gleaned from historical data (Wang et al., 2018, 2019; Agarwal et al., 2020) or pre-trained models (Yang et al., 2023)—into RL algorithms has catalyzed a transformative leap in learning efficiency. This leap is underscored by providing RL systems with a foundational understanding of advantageous actions, even before any environment interaction occurs.

Motivation. The nuanced application of priors within the RL paradigm, especially through Bayesian methodologies, ushers in a sophisticated equilibrium between exploration and exploitation—essential for the practical deployment of RL. By preemptively incorporating knowledge about the environment's dynamics through prior distributions, RL algorithms are primed for more informed exploration. Despite its immense potential, the exploration of priors' role, particularly within the context of function approximation, remains scant. This uncharted domain presents a ripe opportunity for enhancing RL's learning outcomes and applicability in diverse settings, thus pushing the envelope of RL's capabilities and efficiency.

Important Question. Against this backdrop, we are compelled to address a pivotal inquiry:

> *How can the symbiosis of prior knowledge and function approximation be optimized to elevate the adaptability and efficiency of RL algorithms?*

1.1 Key Contributions

Our exploration into this inquiry yields several seminal contributions, especially within the context of linear mixture Markov Decision Processes:

- The introduction of a prior-dependent Bayesian

regret bound (Theorem 1) within the realm of RL with function approximation, elucidating the impact of prior distribution variance on learning efficiency. This milestone underscores a deeper comprehension of RL dynamics in scenarios where environmental distributions are known or can be approximated.

- The unveiling of a prior-free Bayesian regret bound for Posterior Sampling for Reinforcement Learning (PSRL) in linear mixture MDPs, i.e., $\mathcal{O}(dH^{3/2}\sqrt{T \log T})$ Bayesian regret bound for PSRL in linear mixture MDPs (Remark 1), where d is the dimension of feature in basis transition kernels, H is the planning horizon and T is the number of interactions. Our upper bound improves the Bayesian regret bound of PSRL by traditional by a $O(\log T)$ factor.

- The advancement of the analytical landscape of Bayesian regret in RL through methodological novelties, including a decoupling lemma and a variance reduction theorem. These innovations engender a nuanced perspective of regret dynamics, advancing beyond the conventional frameworks reliant on confidence sets and concentration inequalities.

1.2 Preview of Technical Novelty

Our theoretical contributions are underpinned by substantial technical innovations:

Posterior variance reduction. Our analysis features a variance reduction theorem that addresses the heterogeneity of value variance in RL, highlighting a predictable reduction in posterior variance under specific conditions. from the value-targeted perspective of model learning in Section 4. In expectation, the posterior variance of the true model parameters will be reduced in a non-uniform way due to the heteroscedasticity nature of the value variance. The posterior variance is predictable when we have access to the true posterior distribution under correct prior and correct likelihood distribution. Our posterior variance reduction argument provides a way to measure how epistemic uncertainty is reduced after new observations are informed in the learning progresses.

Decoupling argument. We introduce a decoupling argument (Lemmas 5 and 7) that enhances our understanding of Bayesian regret by separating the intertwined dynamics of action selection and value estimation and relate the Bayesian regret to posterior variance. The idea of decoupling is inspired by information-theoretic analysis (Russo and Van Roy, 2016; Kalkanlı and Özgür, 2020) for the linear bandit. The statement in Lemma 7 is slightly stronger than

previous literature, which is essential to prove a better H dependence. In details, this stronger statement is required to upper bound the *absolute value-targeted error* in Proposition 2 and consequently upper bound the *absolute estimation error* in Proposition 3. Extending the previous statement in bandit literature to this stronger statement is not complicated but new.

Integration of prior knowledge: We provide a unique characterization of the relationship between the regret bound and the prior distribution, offering new insights into integrating prior knowledge into RL algorithms. In contrast to previous analysis that utilizes confidence sets, we leverage the decoupling argument and variance reduction argument to upper bound the Bayesian regret, essential for deriving the prior-dependent analysis. Previous Bayesian analysis in RL (Osband and Van Roy, 2014, 2017) relies on a translation of frequentist concentration bound to a Bayesian credit interval, which is essentially a prior-free argument.

Besides, as a by-product of our analysis for the prior-free result, because of going beyond traditional analysis relying on the confidence-sets or the notion of eluder dimension (Osband and Van Roy, 2014; Ayoub et al., 2020) to upper bound regret, the final regret bound will be naturally improved by a $O(\log T)$ factor[1].

1.3 Related Works

Our endeavor distinctly positions itself within the vast expanse of RL research by concentrating on the synergistic integration of prior knowledge and function approximation in linear mixture MDPs. Contrasting with preceding works that predominantly focus on tabular contexts or marginally engage with the potential of prior knowledge, our comprehensive Bayesian analysis introduces both prior-dependent and prior-free regret bounds. This dual-faceted approach meticulously fills a significant lacuna in existing literature, shedding light on the efficacious incorporation of prior knowledge into RL strategies.

Linear function approximation. Linear mixture MDPs (Ayoub et al., 2020) become a widely-accepted benchmark to understand the synergy of exploration and model-based function approximation. Many algorithms including value-targeted regression (Ayoub et al., 2020) and policy optimization (Cai et al., 2020) are developed for understand the statistical complexity of reinforcement learning with linear function approximation under this setup. Existing algorithm for linear mixture MDPs that achieves

---

[1]We also discuss a conjecture on the $\log T$ term in the lower bound Appendix G.

the near-minimax optimal dependency[2] is based on optimism in the face of uncertainty (OFU) principle (Zhou et al., 2021). For another setup of linear MDPs (Yang and Wang, 2019; Jin et al., 2020), a line of works are proposed to understand value-based function approximation and exploration (Jin et al., 2020; Agarwal et al., 2023).

Randomized exploration. An alternative category of algorithms (Strens, 2000), inspired by Thompson sampling (Thompson, 1933), involves randomized exploration that samples a set of statistically plausible action values and selects the maximizing action. Practically, randomized exploration shows promising computational and statistical advantages (Chapelle and Li, 2011; Osband et al., 2019; Li et al., 2022, 2024). Empirical success has prompted a surge of interest in theoretical analysis for randomized exploration, e.g. randomized least-square value iteration (RLSVI) (Zanette et al., 2020; Osband et al., 2019) and its optimistic sampling (Ishfaq et al., 2021) or approximate sampling variant (Ishfaq et al., 2024; Li et al., 2024). These analyses in linear MDPs mostly rely on Azuma–Hoeffding concentration inequality that is unaware of value variance. For linear mixture MDPs, no specific frequentist analyses of randomized exploration exist. Notably, the HyperAgent (Li et al., 2024) demonstrates state-of-the-art practical efficiency as well as provable sublinear regret and scalable per-step computation, bridging theory and practice.

Bayesian regret analysis. The Bayesian regret analysis in (Osband and Van Roy, 2017; Lu and Van Roy, 2019; Li et al., 2024) requires independent Dirichlet prior imposed on the transition probabilities, tailored for tabular MDPs. Lu et al. (2021) requires the independent Beta prior on the transition probabilities since they only analyze a special ring-structured MDPs provided a conjecture holds. Existing Bayesian analysis for RL with function approximation did not give detailed characterization on relationship between regret bound and the prior distribution. Osband and Van Roy (2014) provide the prior-free Bayesian regret for RL with function approximation, borrowing the notion of eluder dimension (Russo and Van Roy, 2014). Specifically, Osband and Van Roy (2014) gives a Bayesian regret bound $\widetilde{\mathcal{O}}(\mathbb{E}\left[K^*\right]\sqrt{d_K d_E H T \log T})$ for PSRL with general function approximation, where K^*, d_K and d_E are the global Lipschitz constant for the future value function, the Kolmogorov and the Eluder dimensions of the model class. In the case of

linear mixture MDPs, this Bayesian regret becomes $\widetilde{\mathcal{O}}(d\sqrt{H^3 T}\log T)$ [3]. In contrast to the existing analyses, our analysis provides the first prior-dependent Bayesian analysis of PSRL and an improved prior-free bound under function approximation; see Theorem 1 and Remark 1.

2 Preliminaries

We consider the problem where the RL agent learns to optimize the *finite horizon MDP* over repeated episodes of interactions. There are two sources of randomness in this repeated process: the *environmental randomness* and the *algorithmic randomness*.

Finite horizon MDP. A finite horizon Markov Decision Process (MDP) [4] (Puterman, 2014) is described by $M = (\mathcal{S}, \mathcal{A}, P, R, H, \rho)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the horizon. In the beginning of an episode, an initial state $s_0 \in \mathcal{S}$ is sampled from ρ. At stage $0 \leqslant h \leqslant H-1$, given state s_h and action a_h, the agent observes a reward r_{h+1} sampled from the reward distribution with mean $R(h, s_h, a_h) \in [0, 1]$ and a next state s_{h+1} sampled from transition kernel $P(h, s_h, a_h)$.

A deterministic policy $\pi = \{\pi_h\}_{h=0}^{H-1}$ is a collection of H functions, each of which maps from a state $s_h \in \mathcal{S}$ to an action $a_h \in \mathcal{A}$. For an MDP $\hat{M}$ and a policy π, we define the h-stage action-value function $Q_{\pi,h}^{\hat{M}}(s,a)$ and the h-stage value function $V_{\pi,h}^M$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$Q_{\pi,h}^{\hat{M}}(s,a) := \mathbb{E}\left[\sum_{j=h}^{H-1} r_{j+1} \,\Big|\, M = \hat{M}, s_h = s, a_h = a, \pi\right],$$
$$V_{\pi,h}^{\hat{M}}(s) := Q_{\pi,h}^{\hat{M}}(s, \pi_h(s)), \quad \forall h = 0, 1, \cdots H-1.$$

The conditional expectation specifies that actions from stage $h+1$ to $H-1$ are determined according to the policy π when interacting with the MDP $\hat{M}$. We define the terminal value $V_{\pi,H}^{\hat{M}}(s) = 0$ for all $s \in \mathcal{S}$. Define the expected value of a policy π under an MDP $\hat{M}$ as $\overline{V}_\pi^{\hat{M}} = \mathbb{E}\left[V_{\pi,0}^{\hat{M}}(s_0) \mid \hat{M}, \pi\right]$. We say a policy $\pi^{\hat{M}}$ is optimal for the MDP $\hat{M}$ if $\pi^{\hat{M}} \in \arg\max_\pi V_{\pi,h}^{\hat{M}}(s)$ for all $(s,h) \in \mathcal{S} \times [H]$ where $[H] = \{0, 1, \ldots, H-1\}$. Let $\overline{V}_\pi := \overline{V}_\pi^M$ be expected value which measures the performance of a policy π under true MDP M.

---

[2]We focus on the time-inhomogeneous MDPs in this work. The minimax lower bound for time-inhomogeneous linear mixture MDPs (Zhou et al., 2021) is $\Omega(Hd\sqrt{T})$. Here d is the dimension of features for basis transition kernels.

[3]By the property of eluder dimension and Kolmogorov dimension, $\mathbb{E}\left[K^*\right] = \mathcal{O}(H), d_K = \mathcal{O}(d), d_E = \dim_E(\mathcal{F}, T^{-1}) = \mathcal{O}(d\log T)$ in the class of linear mixture MDPs where d is the dimension of feature in basis kernels and T is the total interaction rounds.

[4]We do not assume finite state-action spaces. The state space $\mathcal{S}$ and action space $\mathcal{A}$ can be unbounded.

Observations and environmental randomness. The environmental randomness is carried out when nature samples the reward and next state given the current state-action pair. During an episode ℓ, the agent applies a policy π_ℓ and obtains the *observations* up to stage h, which is $O_{\ell,h} := (s_{\ell,0}, a_{\ell,0}, s_{\ell,1}, \ldots, s_{\ell,h-1}, a_{\ell,h-1}, s_{\ell,h})$. The *observational history* before episode ℓ is denoted as $\mathcal{H}_\ell^o = (O_{1,H}, \ldots, O_{\ell-1,H})$, representing all realized environmental randomness that the agent could extract information from.

Algorithm and algorithmic randomness. The agent's behavior is governed by a reinforcement learning algorithm alg. The algorithmic randomness may be introduced when the agent applies a randomized algorithm to select actions in a way that depends on internally generated random numbers. Denote the random numbers used by alg in episode ℓ as U_ℓ. The history of algorithmic randomness before episode ℓ is denoted as $\mathcal{H}_\ell^u = (U_1, \ldots, U_{\ell-1})$. At the beginning of episode ℓ, the algorithm produces a policy $\pi_\ell = \text{alg}(\mathcal{S}, \mathcal{A}, \mathcal{H}_\ell^o, \mathcal{H}_\ell^u, U_\ell)$ based on the state space $\mathcal{S}$ and action space $\mathcal{A}$, the observational history $\mathcal{H}_\ell^o$ before episode ℓ and possibly the history of algorithmic randomness $\mathcal{H}_\ell^u$ before episode ℓ as well as the algorithmic randomness U_ℓ used in episode ℓ.

Notice that the algorithmic randomness is a different source of randomness; for each $k \in \mathbb{N}$, U_k is jointly independent of $\{U_\ell\}_{\ell \neq k}$, the environmental randomness and true MDP M. The algorithm alg usually computes some intermediate quantities in episode ℓ such as the virtual value functions $\hat{V}_{\ell,h}$. Conditioned on $\mathcal{H}_\ell^o$ and $\mathcal{H}_\ell^u$, the policy π_ℓ and the value functions $\hat{V}_{\ell,h}$ at each stage $h \in [H]$ are random only through their dependence on U_ℓ.

Bayesian RL and regret. We work with Bayesian reinforcement learning (RL) framework (Strens, 2000). To deliver our conceptual idea in a clean way, w.l.o.g., we assume[5] the agent understands everything about the MDP but is uncertain about the underlying transition P. The agent's initial knowledge and uncertainty about P is encoded in a prior (representative) distribution $\mathbb{P}(P \in \cdot)$. Thus, P is treated as a random variable in the agent's mind. Therefore, the underlying MDP M is also a random variable with prior distribution $\mathbb{P}(M \in \cdot)$. Let $\pi^* = \pi^M$ denote an optimal policy, where $\pi^M \in \arg\max_\pi \overline{V}_\pi$ is a function of the MDP M and is thus also a random variable.

The *regret* incurred in episode ℓ is the gap between the optimal value $\overline{V}_{\pi^*}$ and the state

value of alg under true MDP M, i.e. $\overline{V}_{\pi^*} - \overline{V}_{\pi_\ell}$. The *expected cumulative regret* over L episodes in the environment with underlying MDP M is $\Re(M, \text{alg}, L) = \sum_{\ell=1}^L \mathbb{E}_{M,\text{alg}} [\overline{V}_{\pi^*} - \overline{V}_{\pi_\ell} \mid M]$, where the expectation integrates over actions, state transitions, and any the algorithmic randomness used by alg, while the MDP M is fixed. It is often useful to analyze the performance of an agent in terms of *Bayesian regret*:

$$\mathfrak{BR}(\text{prior}, \text{alg}, L) = \mathbb{E}_{M \sim \text{prior}} [\Re(M, \text{alg}, L)],$$

where the expectation is taken over the prior distribution prior $:= \mathbb{P}(M \in \cdot)$ over the true MDP.

Posterior sampling algorithm for reinforcement learning (PSRL)(Strens, 2000) serves as a popular huristics to minimize Bayesian regret.

Algorithm 1 PSRL (episode ℓ)

Require: Prior distribution $\mathbb{P}(M \in \cdot)$ for underlying model M.
1: Sample $\hat{M}_\ell \sim \mathbb{P}(M \in \cdot \mid \mathcal{H}_\ell^o)$.
2: Solve optimal policy $\pi_\ell = \pi^{\hat{M}_\ell}$ under $\hat{M}_\ell$.
3: **return** π_ℓ

Linear mixture MDPs. We consider a class of MDPs called *linear mixture MDPs*, where the transition probability kernel P in the class $\mathcal{P}$ satisfying that for any $x = (h, s, a) \in \mathcal{X}$, $P(x) = \langle \theta_h^*, \phi(\cdot \mid x) \rangle$, i.e., the linear mixture of a number of basis kernels (Ayoub et al., 2020; Modi et al., 2020; Zhou et al., 2021). Here $\phi(s' \mid x) : \mathcal{S} \times \mathcal{X} \to \mathbb{R}^d$ is a feature mapping, a.k.a, basis transition kernels. The set $\Theta^* = (\theta_0^*, \ldots, \theta_{H-1}^*)$ includes the underlying model parameters which are random variables that should induce *proper transition probability kernel* $\langle \theta_h^*, \phi(\cdot \mid x) \rangle$ when combined with features.

We sometimes write $\hat{M} := M(\hat{\Theta})$ as a mapping from a parameter set $\hat{\Theta}$ to a linear mixture MDP $\hat{M}$. In the class of linear mixture MDPs, the unknown true MDP $M = M(\Theta^*)$ is random only through its dependence on Θ^*. Therefore, the prior we consider here is over model parameters $\mathbb{P}(\Theta^* \in \cdot)$. The posterior sampling (line 2 in Algorithm 1) is implemented as first sampling the set $\hat{\Theta}_\ell \sim \mathbb{P}(\Theta^* \in \cdot \mid \mathcal{H}_\ell)$ that contains $\hat{\Theta}_\ell = (\hat{\theta}_{\ell,0}, \ldots, \hat{\theta}_{\ell,H-1})$ and then constructing transition kernel $\hat{P}_\ell(x) = \langle \hat{\theta}_{\ell,h}, \phi(\cdot \mid x) \rangle$ for all $h \in [H]$ as well as the corresponding MDP $\hat{M}_\ell = (\mathcal{S}, \mathcal{A}, \hat{P}_\ell, R, H, \rho)$. To summarize, the posterior sampling algorithm for linear mixture MDPs is described in Algorithm 2.

Definition 1 (Value-correlated feature)**.** *For any $x = (h, s, a) \in \mathcal{X}$, we define the* value-correlated feature *induced by any bounded function V and any*

---
[5]Extending the algorithm and analysis in this paper to unknown stochastic rewards poses no real difficulty.

Algorithm 2 PSRL in linear mixture MDPs (episode ℓ)

Require: Prior distribution $\mathbb{P}(\Theta^* \in \cdot)$
 1: Sample $\hat{\Theta}_\ell = (\hat{\theta}_{\ell,0}, \ldots, \hat{\theta}_{\ell,H-1}) \sim \mathbb{P}(\Theta^* \in \cdot \mid \mathcal{H}_\ell^o)$.
 2: Construct $\hat{P}_\ell(h, \cdot, \cdot) = \left\langle \hat{\theta}_{\ell,h}, \phi(\cdot \mid \cdot, \cdot) \right\rangle, \forall h \in [H]$.
 3: Solve optimal policy π_ℓ under $\hat{M}_\ell = (\mathcal{S}, \mathcal{A}, \hat{P}_\ell, R, H, \rho)$.
 4: **return** π_ℓ

---

fixed feature mapping $\phi(s' \mid x) : \mathcal{S} \times \mathcal{X} \to \mathbb{R}^d$, $\phi_V(x) = \sum_{s' \in \mathcal{S}} \phi(s' \mid x) V(s')$. *Note that we do not assume finite state space and thus the summation can be replaced by integration if the state space is unbounded. The integration can be attained by an integration oracle.*

Definition 2 (Covariance matrix of unknown model parameters under posterior distribution). *We use the quantity posterior variance of the unknown model parameters* $\boldsymbol{\Gamma}_{\ell,h} := \mathbb{V}(\theta_h^* \mid \mathcal{H}_\ell^o)$ *as a uncertainty measurement.*

The covariance matrix serves as a uncertainty measurement that how much knowledge of the environment θ^* the agent is still not captured given the history of observations $\mathcal{H}_\ell^o$.

3   Bayesian regret bound for PSRL

Assumption 1. *For any* $x = (h, s, a) \in \mathcal{X}$, *let* $\phi(s' \mid x) : \mathcal{S} \times \mathcal{X} \to \mathbb{R}^d$ *be a feature mapping satisfying that for any bounded function* $V : \mathcal{S} \to [0, 1]$, $\|\phi_V(x)\|_2 \leqslant 1$, *where* ϕ_V *is the value-correlated feature associated with* V *(see Definition 1).*

Assumption 2. *Unknown model parameters* $\theta_0^*, \ldots, \theta_{H-1}^* \in \mathbb{R}^d$ *are mutually independent.*

Assumption 3. *Unknown model parameters have bounded norm* $\|\theta_h^*\|_2 \leqslant B$ *a.s. for all* $h \in [H]$.

Fact 1. *If the support of prior distribution is over norm-bounded model parameters satisfying Assumption 3, i.e.,* $\|\theta_h\|_2 \leqslant B$ *a.s. for all* $h \in [H]$, *we have*

$$\boldsymbol{\Gamma}_{1,h} \preceq \mathbb{E}\left[\|\theta - \mathbb{E}[\theta]\|_2^2\right] \boldsymbol{I}$$
$$= \mathbb{E}\left[\|\theta\|_2^2 - \|\mathbb{E}[\theta]\|_2^2\right] \boldsymbol{I} \preceq B^2 \boldsymbol{I}$$

Note that $T = HL$ is the total interaction steps, we have the two parts of the results separately stated in the following. Theorem 1 is the first prior-dependent regret bound of PSRL with function approximation while Remark 1 is a improved prior-free bound.

Theorem 1 (Prior-dependent analysis). *For any prior over models* $\Theta^* = (\theta_0^*, \ldots, \theta_{H-1}^*)$ *satisfying Assumption 2, PSRL have the Bayesian regret bound* $\mathfrak{BR}(\text{prior}, \text{PSRL}, L)$ *over* L *episodes interaction with the time-inhomogeneous linear mixture MDP satisfying Assumption 1, bounded by*

$$\sqrt{dH^3L \sum_{h=0}^{H-1} \log \det\left(\boldsymbol{I} + L\boldsymbol{\Gamma}_{1,h}\right)}$$

where $\boldsymbol{\Gamma}_{1,h}$ *is the covariance of* θ_h^* *under prior distribution.*

We characterize the dependence of the regret on prior distribution through the variance of the transition kernel. That is to say, if the agent knowledgeable about the environment, represented as small prior variance, then it will soon learn the optimal policy, incurring minimal regret. The proof can be found in Appendix C.

Remark 1 (Prior-free bound). *If Assumption 3 further holds, by the Fact 1, we have the trivial bound* $\log \det(\boldsymbol{I} + dL\boldsymbol{\Gamma}_{1,h}) \leqslant d \log(1 + dLB^2)$ *which leads to* $\mathfrak{BR}(\text{prior}, \text{PSRL}, L)$ *bounded by*

$$\sqrt{3}d\sqrt{H^4L \log(1 + LB^2)} = \mathcal{O}(d\sqrt{H^3T \log L})$$

Comparison with the regret bound of (Osband and Van Roy, 2014) Osband and Van Roy (2014) utilize the technique of eluder dimension and frequentist confidence set (Russo and Van Roy, 2014) to analyze PSRL with general function approximation. Their bound, when specified to linear mixture MDPs, is $\mathcal{O}(H^{3/2}d\sqrt{T} \log T)$. Because of not relying on the union bound over confidence-sets to upper bound regret as they did, our final regret bound will be naturally improved by a $\sqrt{\log T}$ factor. See more discussion of this $\log T$ term in Appendix G.

4   A value-targeted perspective of model learning

Dynamic programming. Given the model $\hat{M}$, value iteration is a classical dynamic programming solver for optimal policy. For any $x = (h, s, a) \in \mathcal{X}$, given the transition kernel $\hat{P}$ and next-state value function $\hat{V}_{h+1}$, we define a short operator form for the expected next-state value,

$$\hat{P}\hat{V}_{h+1}(x) = \mathbb{E}_{s'_x \sim \hat{P}(x)}\left[\hat{V}_{h+1}(s'_x)\right].$$

We use this short notation in the following Algorithm 3.

Algorithm 3 Value iteration

Require: $\hat{M} = (\mathcal{S}, \mathcal{A}, H, \hat{P}, \hat{R}, \rho)$ and init $\hat{V}_H(\cdot) = 0$.
1: **for** Stage $h = H - 1, \ldots, 0$ **do**
2:   $\hat{Q}_h(\cdot, \cdot) = \hat{R}(h, \cdot, \cdot) + \hat{P}\hat{V}_{h+1}(h, \cdot, \cdot)$
3:   $\hat{V}_h(\cdot) = \max_a \hat{Q}_h(\cdot, a)$
4:   $\pi_h(\cdot) = \arg\max_a \hat{Q}_h(\cdot, a)$
5: **end for**
6: **return** Policy $\pi = (\pi_h)_{h \in [H]}$ and value $(\hat{V}_h)_{h \in [H]}$

Model-based regret decomposition. In episode ℓ, a model-based algorithm typically construct a *virtual MDP* $\hat{M}_\ell$ with *virtual transition model* $\hat{P}_\ell$ based on the observational history $\mathcal{H}_\ell^o$ and solve a (near)-optimal policy π_ℓ under $\hat{M}_\ell$. Then, the RL agent acts according to the policy π_ℓ in episode ℓ. The regret in episode ℓ can be decomposed to two terms,

$$\overline{V}_{\pi^*} - \overline{V}_{\pi_\ell} = \underbrace{\overline{V}_{\pi^*}^M - \overline{V}_{\pi_\ell}^{\hat{M}_\ell}}_{\text{Pessimism}} + \underbrace{\overline{V}_{\pi_\ell}^{\hat{M}_\ell} - \overline{V}_{\pi_\ell}^M}_{\text{Estimation error}} \qquad (1)$$

The *pessimism* term is often made non-positive in expectation or with high probability via exploration mechanism induced by the algorithm. Then we suffice to bound the Bayesian regret by the expected summation of estimation error $\mathbb{E}\left[\sum_{\ell=1}^L \overline{V}_{\pi_\ell}^{\hat{M}_\ell} - \overline{V}_{\pi_\ell}^M\right]$. The *estimation error* in episode ℓ is the error of estimating value $\overline{V}_{\pi_\ell}^M$ of a particular policy π_ℓ via the virtual model $\hat{M}_\ell$. Let $\mathcal{X} = [H] \times \mathcal{S} \times \mathcal{A}$ and let $T = HL$ denote the total number of interactions in L episodes. Generally, we can decompose the estimation error into the *value-targeted model error* along the on-policy trajectory as shown in the following lemma (Kearns and Singh, 2002; Osband et al., 2013). One can find the proof in Appendix D.2.

Lemma 1. *Define the short notation* $x_h := (h, s_h, a_h)$. *For any MDP* $\hat{M}$ *and policy* π,

$$\overline{V}_\pi^{\hat{M}} - \overline{V}_\pi^M = \sum_{h=0}^{H-1} \mathbb{E}\Big[\underbrace{\left(\hat{P} - P\right) V_{\pi, h+1}^{\hat{M}}(x_h)}_{\textit{value-targeted model error}} \mid M, \hat{M}, \pi \Big],$$

where $\hat{P}$ *is the transition model under* $\hat{M}$ *and* x_h *is rolled out under the policy* π *over the MDP* M.

4.1 Value-targeted model learning

As shown in Lemma 1 and (1), the value-targeted model error is a sufficient condition for bounding estimation error and thus bounding the regret. Then, we conclude the following rule:

> *A virtual model is a desideratum for a task if it is accurate enough to help predict future high values.*

We can also treat the value-targeted model error as a tracking signal for the learning progress of a model-based RL algorithm. Motivated by this observation, through this work, we take the *value-target perspective* for model-based learning. This perspective is related to the value-targeted regression (Ayoub et al., 2020).

Let $\mathcal{P}$ be a general transition function class. Let $\mathcal{B}(\mathcal{S}, H)$ denotes the set of real-valued measurable functions with domain $\mathcal{S}$ that are bounded by H. Let $\mathcal{V} \subseteq \mathcal{B}$ be the set of value functions. For any tuple $x = (h, s, a) \in \mathcal{X}$, any value function $V \in \mathcal{V}$, and any underlying transition model $P \in \mathcal{P}$, let the next-state be $s'_x \sim P(x)$ and let the value at sampled next-state be $Y = V(s'_x)$. Let the expected value at next-state given the x and V be $f_P(x, V) = \mathbb{E}_{s'_x \sim P(x)}[V(s'_x)]$. Note that the input (x, V) and transition model P can be random variables. We have the conditional expectation and the conditional variance of the outcome Y,

$$\mathbb{E}\left[Y \mid P, x, V\right] = f_P(x, V)$$
$$\mathbb{V}\left(Y \mid P, x, V\right) = \mathbb{V}_{s'_x \sim P(x)}\left(V(s'_x)\right)$$

From the *value-targeted perspective*, the *information acquisition* of transition model can be made possible through the sequence of input (x, V) and observed outcome $Y = V(s'_x)$:

$$Y = f_P(x, V) + Z, \qquad (2)$$

where Z is the noise term with conditional variance $\mathbb{V}\left(Z \mid P, x, V\right) = \mathbb{V}\left(Y \mid P, x, V\right)$. The noise Z involves the transitional noise when sampling next state and the value at the sampled state.

To explicitly represent the input and outcome sequentially, we define a new notion of history. Denote $\hat{V}_{\ell,h} := V_{\pi_\ell,h}^{\hat{M}_\ell}$ for $h \in [H]$ and let $\hat{V}_{\ell,H+1} = 0$ and $a_{\ell,H} = a \in \mathcal{A}$ be constant dummy variables. We are ready to define the *value-augmented observations* up to stage h in episode ℓ,

$$\hat{O}_{\ell,h} := (s_{\ell,0}, a_{\ell,0}, \hat{V}_{\ell,1}, \ldots, s_{\ell,h}, a_{\ell,h}, \hat{V}_{\ell,h+1})$$
$$= (x_{\ell,0}, \hat{V}_{\ell,1}, \ldots, x_{\ell,h}, \hat{V}_{\ell,h+1})$$

where the short notation is $x_{\ell,h} := (h, s_{\ell,h}, a_{\ell,h})$. We define the *history* up to the beginning of episode ℓ as

$$\mathcal{H}_\ell := (\pi_1, \hat{O}_{1,H}, \ldots, \pi_{\ell-1}, \hat{O}_{\ell-1,H})$$

and the *history* up to the stage h of episode ℓ as

$$\mathcal{H}_{\ell,h} := (\mathcal{H}_\ell, \pi_\ell, \hat{O}_{\ell,h}).$$

By definition, the input $(x_{\ell,h}, \hat{V}_{\ell,h+1})$ at stage h of episode ℓ is $\sigma(\mathcal{H}_{\ell,h})$-measurable while the outcome $Y_{\ell,h+1} := \hat{V}_{\ell,h+1}(s_{\ell,h+1})$ is $\sigma(\mathcal{H}_{\ell,h+1})$-measurable. This

is a sequential representation of the information process and will be useful throughout the presentation.

Similarly to Lemma 1, we have a generic regret decomposition from value-targeted perspective.

Lemma 2 (Estimation decomposition conditioned on history). *Define the estimation error in stage h of episode ℓ by $\Delta_{\ell,h}(s) = V_{\pi_\ell,h}^{\hat{M}_\ell}(s) - V_{\pi_\ell,h}^M(s)$ for any fixed state $s \in \mathcal{S}$, we have*

$$\mathbb{E}_{\ell,h}\left[\Delta_{\ell,h}(s_{\ell,h})\right] = \mathbb{E}_{\ell,h}\left[\sum_{j=h}^{H-1}\mathbb{E}_{\ell,j}\left[\left(\hat{P}_\ell - P\right)\hat{V}_{\ell,j+1}(x_{\ell,j})\right]\right]$$

4.2 PSRL from value-targeted perspective

In this section, we state a key observation for analysis and show the posterior distribution for the unknown true MDP M given the *observational history* is the same as the posterior given the *history*.

Lemma 3. *For any $\sigma(\mathcal{H}_\ell^o, \mathcal{H}_\ell^u, U_\ell)$-measurable random variable ξ, we have $\mathbb{P}(M \in \cdot \mid \mathcal{H}_\ell^o) = \mathbb{P}(M \in \cdot \mid \xi, \mathcal{H}_\ell^o)$.*

Lemma 3 is due to the following key observations.

$$\begin{aligned}
\mathbb{P}(M \in \cdot \mid \mathcal{H}_\ell^o) &= \mathbb{P}(M \in \cdot \mid \mathcal{H}_\ell^o, \mathcal{H}_\ell^u, U_\ell) \\
&= \mathbb{P}(M \in \cdot \mid \xi, \mathcal{H}_\ell^o, \mathcal{H}_\ell^u, U_\ell) \\
&= \mathbb{P}(M \in \cdot \mid \xi, \mathcal{H}_\ell^o).
\end{aligned}$$

The first equality is due to the independence between $M \perp\!\!\!\perp \mathcal{H}_\ell^u$, $M \perp\!\!\!\perp U_\ell$ and $U_\ell \perp\!\!\!\perp \mathcal{H}_\ell^u$. The second equality is by definition ξ is $\sigma(\mathcal{H}_\ell^o, \mathcal{H}_\ell^u, U_\ell)$-measurable. The third equality is also due to the independent algorithmic randomness. For any historical episode $\ell' < \ell$, the historical policy $\pi_{\ell'}$ and the historical imagined value functions $\{\hat{V}_{\ell',h}\}_{h=0}^{H+1}$ are $\sigma(\mathcal{H}_{\ell'}^o, \mathcal{H}_{\ell'}^u, U_{\ell'})$-measurable by definition and thus $\sigma(\mathcal{H}_\ell^o, \mathcal{H}_\ell^u, U_\ell)$-measurable. Recall the difference between of $\mathcal{H}_\ell^o$ and $\mathcal{H}_\ell$ is exactly the historical policies and historical imagined value functions. Therefore, by Lemma 3, the posterior distribution of true MDP M conditioned on the observational history is equivalent to the one conditioned on history: $\mathbb{P}(M \in \cdot \mid \mathcal{H}_\ell^o) = \mathbb{P}(M \in \cdot \mid \mathcal{H}_\ell)$. By posterior sampling, we have $\mathbb{P}(\hat{M}_\ell \in \cdot \mid \mathcal{H}_\ell^o) = \mathbb{P}(M \in \cdot \mid \mathcal{H}_\ell^o)$. Therefore, for any $\mathcal{H}_\ell^o$-measurable function g, we conclude $\mathbb{P}(g(M) \in \cdot \mid \mathcal{H}_\ell^o) = \mathbb{P}(g(M^\ell) \in \cdot \mid \mathcal{H}_\ell^o)$. This is noted by (Osband et al., 2013; Russo and Van Roy, 2014) and has an immediate consequence on the pessimism term:

$$\mathbb{E}\left[\overline{V}_{\pi^*}^M - \overline{V}_{\pi_\ell}^{\hat{M}_\ell} \,\middle|\, \mathcal{H}_\ell\right] = \mathbb{E}\left[\overline{V}_{\pi^*}^M - \overline{V}_{\pi_\ell}^{\hat{M}_\ell} \,\middle|\, \mathcal{H}_\ell^o\right] = 0$$

Due to the posterior equivalence on MDP M between $\mathcal{H}_\ell^o$-conditioning and $\mathcal{H}_\ell$-conditioning and to promote the value-targeted perspective of model learning,

we use $\mathcal{H}_\ell$-conditioning later in the presentation. Also, for the purpose of simplicity, we sometimes use the following short notations $\mathbb{P}_{\ell,h}(\cdot), \mathbb{E}_{\ell,h}[\cdot]$ and $\mathbb{V}_{\ell,h}(\cdot)$ to donote conditional probability $\mathbb{P}(\cdot \mid \mathcal{H}_{\ell,h})$, conditional expectation $\mathbb{E}[\cdot \mid \mathcal{H}_{\ell,h}]$ and conditional variance $\mathbb{V}(\cdot \mid \mathcal{H}_{\ell,h})$ respectively.

4.3 Posterior variance reduction

In linear mixture MDPs, to mathematically quantify the reduction of uncertainty of the unknown environment when new information gathered in, we also relate the posterior variance of $\mathbf{\Gamma}_{\ell,h}$ in Definition 2 to value targeted perspective. First, we notice by Lemma 3

$$\mathbf{\Gamma}_{\ell,h} = \mathbb{V}(\theta_h^* \mid \mathcal{H}_\ell^o) = \mathbb{V}(\theta_h^* \mid \mathcal{H}_\ell).$$

Recall the definition of function f_P in Section 4.1. Specifically, for the class of linear mixture MDPs, i.e. $P(x) = \langle\theta_h^*, \phi(\cdot \mid x)\rangle$ for all $x = (h, s, a) \in \mathcal{X}$,

$$\begin{aligned}
f_P(x, V) = PV(x) &= \sum_{s_x' \in \mathcal{S}} \langle\theta_h^*, \phi(s_x' \mid x)\rangle V(s_x') \\
&= \left\langle\theta_h^*, \sum_{s_x' \in \mathcal{S}} \phi(s_x' \mid x)V(s_x')\right\rangle = \langle\theta_h^*, \phi_V(x)\rangle
\end{aligned} \tag{3}$$

Recall the relationship between the input (x, V) and outcome Y in (2), combined with (3), we have the following important observations,

$$Y = \langle\theta_h^*, \phi_V(x)\rangle + Z$$

is a noisy linear model with non-Gaussian prior over θ_h^* and non-Gaussian noise Z.

Definition 3. *Define the short notation $X_{\ell,h} := \phi_{\hat{V}_{\ell,h+1}}(x_{\ell,h}) = \phi_{\hat{V}_{\ell,h+1}}(h, s_{\ell,h}, a_{\ell,h})$.*

Definition 4. *Define $Y_{\ell,h+1} := \hat{V}_{\ell,h+1}(s_{\ell,h+1})$.*

Recall the definition of value-augmented history $\mathcal{H}_\ell$ and $\mathcal{H}_{\ell,h}$ in Section 4, we can see $X_{\ell,h}$ is $\sigma(\mathcal{H}_{\ell,h})$-measurable since $X_{\ell,h}$ is a deterministic function of $(s_{\ell,h}, a_{\ell,h}, \hat{V}_{\ell,h+1})$ which is included in $\mathcal{H}_{\ell,h}$. Similarly, since $Y_{\ell,h+1}$ is a deterministic function of $(\hat{V}_{\ell,h+1}, s_{\ell,h+1})$ which is included in $\mathcal{H}_{\ell,h+1}$, we conclude $Y_{\ell,h+1}$ is $\sigma(\mathcal{H}_{\ell,h+1})$-measurable. Then with the fact that $s_{\ell,h+1} \sim P(x_{\ell,h})$, by (3), we can verify that,

$$\mathbb{E}[Y_{\ell,h+1} \mid \Theta^*, \mathcal{H}_{\ell,h}] = \langle\theta_h^*, X_{\ell,h}\rangle$$

and the variance of $Y_{\ell,h+1}$ conditioned on history $\mathcal{H}_{\ell,h}$ and true model Θ^* is defined as $\sigma_{\ell,h}^2$ which is

$$\mathbb{V}(Y_{\ell,h+1} \mid \Theta^*, \mathcal{H}_{\ell,h}) = \mathbb{V}\left(\hat{V}_{\ell,h+1}(s_{\ell,h+1}) \mid M, \mathcal{H}_{\ell,h}\right)$$

By Assumption 2, conditioned on history $\mathcal{H}_\ell$, the trajectory $(s_{\ell,0}, a_{\ell,0}, \ldots, s_{\ell,h}, a_{\ell,h})$ up to stage h of episode ℓ is independent of true transition kernel $P(h, \cdot, \cdot)$ and thus independent of true model parameter θ_h^*. By definition, conditioned on history $\mathcal{H}_\ell$, the policy π_ℓ and imagined value functions $\{\hat{V}_{\ell,h}\}_{h=1}^H$ are random only through its dependence on algorithmic randomness U_ℓ, which is also independent of θ_h^*. Therefore, we have the following important relationship on the posterior variance,

$$
\begin{aligned}
\mathbb{V}\left(\theta_h^* \mid \mathcal{H}_{\ell,h}\right) &= \mathbb{V}\left(\theta_h^* \mid \mathcal{H}_\ell, \pi_\ell, \hat{O}_{\ell,h}\right) \\
&= \mathbb{V}\left(\theta_h^* \mid \mathcal{H}_\ell, \pi_\ell, s_{\ell,0}, a_{\ell,0}, \hat{V}_{\ell,1}, \ldots, s_{\ell,h}, a_{\ell,h}, \hat{V}_{\ell,h+1}\right) \\
&= \mathbb{V}\left(\theta_h^* \mid \mathcal{H}_\ell\right) =: \mathbf{\Gamma}_{\ell,h}
\end{aligned}
$$

Similarly, conditioned on history $\mathcal{H}_{\ell,h+1}$, the trajectory $(s_{\ell,h+2}, a_{\ell,h+2}, \ldots, s_{\ell,H-1}, a_{\ell,H-1}, s_{\ell,H})$ after step $h+1$ in episode ℓ is independent of θ_h^*. Then, we have,

$$
\begin{aligned}
\mathbb{V}\left(\theta_h^* \mid \mathcal{H}_{\ell,h+1}\right) &= \mathbb{V}\left(\theta_h^* \mid \mathcal{H}_\ell, \pi_\ell, \hat{O}_{\ell,H}\right) \\
&= \mathbb{V}\left(\theta_h^* \mid \mathcal{H}_{\ell+1}\right) =: \mathbf{\Gamma}_{\ell+1,h}
\end{aligned}
$$

Now we are ready to apply Lemma 11 (posterior variance reduction lemma for general prior and general likelihood in noisy linear model) to the above noise linear system and obtain the Theorem 2. It shows that, in expectation, the posterior covariance of the true model parameters will be reduced. The reduction depends on the variance in the direction of the value-correlated feature vector quantified with the prior distribution of true model parameters. The reduction also depends on the variance of value due to the transitional noise. Denote short notations $\sigma_{\ell,h}^2 := \mathbb{V}_{\ell,h}\left(\hat{V}_{\ell,h+1}(s_{\ell,h+1}) \mid M\right)$ and $X_{\ell,h} := \phi_{\hat{V}_{\ell,h+1}}(x_{\ell,h})$.

Theorem 2 (Posterior variance reduction).

$$
\mathbb{E}_{\ell,h}\left[\mathbf{\Gamma}_{\ell+1,h}\right] \preceq \mathbf{\Gamma}_{\ell,h} - \frac{\mathbf{\Gamma}_{\ell,h} X_{\ell,h} X_{\ell,h}^\top \mathbf{\Gamma}_{\ell,h}}{\mathbb{E}_{\ell,h}\left[\sigma_{\ell,h}^2\right] + X_{\ell,h}^\top \mathbf{\Gamma}_{\ell,h} X_{\ell,h}},
$$

where $A \preceq B$ for any two matrices A and B iff $B - A$ is a positive semi-definite matrix.

Remark 2. For any $\bar{\sigma}_{\ell,h}$ such that $\bar{\sigma}_{\ell,h}^2 \geqslant \mathbb{E}_{\ell,h}\left[\sigma_{\ell,h}^2\right]$ almost surely, Theorem 2 implies

$$
\mathbb{E}_{\ell,h}\left[\mathbf{\Gamma}_{\ell+1,h}\right] \preceq \mathbf{\Gamma}_{\ell,h} - \frac{\mathbf{\Gamma}_{\ell,h} X_{\ell,h} X_{\ell,h}^\top \mathbf{\Gamma}_{\ell,h}}{\bar{\sigma}_{\ell,h}^2 + X_{\ell,h}^\top \mathbf{\Gamma}_{\ell,h} X_{\ell,h}}.
$$

Further, by Sherman-Morrison formula, Theorem 2 implies (assume invertible)

$$
\mathbb{E}_{\ell,h}\left[\mathbf{\Gamma}_{\ell+1,h}\right]^{-1} \succeq \mathbf{\Gamma}_{\ell,h}^{-1} + \frac{1}{\bar{\sigma}_{\ell,h}^2} X_{\ell,h} X_{\ell,h}^\top.
$$

In the analysis, we choose $\bar{\sigma}_{\ell,h}$ s.t. $\bar{\sigma}_{\ell,h}^2 := \mathbb{E}_{\ell,h}\left[\sigma_{\ell,h}^2\right] \vee \sigma_{\min}^2$. We choose $\sigma_{\min} = H$. Note that the Theorem 2 still holds if the parameter $\hat{\theta}_{\ell,h}$ does not induce a proper transition kernel.

4.4 Decouple regret to posterior variance

We are now ready to see how the regret is being translated to the posterior variance over models from the value-targeted perspective. All details of analysis can be found in Appendix B. Denote the estimation error in stage h of episode ℓ shortly by $\Delta_{\ell,h}(s) = V_{\pi_\ell,h}^{\hat{M}_\ell}(s) - V_{\pi_\ell,h}^M(s)$ for any fixed state $s \in \mathcal{S}$. The first step is to translated regret into inner product between model parameters and value-correlated features under linear mixture model assumptions.

Lemma 4 (Regret to model posterior deviation over feature directions).

$$
\mathbb{E}_{\ell,h}\left[|\Delta_{\ell,h}(s_{\ell,h})|\right] \leqslant \mathbb{E}_{\ell,h}\left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j}\left[\left|\left\langle \hat{\theta}_{\ell,j} - \theta_j^*, X_{\ell,j}\right\rangle\right|\right]\right]
$$

The key to relate regret to posterior variance is the decoupling argument.

Lemma 5 (Decoupling lemma). *Let θ^*, $\hat{\theta}$ and ϕ be arbitrary random vectors in $\mathbb{R}^d$. If $\hat{\theta}$ is an i.i.d copy of θ^*, then*

$$
\mathbb{E}\left[\left|\left\langle \hat{\theta} - \theta^*, \phi\right\rangle\right|\right]^2 \leqslant 2d\mathbb{E}\left[\phi^\top \mathbb{V}\left(\theta^*\right)\phi\right].
$$

Note that the above inequality holds even when $\phi \in \mathbb{R}^d$ is dependent on θ^* or $\hat{\theta}$.

With Lemmas 4 and 5 in hand, we now give the general regret decomposition that is essential for prior-dependent Bayesian analysis of PSRL.

Proposition 1 (General regret bound (Informal)). *For any episode ℓ and any stage h, $\mathbb{E}\left[\sum_{\ell=1}^L |\Delta_{\ell,h}(s_{\ell,h})|\right]$ is bounded by a term including the sum of posterior variance in the on-policy trajectory,*

$$
\sqrt{d\mathbb{E}\left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} \|X_{\ell,h}\|_{\mathbf{\Gamma}_{\ell,h}}^2\right]}
$$

Notice that this relationship does not require any assumption additional to linear mixture MDPs.

With the general regret bound and the posterior variance reduction argument to track the learning process, we are ready to prove Theorem 1. The sketch and detail of the proof can be found in Appendix C.

5 Discussions

We now discuss the **generality of the techniques** we developed and their possible applications for analyzing other problems:

- The posterior variance reduction technique in Section 4.3 and Appendix F.1 is applicable to any prior-likelihood model as well as any learning algorithm. It measures the posterior variance reduction of model parameters only through features in the history (see the comparison between Remarks 4 and 5) which in some sense recover the property of Gaussian posterior. This key observation may also enable mis-specification analysis of Gaussian approximation of environment that is not Gaussian.

- The ability of posterior variance reduction lemma to characterize heterogeneous noise likelihood in a predictable way for the sequential decision problem may enable some rigorous understanding of the benefit of information directed sampling compared to thompson sampling.

- The decoupling technique Lemmas 5 and 6 is possibly applicable to other *randomized* algorithms when the imagined virtual model $\hat{\theta}$ is not exact identically distributed as the true model θ^*.

Acknowledgements

References

Agarwal, A., Jin, Y., and Zhang, T. (2023). Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063. PMLR.

Agarwal, R., Schuurmans, D., and Norouzi, M. (2020). An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1220–III–1228. JMLR.org.

Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3(null):397–422.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474.

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349.

Bertsekas, D. and Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, pages 355–366.

Hamidi, N. and Bayati, M. (2021). The randomized elliptical potential lemma with an application to linear thompson sampling. *arXiv*, 2102.07987.

Ishfaq, H., Cui, Q., Nguyen, V., Ayoub, A., Yang, Z., Wang, Z., Precup, D., and Yang, L. (2021). Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR.

Ishfaq, H., Lan, Q., Xu, P., Mahmood, A. R., Precup, D., Anandkumar, A., and Azizzadenesheli, K. (2024). Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. In *The Twelfth International Conference on Learning Representations*.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143.

Kalkanlı, C. and Özgür, A. (2020). An improved regret bound for thompson sampling in the gaussian linear bandit setting. In *IEEE International Symposium on Information Theory*, pages 2783–2788.

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232.

Lattimore, T. and Hutter, M. (2012). Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pages 320–334.

Li, Y., Wang, Y., and Zhou, Y. (2019). Nearly minimax-optimal regret for linearly parameterized bandits. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2173–2174. PMLR.

Li, Y., Xu, J., Han, L., and Luo, Z.-Q. (2024). Hyperagent: A simple, scalable, efficient and provable reinforcement learning framework for complex environments. *arXiv preprint arXiv:2402.10228*.

Li, Z., Li, Y., Zhang, Y., Zhang, T., and Luo, Z.-Q. (2022). Hyperdqn: A randomized exploration method for deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*.

Lu, X. and Van Roy, B. (2019). Information-theoretic confidence bounds for reinforcement learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lu, X., Van Roy, B., Dwaracherla, V., Ibrahimi, M., Osband, I., and Wen, Z. (2021). Reinforcement learning, bit by bit. *arXiv preprint arXiv:2103.04047*.

Modi, A., Jiang, N., Tewari, A., and Singh, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020.

Munos, R. and Moore, A. (1999). Influence and variance of a markov chain: Application to adaptive discretization in optimal control. In *IEEE Conference on Decision and Control*, pages 1464–1469.

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474.

Osband, I. and Van Roy, B. (2017). Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pages 2701–2710.

Osband, I., Van Roy, B., and Russo, D. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011.

Osband, I., Van Roy, B., Russo, D. J., and Wen, Z. (2019). Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

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

Strens, M. (2000). A bayesian framework for reinforcement learning. In *International Conference on Machine Learning*, pages 943–950.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Wang, Q., Li, Y., Xiong, J., and Zhang, T. (2019). Divergence-augmented policy optimization. *Advances in Neural Information Processing Systems*, 32.

Wang, Q., Xiong, J., Han, L., Liu, H., Zhang, T., et al. (2018). Exponentially weighted imitation learning for batched historical data. *Advances in Neural Information Processing Systems*, 31.

Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International conference on machine learning*, pages 6995–7004. PMLR.

Yang, S., Nachum, O., Du, Y., Wei, J., Abbeel, P., and Schuurmans, D. (2023). Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. (2020). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964.

Zhang, T. (2021). Feel-Good Thompson Sampling for Contextual Bandits and Reinforcement Learning. *arXiv e-prints*, page arXiv:2110.00871.

Zhou, D., Gu, Q., and Szepesvari, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4532–4576. PMLR.

A   Notations

In this section, we collect all appeared symbols in the Table 1.

Table 1: Symbols

Symbol	Meaning
M	Underlying true MDP for the environment.
Θ^*	The set of model parameters for underlying linear mixture MDP $\Theta^* = (\theta_0^*, \ldots, \theta_{H-1}^*)$.
$\hat{M}_\ell$	Virtual MDP constructed by the agent in episode ℓ.
$\hat{\Theta}_\ell$	The set of model parameters $\hat{\Theta}_\ell = (\hat{\theta}_{\ell,0}, \ldots, \hat{\theta}_{\ell,H-1})$ in episode ℓ sampled by PSRL.
$s_{\ell,h}$	state encountered in stage h of episode ℓ by the agent.
$a_{\ell,h}$	action encountered in stage h of episode ℓ by the agent.
$x_{\ell,h}$	short notation for $(h, s_{\ell,h}, a_{\ell,h})$.
$\hat{V}_{\ell,h}$	the virtual value function $V_{\pi_\ell,h}^{\hat{M}_\ell}$ at stage h computed in episode ℓ.
$X_{\ell,h}$	future-virtual-value-correlated feature $\phi_{\hat{V}_{\ell,h+1}}(x_{\ell,h}) = \sum_{s'} \phi(s' \mid x_{\ell,h}) \hat{V}_{\ell,h+1}(s')$.
$Y_{\ell,h}$	virtual value $\hat{V}_{\ell,h}(s_{\ell,h})$ at the state encountered in stage h and episode ℓ.
$\Delta_{\ell,h}(s)$	Estimation error of policy π_ℓ using virtual MDP $\hat{M}_\ell$, i.e. $V_{\pi_\ell,h}^{\hat{M}_\ell}(s) - V_{\pi_\ell,h}^M(s)$.
$O_{\ell,h}$	Observations $(s_{\ell,0}, a_{\ell,0}, \ldots, s_{\ell,h}, a_{\ell,h})$.
$\mathcal{H}_\ell^o$	Observational history $\mathcal{H}_\ell^o = (O_{1,H}, \ldots, O_{\ell,H})$ up to the beginning of episode ℓ.
$\mathcal{H}_{\ell,h}^o$	Observational history $\mathcal{H}_\ell^o = (\mathcal{H}_\ell^o, O_{\ell,h})$ up to the stage h of episode ℓ.
$\hat{O}_{\ell,h}$	Value-augmented observations $(s_{\ell,0}, a_{\ell,0}, \hat{V}_{\ell,1}, \ldots, s_{\ell,h}, a_{\ell,h}, \hat{V}_{\ell,h+1})$.
$\mathcal{H}_\ell$	History $(\pi_1, \hat{O}_{1,H}, \ldots, \pi_{\ell-1}, \hat{O}_{\ell-1,H})$ up to the beginning of episode ℓ.
$\mathcal{H}_{\ell,h}$	History $(\mathcal{H}_\ell, \pi_\ell, \hat{O}_{\ell,h})$ up to the stage h of episode ℓ. Note $\mathcal{H}_{\ell,H}$ is exactly $\mathcal{H}_{\ell+1}$
$\mathbb{P}_\ell(\cdot)$	short notation for conditional probability $\mathbb{P}(\cdot \mid \mathcal{H}_\ell)$.
$\mathbb{P}_{\ell,h}(\cdot)$	short notation for conditional probability $\mathbb{P}(\cdot \mid \mathcal{H}_{\ell,h})$.
$\mathbb{E}_\ell[\cdot]$	short notation for conditional expectation $\mathbb{E}[\cdot \mid \mathcal{H}_\ell]$.
$\mathbb{E}_{\ell,h}[\cdot]$	short notation for conditional expectation $\mathbb{E}[\cdot \mid \mathcal{H}_{\ell,h}]$.
$\mathbb{V}_\ell(\cdot)$	short notation for conditional variance (covariance) $\mathbb{V}(\cdot \mid \mathcal{H}_\ell)$.
$\mathbb{V}_{\ell,h}(\cdot)$	short notation for conditional variance (covariance) $\mathbb{V}(\cdot \mid \mathcal{H}_{\ell,h})$.
$\mathbb{E}[\cdot]^p$	The p-th power to the expectation $(\mathbb{E}[\cdot])^p$.
$\mathbf{\Gamma}_{\ell,h}$	the posterior covariance matrix $\mathbb{V}_\ell(\theta_h^*)$ of the true parameter θ_h^*.
$\sigma_{\ell,h}^2$	virtual-value variance only due to transitional noise $\mathbb{V}_{\ell,h}\left(\hat{V}_{\ell,h+1}(s_{\ell,h+1}) \mid M\right)$.
$\sigma_{\min}$	a deterministic constant for the convenient of analysis.
$\bar{\sigma}_{\ell,h}$	random variable $\sqrt{\mathbb{E}_{\ell,h}\left[\sigma_{\ell,h}^2\right]} \vee \sigma_{\min}$.

B   Regret decomposition

Define the estimation error in stage h of episode ℓ by $\Delta_{\ell,h}(s) = V_{\pi_\ell,h}^{\hat{M}_\ell}(s) - V_{\pi_\ell,h}^M(s)$ for any fixed state $s \in \mathcal{S}$. Recall the definition of $X_{\ell,h}$ in Definition 3 and $\bar{\sigma}_{\ell,h}$ in Remark 2, we have the regret decomposition

Proposition 2 (Regret decomposition). *For any episode ℓ and any stage h,*

$$\mathbb{E}\left[\sum_{\ell=1}^L |\Delta_{\ell,h}(s_{\ell,h})|\right] \leqslant \sqrt{2d\mathbb{E}\left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} \bar{\sigma}_{\ell,h}^2\right] \mathbb{E}\left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} 1 \wedge \left\|\frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}}\right\|_{\mathbf{\Gamma}_{\ell,h}}^2\right]}$$

Furthermore, if Assumption 2 holds,

$$\mathbb{E}\left[\sum_{\ell=1}^L \overline{V}_{\pi_\ell}^{\hat{M}_\ell} - \overline{V}_{\pi_\ell}^M\right] \leqslant \sqrt{d\mathbb{E}\left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} \bar{\sigma}_{\ell,h}^2\right] \mathbb{E}\left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} 1 \wedge \left\|\frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}}\right\|_{\mathbf{\Gamma}_{\ell,h}}^2\right]}$$

B.1   Regret decomposition

Recall the short notation of estimation error in stage h of episode ℓ, $\Delta_{\ell,h}(s) = V_{\pi_\ell,h}^{\hat{M}_\ell}(s) - V_{\pi_\ell,h}^M(s)$. We have the following lemmas prepared for Proposition 2.

Lemma 6 (Regret to model posterior deviation over feature directions).

$$\mathbb{E}_{\ell,h}\left[\Delta_{\ell,h}(s_{\ell,h})\right] = \mathbb{E}_{\ell,h}\left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j}\left[\left\langle \hat{\theta}_{\ell,j} - \theta_j^*, X_{\ell,j}\right\rangle\right]\right]$$

$$\mathbb{E}_{\ell,h}\left[|\Delta_{\ell,h}(s_{\ell,h})|\right] \leqslant \mathbb{E}_{\ell,h}\left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j}\left[\left|\left\langle \hat{\theta}_{\ell,j} - \theta_j^*, X_{\ell,j}\right\rangle\right|\right]\right]$$

Corollary 1. *If Assumption 2 (mutual-independence of unknown model parameters) holds, a corollary of Lemma 6 is*

$$\mathbb{E}_{\ell,h}\left[\Delta_{\ell,h}(s_{\ell,h})\right] = \mathbb{E}_{\ell,h}\left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j}\left[\left\langle \hat{\theta}_{\ell,j} - \mathbb{E}_\ell\left[\hat{\theta}_{\ell,j}\right], X_{\ell,j}\right\rangle\right]\right]$$

B.2   Decoupling

In order to bound the RHS of Lemma 6, we utilize the idea in (Russo and Van Roy, 2016; Kalkanlı and Özgür, 2020) for linear bandit. The Lemma 7 shown here is slightly stronger than previous literature. This stronger statement is essential to bound the absolute value-targeted error in Proposition 2 which is required for proving a better H dependence by bounding the absolute estimation error $|\Delta_{\ell,h}|$ in Proposition 3. Extending previous statement in bandit literature to this stronger statement is not difficult but new. The proof of Lemma 7 can be found in Appendix F.2.

Lemma 7. *Let X_1 and X_2 be arbitrary i.i.d., $\mathbb{R}^m$ valued random variables and f_1, f_2 measurable maps such that $f_1, f_2 : \mathbb{R}^m \to \mathbb{R}^d$ with $\mathbb{E}\left[\|f_1(X_1)\|_2^2\right], \mathbb{E}\left[\|f_2(X_1)\|_2^2\right] < \infty$, then*

$$\mathbb{E}\left[\left|f_1(X_1)^\top f_2(X_1)\right|\right]^2 \leqslant d\mathbb{E}\left[\left(f_1(X_1)^\top f_2(X_2)\right)^2\right].$$

Specifically, let $X_1 = (X, Z)$, $f_1(x,z) = x, f_2(x,z) = z$, and X_2 be a i.i.d copy of X_1,

$$\mathbb{E}\left[\left|X^\top Z\right|\right]^2 \leqslant d\operatorname{Tr}\left(\mathbb{E}\left[XX^\top\right]\mathbb{E}\left[ZZ^\top\right]\right)$$

Corollary 2. *Let θ^*, $\hat\theta$ and ϕ be arbitrary random vectors in $\mathbb{R}^d$. If $\hat\theta$ is an i.i.d copy of θ^*, then*

$$\mathbb{E}\left[\left|\left\langle\hat\theta-\mathbb{E}\left[\hat\theta\right],\phi\right\rangle\right|\right]^2 \leqslant d\mathbb{E}\left[\phi^\top\mathbb{V}\left(\theta^*\right)\phi\right]$$

$$\mathbb{E}\left[\left|\left\langle\hat\theta-\theta^*,\phi\right\rangle\right|\right]^2 \leqslant 2d\mathbb{E}\left[\phi^\top\mathbb{V}\left(\theta^*\right)\phi\right].$$

Note that the above inequality holds even when $\phi \in \mathbb{R}^d$ is dependent on θ^* or $\hat\theta$.

Proof. Notice the Lemma 7 do not require the independence between X and Z. Therefore, we can apply Lemma 7 with setting $X=\hat\theta-\mathbb{E}\left[\hat\theta\right]$ and $Z=\phi$ to yield

$$\mathbb{E}\left[\left|\left\langle\hat\theta-\mathbb{E}\left[\hat\theta\right],\phi\right\rangle\right|\right]^2 \leqslant d\operatorname{Tr}\left(\mathbb{V}\left(\hat\theta\right)\mathbb{E}\left[\phi\phi^\top\right]\right) \overset{(b)}{=} d\mathbb{E}\left[\phi^\top\mathbb{V}(\theta^*)\phi\right]$$

where (b) is from $\mathbb{V}\left(\theta^*\right)=\mathbb{V}\left(\hat\theta\right)$ by the i.i.d condition and the linearity of expectation and trace operator.

Similarly, to upper bound the RHS of the second inequality in the corollary, we can apply Lemma 7 with setting $X=\hat\theta-\theta^*$ and $Z=\phi$. Notice that by i.i.d condition between θ^* and $\hat\theta$,

$$\mathbb{E}\left[\left(\theta^*-\hat\theta\right)\left(\theta^*-\hat\theta\right)^\top\right] = \mathbb{E}\left[\left(\theta^*-\mathbb{E}[\theta^*]+\mathbb{E}\left[\hat\theta\right]-\hat\theta\right)\left(\theta^*-\mathbb{E}[\theta^*]+\mathbb{E}\left[\hat\theta\right]-\hat\theta\right)^\top\right]$$

$$= \mathbb{V}\left(\theta^*-\mathbb{E}[\theta^*]\right)+\mathbb{V}\left(\hat\theta-\mathbb{E}\left[\hat\theta\right]\right) = 2\mathbb{V}\left(\theta^*\right)$$

Then we derive, $\mathbb{E}\left[\left|\left\langle\hat\theta-\theta^*,\phi\right\rangle\right|\right]^2 \leqslant 2d\mathbb{E}\left[\phi^\top\mathbb{V}(\theta^*)\phi\right]$. □

B.3 Proof of Proposition 2

Proof. Recall the definition of $\mathcal{H}_{\ell,h}=(\mathcal{H}_\ell,\pi_\ell,\hat O_{\ell,h})$, we have

$$\mathbb{E}\left[|\Delta_{\ell,h}(s_{\ell,h})|\mid\mathcal{H}_\ell\right]=\mathbb{E}\left[\mathbb{E}\left[|\Delta_{\ell,h}(s_{\ell,h})|\mid\mathcal{H}_{\ell,h}\right]\mid\mathcal{H}_\ell\right]=\mathbb{E}_\ell\left[\mathbb{E}_{\ell,h}\left[|\Delta_{\ell,h}(s_{\ell,h})|\right]\right] \quad (4)$$

With application of Lemma 6, we have

$$\mathbb{E}_\ell\left[\mathbb{E}_{\ell,h}\left[|\Delta_{\ell,h}(s_{\ell,h})|\right]\right] \leqslant \mathbb{E}_\ell\left[\mathbb{E}_{\ell,h}\left[\sum_{j=h}^{H-1}\mathbb{E}_{\ell,j}\left[\left|\left\langle\hat\theta_{\ell,j}-\theta_j^*,X_{\ell,j}\right\rangle\right|\right]\right]\right]=\sum_{j=h}^{H-1}\mathbb{E}_\ell\left[\left|\left\langle\hat\theta_{\ell,j}-\theta_j^*,X_{\ell,j}\right\rangle\right|\right]. \quad (5)$$

Combining (4) and (5) yields,

$$\mathbb{E}_\ell\left[|\Delta_{\ell,h}(s_{\ell,h})|\right] \leqslant \sum_{j=h}^{H-1}\mathbb{E}_\ell\left[\left|\left\langle\hat\theta_{\ell,j}-\theta_j^*,X_{\ell,j}\right\rangle\right|\right] \quad (6)$$

Recall the definition $\Delta_{\ell,0}(s)=V_{\pi_\ell,h}^{\hat M_\ell}(s)-V_{\pi_\ell,h}^M(s)$. Since the initial state $s_{\ell,0}\sim\rho$, we have $\overline V_{\pi_\ell}^{\hat M_\ell}-\overline V_{\pi_\ell}^M=\mathbb{E}\left[\Delta_{\ell,0}(s_{\ell,0})\mid\hat M_\ell,M,\pi_\ell\right]$. It follows that,

$$\mathbb{E}_\ell\left[\overline V_{\pi_\ell}^{\hat M_\ell}-\overline V_{\pi_\ell}^M\right]=\mathbb{E}_\ell\left[\Delta_{\ell,0}(s_{\ell,0})\right]=\mathbb{E}_\ell\left[\mathbb{E}_{\ell,0}\left[\Delta_{\ell,0}(s_{\ell,0})\right]\right] \quad (7)$$

By applying Lemma 6,

$$\mathbb{E}_\ell\left[\mathbb{E}_{\ell,0}\left[\Delta_{\ell,0}(s_{\ell,0})\right]\right]=\mathbb{E}_\ell\left[\mathbb{E}_{\ell,0}\left[\sum_{j=0}^{H-1}\mathbb{E}_{\ell,j}\left[\left\langle\hat\theta_{\ell,j}-\theta_j^*,X_{\ell,j}\right\rangle\right]\right]\right]=\sum_{h=0}^{H-1}\mathbb{E}_\ell\left[\left\langle\hat\theta_{\ell,h}-\theta_h^*,X_{\ell,h}\right\rangle\right] \quad (8)$$

If Assumption 2 holds, by Corollary 1, we can further derive,

$$\mathbb{E}_\ell \left[\mathbb{E}_{\ell,0} \left[\Delta_{\ell,0}(s_{\ell,0}) \right] \right] = \mathbb{E}_\ell \left[\mathbb{E}_{\ell,0} \left[\sum_{j=0}^{H-1} \mathbb{E}_{\ell,j} \left[\left\langle \hat{\theta}_{\ell,j} - \mathbb{E}_\ell \left[\hat{\theta}_{\ell,j} \right], X_{\ell,j} \right\rangle \right] \right] \right] = \sum_{h=0}^{H-1} \mathbb{E}_\ell \left[\left\langle \hat{\theta}_{\ell,h} - \mathbb{E}_\ell \left[\hat{\theta}_{\ell,h} \right], X_{\ell,h} \right\rangle \right] \quad (9)$$

Combining (7) and (9) yields,

$$\mathbb{E}_\ell \left[\overline{V}_{\pi_\ell}^{\hat{M}_\ell} - \overline{V}_{\pi_\ell}^M \right] = \sum_{h=0}^{H-1} \mathbb{E}_\ell \left[\left\langle \hat{\theta}_{\ell,h} - \mathbb{E}_\ell \left[\hat{\theta}_{\ell,h} \right], X_{\ell,h} \right\rangle \right] \quad (10)$$

Notice that $\hat{\theta}_{\ell,h}$ is an i.i.d copy of θ_h^* conditioned on history $\mathcal{H}_\ell$ by the algorithm of posterior sampling. Recall the definition of $\mathbf{\Gamma}_{\ell,h} = \mathbb{V}_\ell(\theta_h^*)$, with Corollary 2, we have

$$\mathbb{E}_\ell \left[\left| \left\langle \hat{\theta}_{\ell,h} - \mathbb{E}_\ell \left[\hat{\theta}_{\ell,h} \right], X_{\ell,h} \right\rangle \right| \right]^2 \leqslant d\mathbb{E}_\ell \left[X_{\ell,h}^\top \mathbf{\Gamma}_{\ell,h} X_{\ell,h} \right] \quad (11)$$

$$\mathbb{E}_\ell \left[\left| \left\langle \hat{\theta}_{\ell,h} - \theta_h^*, X_{\ell,h} \right\rangle \right| \right]^2 \leqslant 2d\mathbb{E}_\ell \left[X_{\ell,h}^\top \mathbf{\Gamma}_{\ell,h} X_{\ell,h} \right]. \quad (12)$$

The RHS of the above inequality is possibly larger than the valid range of LHS. Therefore, we try to tighten the upper bound by clipping. First, note that for any $x = (h,s,a) \in \mathcal{X}$, any MDP $\hat{M}$ and any policy π, we have the following equality holds

$$\left| (\hat{P} - P) V_{\pi,h+1}^{\hat{M}}(x) \right| = \left| \hat{P} V_{\pi,h+1}^{\hat{M}}(x) - P V_{\pi,h+1}^{\hat{M}}(x) \right| \leqslant H \quad a.s.$$

Recall that $(\hat{\theta}_{\ell,h} - \theta_h^*)^\top X_{\ell,h} = (\hat{P} - P)\hat{V}_{\ell,h}(x_{\ell,h}) \leqslant H$. Therefore, for any constant $\sigma_{\min} \geqslant H/\sqrt{d}$ and any random variable $\bar{\sigma}_{\ell,h} \geqslant \sigma_{\min}$ almost surely,

$$\mathbb{E}_\ell \left[\left| \left\langle \hat{\theta}_{\ell,h} - \mathbb{E}_\ell \left[\hat{\theta}_{\ell,h} \right], X_{\ell,h} \right\rangle \right| \right] \leqslant \sqrt{d} \mathbb{E}_\ell \left[\bar{\sigma}_{\ell,h}^2 \cdot 1 \wedge (X_{\ell,h}/\bar{\sigma}_{\ell,h})^\top \mathbf{\Gamma}_{\ell,h} (X_{\ell,h}/\bar{\sigma}_{\ell,h}) \right]^{1/2} \quad (13)$$

$$\mathbb{E}_\ell \left[\left| \left\langle \hat{\theta}_{\ell,h} - \theta_h^*, X_{\ell,h} \right\rangle \right| \right] \leqslant \sqrt{2d} \mathbb{E}_\ell \left[\bar{\sigma}_{\ell,h}^2 \cdot 1 \wedge (X_{\ell,h}/\bar{\sigma}_{\ell,h})^\top \mathbf{\Gamma}_{\ell,h} (X_{\ell,h}/\bar{\sigma}_{\ell,h}) \right]^{1/2} \quad (14)$$

Recall we choose $\bar{\sigma}_{\ell,h}^2 := \mathbb{E}_{\ell,h} \left[\sigma_{\ell,h}^2 \right] \vee \sigma_{\min}^2$ for analysis and thus Equation (13) and Equation (14) are valid. Finally, by plugging (13) into (10),

$$\mathbb{E} \left[\sum_{\ell=1}^L \overline{V}_{\pi_\ell}^{\hat{M}_\ell} - \overline{V}_{\pi_\ell}^M \right] = \mathbb{E} \left[\sum_{\ell=1}^L \mathbb{E}_\ell \left[\overline{V}_{\pi_\ell}^{\hat{M}_\ell} - \overline{V}_{\pi_\ell}^M \right] \right]$$

$$\leqslant \sqrt{d} \mathbb{E} \left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} \mathbb{E}_\ell \left[\bar{\sigma}_{\ell,h}^2 \cdot 1 \wedge (X_{\ell,h}/\bar{\sigma}_{\ell,h})^\top \mathbf{\Gamma}_{\ell,h} (X_{\ell,h}/\bar{\sigma}_{\ell,h}) \right]^{1/2} \right]$$

$$\leqslant \sqrt{d} \sum_{\ell=1}^L \sum_{h=0}^{H-1} \mathbb{E} \left[\bar{\sigma}_{\ell,h}^2 \cdot 1 \wedge (X_{\ell,h}/\bar{\sigma}_{\ell,h})^\top \mathbf{\Gamma}_{\ell,h} (X_{\ell,h}/\bar{\sigma}_{\ell,h}) \right]^{1/2}$$

$$\overset{(\star)}{\leqslant} \sqrt{d} \sum_{\ell=1}^L \sum_{h=0}^{H-1} \mathbb{E} \left[\bar{\sigma}_{\ell,h}^2 \right]^{1/2} \cdot \mathbb{E} \left[1 \wedge (X_{\ell,h}/\bar{\sigma}_{\ell,h})^\top \mathbf{\Gamma}_{\ell,h} (X_{\ell,h}/\bar{\sigma}_{\ell,h}) \right]^{1/2}$$

$$\leqslant \sqrt{d\mathbb{E} \left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} \bar{\sigma}_{\ell,h}^2 \right] \cdot \mathbb{E} \left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} 1 \wedge (X_{\ell,h}/\bar{\sigma}_{\ell,h})^\top \mathbf{\Gamma}_{\ell,h} (X_{\ell,h}/\bar{\sigma}_{\ell,h}) \right]}$$

Here, $\sigma_{\min} = H$.

□

C   Proof sketch of Theorem 1.

C.1   Quantify information gain of the environment via posterior variance reduction

The first step is to mathematically quantify the reduction of uncertainty of the unknown environment when new information gathered in. We use the quantity *posterior variance* of the unknown model parameters $\mathbf{\Gamma}_{\ell,h} := \mathbb{V}(\theta_h^* \mid \mathcal{H}_\ell)$ defined in Definition 1 as a uncertainty measurement. From the perspective of value-targeted model learning stated in Section 4, with the input as $X_{\ell,h} := \phi_{\hat{V}_{\ell,h+1}}(x_{\ell,h}) = \sum_{s' \in \mathcal{S}} \phi(s' \mid x_{\ell,h}) \hat{V}_{\ell,h+1}(s')$ and response as $Y_{\ell,h+1} := \hat{V}_{\ell,h+1}(s_{\ell,h+1})$, we establish Theorem 2 on posterior variance reduction on the unknown model parameters as following:

$$\mathbb{E}\left[\mathbf{\Gamma}_{\ell+1,h} \mid \mathcal{H}_{\ell,h}\right]^{-1} \succeq \mathbf{\Gamma}_{\ell,h}^{-1} + \frac{1}{\mathbb{E}\left[\sigma_{\ell,h}^2 \mid \mathcal{H}_{\ell,h}\right]} X_{\ell,h} X_{\ell,h}^\top,$$

where $\sigma_{\ell,h}^2 = \mathbb{V}\left(\hat{V}_{\ell,h+1}(s_{\ell,h+1}) \mid M, \mathcal{H}_{\ell,h}\right)$ is the variance of value due to the transitional noise for sampling $s_{\ell,h+1} \sim P(x_{\ell,h})$ given the true MDP M and history $\mathcal{H}_{\ell,h} = (\mathcal{H}_\ell, \pi_\ell, x_{\ell,0}, \hat{V}_{\ell,1}, \ldots, x_{\ell,h}, \hat{V}_{\ell,h+1})$. This is the key enabling technique with which we can measure the reduction involving with the information of value variance under posterior probability measure without using any concentration bounds.

C.2   Regret decomposition to cumulative posterior value variance and cumulative potential

The second step is to decompose the regret into the the so-called cumulative variance and cumulative potentials which will be explained later.

- To facilitate the analysis, we define a bounded-below sequence $\bar{\sigma}_{\ell,h}$ such that

$$\bar{\sigma}_{\ell,h}^2 := \mathbb{E}\left[\sigma_{\ell,h}^2 \mid \mathcal{H}_{\ell,h}\right] \vee \sigma_{\min}^2.$$

  In the analysis, we have twp different choice of $\sigma_{\min}$, leading to two different bounds.

  – Choice 1: choose $\sigma_{\min} := H$

- Since the regret is equal to the estimation error in expectation. By Proposition 2, we bound the estimation error by

$$\mathbb{E}\underbrace{\left[\sum_{\ell=1}^{L} \overline{V}_{\pi_\ell}^{\hat{M}_\ell} - \overline{V}_{\pi_\ell}^{M}\right]}_{\text{Estimation error}} \leqslant \sqrt{d\ \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1} \bar{\sigma}_{\ell,h}^2\right]}_{\text{Cumulative variance}} \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1} 1 \wedge \left\|\frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}}\right\|_{\mathbf{\Gamma}_{\ell,h}}^2\right]}_{\text{Cumulative potential}}}$$

- The cumulative variance term is approximately the expected summation of the value variance $\mathbb{V}\left(\hat{V}_{\ell,h+1}(s_{\ell,h+1}) \mid M, \mathcal{H}_{\ell,h}\right)$ due to the transitional noise.

- The cumulative potential is the expected summation of the estimation error of θ_h^* in the direction of the variance-normalized feature $X_{\ell,h}/\bar{\sigma}_{\min}$ conditioned on the historical data available in the beginning of episode ℓ, i.e. $\left\|\frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}}\right\|_{\mathbf{\Gamma}_{\ell,h}}^2$.

C.3   Bounding cumulative variance

Proposition 3 (Cumulative variance of value under virtual model)**.** *Recall the definition* $\bar{\sigma}_{\ell,h}^2 = \mathbb{E}_{\ell,h}\left[\sigma_{\ell,h}^2\right] \vee \sigma_{\min}^2$ *where* $\sigma_{\ell,h}^2 = \mathbb{V}\left(\hat{V}_{\ell,h+1}(s_{\ell,h+1}) \mid M, \mathcal{H}_{\ell,h}\right)$.

1. *If $\sigma_{\min} = H$, we have*

$$\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1}\bar{\sigma}_{\ell,h}^2\right] \leqslant LH^3$$

For choice 2 of $\sigma_{\min} = H/\sqrt{d}$: We bound the cumulative variance with $\mathcal{O}(LH^2)$ by Proposition 3 provided the condition $d \gtrsim H$. The proof is to relate the variance of virtual value functions $\mathbb{V}\left(\hat{V}_{\ell,h}(s_{\ell,h+1}) \mid M, \mathcal{H}_{\ell,h}\right)$ to the variance of true value functions $\mathbb{V}\left(V_{\pi_\ell,h}^M(s_{\ell,h+1}) \mid M, \mathcal{H}_{\ell,h}\right)$. The proof sketch is as follows

- By Lemma 13, we upper bound the variance difference

$$\mathbb{V}\left(V_{\pi_\ell,h}^M(s_{\ell,h+1}) \mid M, \mathcal{H}_{\ell,h}\right) - \mathbb{V}\left(\hat{V}_{\ell,h}(s_{\ell,h+1}) \mid M, \mathcal{H}_{\ell,h}\right),$$

  which is a dominated term.

- (**Dominant term.**) Then, by Lemma 12, we use a conditioned version of law of total variance to relate the expected summation of variance of value to the variance of summation of rewards:

$$\mathbb{E}\left[\sum_{h=0}^{H-1}\mathbb{V}\left(V_{\pi_\ell,h+1}^M(s_{\ell,h+1}) \mid M, \mathcal{H}_{\ell,h}\right)\right] = \mathbb{E}\left[\mathbb{V}\left(\sum_{j=0}^{H-1}R(x_{\ell,j}) \mid M, \mathcal{H}_{\ell,0}\right)\right] \leqslant H^2. \tag{15}$$

  A naive upper bound the summation of the value variance will be H^3. Therefore, with the time-varying variance information in hand, (15) has a direct consequence on the $\sqrt{H}$ factor improvement in the final regret bound.

C.4 Bounding cumulative potential

Proposition 4 (Cumulative potential). *Under the above assumptions and definitions, for any $L \in \mathbb{Z}_+$ and any $h \in [H]$,*

$$\mathbb{E}\left[\sum_{\ell=1}^{L} 1 \wedge \left\|\frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}}\right\|_{\boldsymbol{\Gamma}_{\ell,h}}^2\right] \leqslant 2\log\det\left(\boldsymbol{I} + \frac{LH^2}{\sigma_{\min}^2}\boldsymbol{\Gamma}_{1,h}\right)$$

where $\boldsymbol{\Gamma}_{1,h}$ is the prior covariance of θ_h^*.

If Assumption 3 holds, we can further upper bound the RHS by the Fact 1

$$\log\det(\boldsymbol{I} + (LH^2/\sigma_{\min}^2)\boldsymbol{\Gamma}_{1,h}) \leqslant d\log(1 + LH^2B^2/\sigma_{\min}^2) \tag{16}$$

1. $\sigma_{\min} = H$, the upper bound becomes $\mathcal{O}(d\log L)$

C.4.1 Put everything together

Putting everything together, by Proposition 2

1. $\sigma_{\min} = H$, we obtained the upper bound $\mathcal{O}(dH^{3/2}\sqrt{T\log T})$ in Theorem 1 and Remark 1.

D Technical details for regret decomposition

D.1 Generic estimation error decomposition

Lemma 8 (Estimation error decomposition). *For any MDP $\hat{M}$ and policy π, let $\hat{P}$ be the transition model under $\hat{M}$ and $s_1, \ldots, s_{H-1}$ is rolled out under the policy π over the MDP M. Define $\Delta_h(s) = V_{\pi,h}^{\hat{M}}(s) - V_{\pi,h}^M(s)$ and we have,*

$$\Delta_h(s_h) = \left(\hat{P} - P\right)V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h)) + \Delta_{h+1}(s_{h+1}) + d_h(\Delta_{h+1})$$

$$|\Delta_h(s_h)| \leqslant \left|\left(\hat{P} - P\right)V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h))\right| + |\Delta_{h+1}(s_{h+1})| + d_h(|\Delta_{h+1}|).$$

where $d_h(\Delta) = [P\Delta](h, s_h, \pi(s_h)) - \Delta(s_{h+1})$.

Proof. Now we derive the decomposition of the value gap,

$$
\begin{aligned}
\Delta_h(s_h) &= V_{\pi,h}^{\hat{M}}(s_h) - V_{\pi,h}^{M}(s_h) \\
&= \hat{P} V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h)) - P V_{\pi,h+1}^{M}(h, s_h, \pi(s_h)) \\
&= \left(\hat{P} - P \right) V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h)) + P \left(V_{\pi,h+1}^{\hat{M}} - V_{\pi,h+1}^{M} \right)(h, s_h, \pi(s_h)) \\
&= \left(\hat{P} - P \right) V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h)) + P \Delta_{h+1}(h, s_h, \pi(s_h)) \\
&= \left(\hat{P} - P \right) V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h)) + \Delta_{h+1}(s_{h+1}) + d_h(\Delta_{h+1}).
\end{aligned}
$$

And similarly the decomposition for absolute value gap,

$$
\begin{aligned}
|\Delta_h(s_h)| &\leqslant \left| \left(\hat{P} - P \right) V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h)) \right| + \left| P \left(V_{\pi,h+1}^{\hat{M}} - V_{\pi,h+1}^{M} \right)(h, s_h, \pi(s_h)) \right| \\
&\leqslant \left| \left(\hat{P} - P \right) V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h)) \right| + P \left| \left(V_{\pi,h+1}^{\hat{M}} - V_{\pi,h+1}^{M} \right) \right| (h, s_h, \pi(s_h)) \\
&= \left| \left(\hat{P} - P \right) V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h)) \right| + P \left| \Delta_{h+1} \right| (h, s_h, \pi(s_h)) \\
&= \left| \left(\hat{P} - P \right) V_{\pi,h+1}^{\hat{M}}(h, s_h, \pi(s_h)) \right| + |\Delta_{h+1}(s_{h+1})| + d_h(|\Delta_{h+1}|).
\end{aligned}
$$

where $|\Delta|(s) = |\Delta(s)|$. $\qquad\square$

Lemma 9. *If a function Δ is measurable w.r.t the sigma-algebra generated by random variables Z and $P(h, \cdot, \cdot)$, then, for any $x = (h, s, a) \in \mathcal{X}_h$ and $s'_x \sim P(x)$, the conditional expectation*

$$
\mathbb{E}\left[P\Delta(x) - \Delta(s'_x) \mid Z, P(h, \cdot, \cdot) \right] = 0,
$$

where $P\Delta(x) := \mathbb{E}_{s'_x \sim P(x)}[\Delta(x)]$.

D.2 Proof of Lemma 1 in Section 4

Proof. Note $\overline{V}_{\pi}^{\hat{M}} - \overline{V}_{\pi}^{M} = \mathbb{E}\left[\Delta_0(s_0) \mid \hat{M}, M, \pi \right]$. We can verify that $\Delta_h = V_{\pi,h}^{\hat{M}} - V_{\pi,h}^{M}$ is measurable w.r.t the sigma-algebra generated by $\pi, P([h:H], \cdot, \cdot)$ and $\hat{P}([h:H], \cdot, \cdot)$. Also note that conditioned on π, the action is $\pi(s_h) = a_h$. Therefore, together with Lemma 9, the conditional expectation is

$$
\mathbb{E}\left[[P\Delta_{h+1}](h, s_h, \pi(s_h))) \mid \hat{M}, M, \pi \right] = \mathbb{E}\left[[P\Delta_{h+1}](x_h) \mid \hat{M}, M, \pi \right] = \mathbb{E}\left[\Delta_{h+1}(s_{h+1}) \mid \hat{M}, M, \pi \right].
$$

By recursive application of Lemma 8 form $h = 0$ to $h = H - 1$ and noticing

$$
\mathbb{E}\left[d_h(\Delta_{h+1}) \mid \hat{M}, M, \pi \right] = 0,
$$

we conclude the result. $\qquad\square$

D.3 Proof of Lemma 2 in Section 4.1

Proof. Recall the definition of value-augmented observations $\hat{O}_{\ell,h}$ and the definition of history $\mathcal{H}_\ell$ and $\mathcal{H}_{\ell,h}$ in Section 4. Also recall the short notation $\mathbb{E}_{\ell,h}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{H}_{\ell,h}]$ and $\mathbb{E}_\ell[\cdot] := \mathbb{E}[\cdot \mid \mathcal{H}_\ell]$,

$$
\mathbb{E}_{\ell,h}[\Delta_{\ell,h}(s_{\ell,h})] = \mathbb{E}[\Delta_{\ell,h}(s_{\ell,h}) \mid \mathcal{H}_{\ell,h}] = \mathbb{E}\left[\Delta_{\ell,h}(s_{\ell,h}) \mid \mathcal{H}_\ell, \pi_\ell, \hat{O}_{\ell,h} \right] \tag{17}
$$

An application of Lemma 8 yields,

$$
\begin{aligned}
\mathbb{E}_{\ell,h}[\Delta_{\ell,h}(s_{\ell,h})] &= \mathbb{E}_\ell\left[\left(\hat{P}_\ell - P \right) \hat{V}_{\ell,h+1}(h, s_{\ell,h}, \pi_\ell(s_{\ell,h})) + \Delta_{\ell,h+1}(s_{\ell,h+1}) + d_{\ell,h}(\Delta_{\ell,h+1}) \mid \pi_\ell, \hat{O}_{\ell,h} \right] \\
&= \mathbb{E}_\ell\left[\left(\hat{P}_\ell - P \right) \hat{V}_{\ell,h+1}(x_{\ell,h}) + \Delta_{\ell,h+1}(s_{\ell,h+1}) + d_{\ell,h}(\Delta_{\ell,h+1}) \mid \pi_\ell, \hat{O}_{\ell,h} \right] \\
&\overset{(a)}{=} \mathbb{E}_\ell\left[\left(\hat{P}_\ell - P \right) \hat{V}_{\ell,h+1}(x_{\ell,h}) + \Delta_{\ell,h+1}(s_{\ell,h+1}) \mid \pi_\ell, \hat{O}_{\ell,h} \right] \\
&= \mathbb{E}_{\ell,h}\left[\left(\hat{P}_\ell - P \right) \hat{V}_{\ell,h+1}(x_{\ell,h}) + \mathbb{E}_{\ell,h+1}[\Delta_{\ell,h+1}(s_{\ell,h+1})] \right] \tag{18}
\end{aligned}
$$

where the second equality is due to π_ℓ-conditioning with short notation $x_{\ell,h} = (h, s_{\ell,h}, a_{\ell,h})$. The equality (a) holds by Lemma 9 with the fact that the function $\Delta_{\ell,h+1} = V_{\pi_\ell,h+1}^{\hat{M}} - V_{\pi_\ell,h+1}^M = \hat{V}_{\ell,h+1} - V_{\pi_\ell,h+1}^M$ is measurable w.r.t the sigma-algebra generated by $M, \pi_\ell, \hat{V}_{\ell,h+1}$ and the fact M contains h-stage transition kernel $P(h, \cdot, \cdot)$,

$$
\begin{aligned}
\mathbb{E}_\ell \left[d_{\ell,h}(\Delta_{\ell,h+1}) \mid M, \pi_\ell, O_{\ell,h} \right] &= \mathbb{E}_\ell \left[P\Delta_{\ell,h+1}(x_{\ell,h}) - \Delta_{\ell,h+1}(s_{\ell,h+1}) \mid M, \pi_\ell, O_{\ell,h-1}, s_{\ell,h}, a_{\ell,h}, \hat{V}_{\ell,h+1} \right] \\
&= \mathbb{E}_\ell \left[P\Delta_{\ell,h+1}(x_{\ell,h}) - \Delta_{\ell,h+1}(s_{\ell,h+1}) \mid M, \pi_\ell, \hat{V}_{\ell,h+1}, O_{\ell,h-1}, x_{\ell,h} \right] \\
&= 0.
\end{aligned}
$$

Further expand the recursion of Equation (18), we obtain

$$
\mathbb{E}_{\ell,h} \left[\Delta_{\ell,h}(s_{\ell,h}) \right] = \mathbb{E}_{\ell,h} \left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j} \left[\left(\hat{P}_\ell - P \right) \hat{V}_{\ell,j+1}(x_{\ell,j}) \right] \right].
$$

Similarly, by applying Lemma 8, we have

$$
\mathbb{E}_{\ell,h} \left[|\Delta_{\ell,h}(s_{\ell,h})| \right] \leqslant \mathbb{E}_{\ell,h} \left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j} \left[\left| \left(\hat{P}_\ell - P \right) \hat{V}_{\ell,j+1}(x_{\ell,j}) \right| \right] \right].
$$

$\square$

D.4 Proof of Lemma 6: Error decomposition in linear mixture MDPs from value-targeted perspective

Recall the short notation of estimation error in stage h of episode ℓ, $\Delta_{\ell,h}(s) = V_{\pi_\ell,h}^{\hat{M}_\ell}(s) - V_{\pi_\ell,h}^M(s)$. We have the following lemmas prepared for Proposition 2.

Proof. For any $x = (h, s, a) \in \mathcal{X}$, recall that $\phi_V(x) = \sum_{s'} \phi(s' \mid x) V(s')$. For any MDP $\hat{M}$ in the class of linear mixture MDPs, the transition kernel can be represented as $\hat{P}(x) = \langle \hat{\theta}_h, \phi(\cdot \mid x) \rangle$. Then, we have the following property

$$
\hat{P}V(x) = \sum_{s' \in \mathcal{S}} \left\langle \hat{\theta}_h, \phi(s' \mid x) \right\rangle V(s') = \left\langle \hat{\theta}_h, \sum_{s' \in \mathcal{S}} \phi(s' \mid x) V(s') \right\rangle = \left\langle \hat{\theta}_h, \phi_V(x) \right\rangle.
$$

With the short notation $X_{\ell,h} = \phi_{\hat{V}_{\ell,h+1}}(x_{\ell,h})$, we have $\left(\hat{P}_\ell - P \right) \hat{V}_{\ell,h+1}(x_{\ell,h}) = \left\langle \hat{\theta}_{\ell,h} - \theta_h^*, X_{\ell,h} \right\rangle$. As a corollary of Lemma 2,

$$
\mathbb{E}_{\ell,h} \left[\Delta_{\ell,h}(s_{\ell,h}) \right] = \mathbb{E}_{\ell,h} \left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j} \left[\left(\hat{P}_\ell - P \right) \hat{V}_{\ell,j+1}(x_{\ell,j}) \right] \right] = \mathbb{E}_{\ell,h} \left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j} \left[\left\langle \hat{\theta}_{\ell,j} - \theta_j^*, X_{\ell,j} \right\rangle \right] \right] \tag{19}
$$

and similarly from Lemma 2,

$$
\mathbb{E}_{\ell,h} \left[|\Delta_{\ell,h}(s_{\ell,h})| \right] \leqslant \mathbb{E}_{\ell,h} \left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j} \left[\left| \left\langle \hat{\theta}_{\ell,j} - \theta_j^*, X_{\ell,j} \right\rangle \right| \right] \right]
$$

$\square$

D.5 Proof of Corollary 1

Proof. Since $X_{\ell,h} = \phi_{\hat{V}_{\ell,h+1}}(x_{\ell,h})$ is a deterministic function of $x_{\ell,h}$ and $\hat{V}_{\ell,h+1}$, by recalling the definition of $\mathcal{H}_{\ell,h}$, we conclude $X_{\ell,h}$ is $\sigma(\mathcal{H}_{\ell,h})$-measurable, resulting the following equation,

$$
\mathbb{E}_{\ell,h} \left[\langle \theta_h^*, X_{\ell,h} \rangle \right] = \langle \mathbb{E}_{\ell,h} \left[\theta_h^* \right], X_{\ell,h} \rangle \tag{20}
$$

Notice $x_{\ell,h}$ is rolled out from initial stage to the stage h by policy π_ℓ under M. Therefore, conditioned on $\mathcal{H}_\ell$, the state-action pair $x_{\ell,h} = (h, s_{\ell,h}, a_{\ell,h})$ is dependent on $(\theta_0^*, \ldots, \theta_{h-1}^*)$ and $\hat{\Theta}_\ell$ but is independent of θ_h^* by Assumption 2. Also notice the fact that conditioned on $\mathcal{H}_\ell$, the imagined value function $\hat{V}_{\ell,h+1}$ is dependent on $\hat{\Theta}_\ell$ but is independent of Θ^*. We conclude that $X_{\ell,h} = \phi_{\hat{V}_{\ell,h+1}}(x_{\ell,h})$ is independent of θ_h^* conditioned on $\mathcal{H}_\ell$.

Similarly, conditioned on $\mathcal{H}_\ell$, the random variables $x_{\ell,0}, \hat{V}_{\ell,1}, \ldots, x_{\ell,h-1}, \hat{V}_{\ell,h}, x_{\ell,h}, \hat{V}_{\ell,h+1}$ and π_ℓ are all independent of θ_h^*. Thus, we also have

$$\mathbb{E}_{\ell,h}\left[\theta_h^*\right] = \mathbb{E}\left[\theta_h^* \mid \mathcal{H}_\ell, \pi_\ell, \hat{O}_{\ell,h}\right] = \mathbb{E}\left[\theta_h^* \mid \mathcal{H}_\ell\right] = \mathbb{E}_\ell\left[\theta_h^*\right]$$

By posterior sampling, we have

$$\mathbb{E}_\ell\left[\theta_h^*\right] = \mathbb{E}_\ell\left[\hat{\theta}_{\ell,h}\right].$$

It follows that

$$\mathbb{E}_{\ell,h}\left[\langle\hat{\theta}_{\ell,h} - \theta_h^*, X_{\ell,h}\rangle\right] = \mathbb{E}_{\ell,h}\left[\langle\hat{\theta}_{\ell,h} - \mathbb{E}_\ell\left[\hat{\theta}_{\ell,h}\right], X_{\ell,h}\rangle\right].$$

Then, one can derive from (19),

$$\mathbb{E}_{\ell,h}\left[\Delta_{\ell,h}(s_{\ell,h})\right] = \mathbb{E}_{\ell,h}\left[\sum_{j=h}^{H-1} \mathbb{E}_{\ell,j}\left[\langle\hat{\theta}_{\ell,j} - \mathbb{E}_\ell\left[\hat{\theta}_{\ell,j}\right], X_{\ell,j}\rangle\right]\right]$$

$\square$

E Proofs of propositions in Appendix C

E.1 Proof of Proposition 3: Bounding cumulative value variance

Proof. For the first case $\sigma_{\min} = H$, since $\hat{V}_{\ell,h+1} \leqslant H$ almost surely, we have $\bar{\sigma}_{\ell,h}^2 = H^2$ almost surely, then the cumulative variance is trivially bounded.

For the second case, the Lemma and Corollary used in the proof can be found in Appendix F.3. To upper bound the cumulative variance of value under virtual model, we need to first bound the variance of true value by Corollary 3 and then relate the variance of virtual value to the variance of the true value by Lemma 13 and Proposition 2.

The first step is to upper bound the LHS to the summation of posterior variance of virtual variance. Since $\bar{\sigma}_{\ell,h}^2 := \mathbb{E}_{\ell,h}\left[\sigma_{\ell,h}^2\right] \vee \sigma_{\min}^2$,

$$\mathbb{E}\left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} \bar{\sigma}_{\ell,h}^2\right] \leqslant \mathbb{E}\left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} \sigma_{\ell,h}^2 + \sigma_{\min}^2\right]$$

$$= \mathbb{E}\left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} \mathbb{V}_{\ell,h}\left(\hat{V}_{\ell,h+1}(s_{\ell,h+1}) \mid M\right)\right] + LH\sigma_{\min}^2 \tag{21}$$

By Corollary 3, there is an sharp upper bound on the cumulative variance of true value,

$$\mathbb{E}\left[\sum_{\ell=1}^L \sum_{h=0}^{H-1} \mathbb{V}_{\ell,h}\left(V_{\pi_\ell,h+1}^M(s_{\ell,h+1}) \mid M\right)\right] \leqslant LH^2. \tag{22}$$

Note that an naive upper bound on LHS of Equation (22) would be LH^3. The sharp LH^2 upper is indeed the source of improvement on the dependence of horizon factor H in the final regret bound. Therefore, the second step is to relate the variance of $\hat{V}_{\ell,h} = V_{\pi_\ell,h}^{\hat{M}^\ell}$ to the variance of $V_{\pi_\ell,h}^M$, i.e.

$$\mathbb{V}_{\ell,h}\left(V_{\pi_\ell,h+1}^{\hat{M}^\ell}(s_{\ell,h+1}) \mid M\right) = \mathbb{V}_{\ell,h}\left(V_{\pi_\ell,h+1}^M(s_{\ell,h+1}) \mid M\right) + \delta_{\ell,h},$$

where $\delta_{\ell,h} := \mathbb{V}_{\ell,h}\left(V_{\pi_\ell,h+1}^{\hat{M}^\ell}(s_{\ell,h+1} \mid M)\right) - \mathbb{V}_{\ell,h}\left(V_{\pi_\ell,h+1}^M(s_{\ell,h+1}) \mid M\right)$ is bounded by Lemma 13,

$$\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1}\delta_{\ell,h}\right] \leqslant 2H\mathbb{E}\left[\sum_{\ell=0}^{L-1}\sum_{h=0}^{H-1}|\Delta_{\ell,h+1}(s_{\ell,h+1})|\right].$$

By Proposition 2,

$$\mathbb{E}\left[\sum_{\ell=0}^{L-1}\sum_{h=0}^{H-1}|\Delta_{\ell,h+1}(s_{\ell,h+1})|\right] \leqslant H\sqrt{2d\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1}\bar{\sigma}_{\ell,h}^2\right]\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1}1\wedge\left\|\frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}}\right\|_{\Gamma_{\ell,h}}^2\right]}$$

Combine the above results and (21), we have

$$\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1}\bar{\sigma}_{\ell,h}^2\right] \leqslant LH\sigma_{\min}^2 + LH^2 + 2H^2\sqrt{2d\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1}\bar{\sigma}_{\ell,h}^2\right]\underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1}1\wedge\left\|\frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}}\right\|_{\Gamma_{\ell,h}}^2\right]}_{S}}$$

Since $x \leqslant a\sqrt{x} + b$ implies $x \leqslant \frac{3}{2}(a^2 + b)$, finally we derive

$$\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1}\bar{\sigma}_{\ell,h}^2\right] \leqslant \frac{3}{2}\left(HL\sigma_{\min}^2 + H^2L + 8dH^4S\right),$$

where $S = \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1}1\wedge\left\|\frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}}\right\|_{\Gamma_{\ell,h}}^2\right]$. □

E.2 Proof of Proposition 4: Bounding cumulative potential

The proof of Proposition 4 is by the recursive application of Theorem 2 and the following Lemma 10. Note that various version of the potential lemma (Auer, 2003; Dani et al., 2008; Agrawal and Goyal, 2013; Li et al., 2019; Hamidi and Bayati, 2021) is used for analyzing linear bandit and reinforcement learning with function approximation. We apply the following version of potential lemma in the proof of Proposition 4. The proof of Lemma 10 can be found in Appendix F.4.

Lemma 10 (Potential lemma (Hamidi and Bayati, 2021)). *For $x > 0$ and positive semi-definite matrix Σ, define $f(\Sigma, x) = \log \det(I + x\Sigma)$. Then, for any vector $V \in \mathbb{R}^d$, we have*

$$\log\left(1 + V^\top \Sigma V\right) + f\left(\Sigma', x\right) \leqslant f\left(\Sigma, x + V^\top V\right)$$

where

$$\Sigma' := \Sigma - \frac{\Sigma V V^\top \Sigma}{1 + V^\top \Sigma V} = \Sigma^{\frac{1}{2}}\left(I - \frac{\Sigma^{\frac{1}{2}} V V^\top \Sigma^{\frac{1}{2}}}{1 + V^\top \Sigma V}\right)\Sigma^{\frac{1}{2}}$$

and equivalently $\Sigma'^{-1} = \Sigma^{-1} + VV^\top$, using Sherman-Morrison formula.

By the assumption of the feature vector and the fact value is bounded by [0,H], for each episode ℓ and each stage

$$\left\|\phi_{V_{M^\ell,h+1}^{\pi_\ell}}(s_h^\ell, a_h^\ell)\right\|^2 = \left\|\sum_{s'\in\mathcal{S}}\phi(s' \mid s_h^\ell, a_h^\ell)V_{M^\ell,h+1}^{\pi_\ell}(s')\right\|^2 \leqslant H^2$$

Since $\bar{\sigma}_{\ell,h}^2 = \max(\sigma_{\min}^2, \mathbb{E}_{\ell,h}\left[\sigma_{\ell,h}^2\right]) \geqslant \sigma_{\min}^2$, we have

$$\left\|\frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}}\right\| \leqslant \frac{H}{\sigma_{\min}}$$

For any $x \in (-1, +\infty)$, we have $1 \wedge x \leqslant 2\log_2(1+x)$ and therefore

$$1 \wedge \left\| \frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}} \right\|^2_{\mathbf{\Gamma}_{\ell,h}} \leqslant 2\log\left(1 + \left\| \frac{X_{\ell,h}}{\bar{\sigma}_{\ell,h}} \right\|^2_{\mathbf{\Gamma}_{\ell,h}}\right)$$

Define the short notation $\bar{X}_{\ell,h} = X_{\ell,h}/\bar{\sigma}_{\ell,h}$, we will prove the proposition by mathematical induction: For base case $L = 1$,

$$\log(1 + \left\|\bar{X}_{1,h}\right\|^2_{\mathbf{\Gamma}_{1,h}}) = \log\det\left(\mathbf{I} + \mathbf{\Gamma}^{1/2}_{1,h}\bar{X}_{1,h}\bar{X}^\top_{1,h}\mathbf{\Gamma}^{1/2}_{1,h}\right)$$

$$\leqslant \log\det\left(\mathbf{I} + \frac{H^2}{\sigma^2_{\min}}\mathbf{\Gamma}^{1/2}_{1,h}\mathbf{I}\mathbf{\Gamma}^{1/2}_{1,h}\right)$$

$$= \log\det\left(\mathbf{I} + \frac{H^2}{\sigma^2_{\min}}\mathbf{\Gamma}_{1,h}\right)$$

by the fact $(1 + X^\top X) = \det(\mathbf{I} + XX^\top)$ and $AA^\top \preceq \|A\|^2 I$.

We use the induction hypothesis for $L - \ell$ with prior $\mathbb{P}(\cdot \mid \mathcal{H}_{\ell+1})$, that is

$$\mathbb{E}_{\ell+1}\left[\sum_{k=\ell+1}^{L} 1 \wedge \left\|\bar{X}_{k,h}\right\|^2_{\mathbf{\Gamma}_{k,h}}\right] \leqslant 2\log\det\left(\mathbf{I} + \frac{(L-\ell)H^2}{\sigma^2_{\min}}\mathbf{\Gamma}_{\ell+1,h}\right)$$

Define function $f(\mathbf{\Sigma}, x) = \log\det(\mathbf{I} + x\mathbf{\Sigma})$ and denote

$$\mathbf{\Gamma}'_{\ell,h} := \mathbf{\Gamma}_{\ell,h} - \frac{\mathbf{\Gamma}_{\ell,h}\bar{X}_{\ell,h}\bar{X}^\top_{\ell,h}\mathbf{\Gamma}_{\ell,h}}{1 + \bar{X}^\top_{\ell,h}\mathbf{\Gamma}_{\ell,h}\bar{X}_{\ell,h}}.$$

We have

$$\mathbb{E}_{\ell,h}\left[\sum_{k=\ell+1}^{L} 1 \wedge \left\|\bar{X}_{k,h}\right\|^2_{\mathbf{\Gamma}_{k,h}}\right] = \mathbb{E}_{\ell,h}\left[\mathbb{E}_{\ell+1}\left[\sum_{k=\ell+1}^{L} 1 \wedge \left\|\bar{X}_{k,h}\right\|^2_{\mathbf{\Gamma}_{k,h}}\right]\right]$$

$$\leqslant 2\log\det\left(\mathbf{I} + \frac{(L-\ell)H^2}{\sigma^2_{\min}}\mathbb{E}_{\ell,h}\left[\mathbf{\Gamma}_{\ell+1,h}\right]\right)$$

$$\overset{(a)}{\leqslant} 2\log\det\left(\mathbf{I} + \frac{(L-\ell)H^2}{\sigma^2_{\min}}\left(\mathbf{\Gamma}_{\ell,h} - \frac{\mathbf{\Gamma}_{\ell,h}\bar{X}_{\ell,h}\bar{X}^\top_{\ell,h}\mathbf{\Gamma}_{\ell,h}}{1 + \bar{X}^\top_{\ell,h}\mathbf{\Gamma}_{\ell,h}\bar{X}_{\ell,h}}\right)\right)$$

$$= 2f(\mathbf{\Gamma}'_{\ell,h}, (L-\ell)H^2/\sigma^2_{\min})$$

where (a) is by the condition $\bar{\sigma}_{\ell,h} \geqslant \mathbb{E}_{\ell,h}\left[\sigma_{\ell,h}\right]$ and Theorem 2.

Now we use potential lemma (Lemma 10) to step forward,

$$\mathbb{E}_{\ell,h}\left[\sum_{k=\ell}^{L} 1 \wedge \left\|\bar{X}_{k,h}\right\|^2_{\mathbf{\Gamma}_{k,h}}\right] = 1 \wedge \left\|\bar{X}_{\ell,h}\right\|^2_{\mathbf{\Gamma}_{\ell,h}} + \mathbb{E}_{\ell,h}\left[\sum_{k=\ell+1}^{L} 1 \wedge \left\|\bar{X}_{k,h}\right\|^2_{\mathbf{\Gamma}_{k,h}}\right]$$

$$\leqslant 2\left(\log(1 + \bar{X}^\top_{\ell,h}\mathbf{\Gamma}_{\ell,h}\bar{X}_{\ell,h}) + f(\mathbf{\Gamma}'_{\ell,h}, (L-\ell)H^2/\sigma^2_{\min})\right)$$

$$\overset{(b)}{\leqslant} 2f(\mathbf{\Gamma}_{\ell,h}, (L-\ell+1)H^2/\sigma^2_{\min})$$

where (b) is by Lemma 10. Notice the fact that RHS is $\sigma(\mathcal{H}_\ell)$-measurable, by taking $\mathbb{E}_\ell[\cdot]$ on both sides of the inequality, we conclude that the statement holds for $L - \ell + 1$ with prior $\mathbb{P}(\cdot \mid \mathcal{H}_\ell)$.

F   Technical lemmas and facts

F.1   Posterior variance reduction for general prior and noise distributions

Posterior variance reflects how uncertain we are about a random variable based on existed information. The progress of learning can be measured by how much variance is reduced when new information comes in. This

section provides a tool to track the variance reduction when prior and noise distribution is *not* Gaussian. For the purpose of simplicity and generality, we state the lemma in a different set of short notations.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$ be an increasing sequence of σ-algebras that are meant to encode the information available up to time t. Let $\theta^\star : \Omega \to \mathbb{R}^d$ be the true parameters vector. Let X_t be a random variable that is possibly dependent on the history and some external source of randomness.

Assumption 4. *Assume X_t is $\mathcal{F}_t$-measurable.*

More information about $\theta^\star$ is then made available sequentially through a sequence of inputs $X_0, X_1, \cdots : \Omega \to \mathbb{R}^d$ and a sequence of observations $Y_1, Y_2, \cdots : \Omega \to \mathbb{R}$.

Assumption 5. *Assume Y_{t+1} is $\mathcal{F}_{t+1}$-measurable. Assume that for all $t \geqslant 0$,*

$$\mathbb{E}_t \left[Y_{t+1} \mid \theta^\star \right] = \langle \theta^\star, X_t \rangle$$

and there exists some random variable $\sigma_t : \Omega \to \mathbb{R}$,

$$\mathbb{V}_t \left(Y_{t+1} \mid \theta^\star \right) = \sigma_t^2 \quad a.s.$$

One possible way to generate the filtration satisfying the above assumptions is as following,

$$\mathcal{F}_t = \sigma \left(X_0, Y_1, X_1, \ldots, X_{t-1}, Y_t, X_t \right).$$

> **Lemma 11** (Variance reduction). *Define $\mathbf{\Gamma}_t := \mathbb{V}_t \left(\theta^* \right)$ as the posterior covariance of θ^*. When Assumption 4 and Assumption 5 hold, for all $t \geqslant 0$*
>
> $$\mathbb{E} \left[\mathbf{\Gamma}_{t+1} \mid \mathcal{F}_t \right] \preceq \mathbf{\Gamma}_t - \frac{\mathbf{\Gamma}_t^\top X_t X_t^\top \mathbf{\Gamma}_t}{\mathbb{E} \left[\sigma_t^2 \mid \mathcal{F}_t \right] + X_t^\top \mathbf{\Gamma}_t X_t} \quad a.s.$$

Remark 3. *Note that (Hamidi and Bayati, 2021) give a similar lemma as Lemma 11. But they consider a fixed constant upper bound for noise variance instead of the predictable heterogeneous noise variance $\mathbb{E} \left[\sigma_t^2 \mid \mathcal{F}_t \right]$, which is essential for the purpose of this work.*

Proof of Lemma 11. Denote the short notation for conditional expectation $\mathbb{E}_t \left[\cdot \right] = \mathbb{E} \left[\cdot \mid \mathcal{F}_t \right]$ and conditional variance $\mathbb{V}_t \left(\cdot \right) = \mathbb{V} \left(\cdot \mid \mathcal{F}_t \right)$. To make things clear, we first prove the claim for $\theta^\star$ assuming $\mathbb{E}_t \left[\theta^\star \right] = 0$. It suffices to prove that for any d-dim real vector $v \in \mathbb{R}^d$,

$$v^\top \mathbb{E}_t \left[\mathbf{\Gamma}_{t+1} \right] v = \mathbb{E}_t \left[\mathrm{Var}_{t+1} \left(\langle \theta^\star, v \rangle \right) \right]$$

$$\leqslant v^\top \mathbf{\Gamma}_t v - \frac{\left(X_t^\top \mathbf{\Gamma}_t v \right)^2}{\mathbb{E}_t \left[\sigma_t^2 \right] + X_t^\top \mathbf{\Gamma}_t X_t}$$

Denoting by $\mathcal{F}_t^A$ the set of $\mathcal{F}_t$-measurable random variables. For any fixed $v \in \mathbb{R}^d$,

$$\mathbb{E}_t \left[\mathrm{Var}_{t+1} \left(\langle \theta^\star, v \rangle \right) \right] = \mathbb{E}_t \left[\inf_{\xi \in \mathcal{F}_{t+1}^A} \mathbb{E}_{t+1} \left[\left(\langle \theta^\star, v \rangle - \xi \right)^2 \right] \right]$$

$$\leqslant \mathbb{E}_t \left[\inf_{a \in \mathbb{R}} \mathbb{E}_{t+1} \left[\left(\langle \theta^\star, v \rangle - a Y_{t+1} \right)^2 \right] \right]$$

$$\leqslant \inf_{a \in \mathbb{R}} \mathbb{E}_t \left[\mathbb{E}_{t+1} \left[\left(\langle \theta^\star, v \rangle - a Y_{t+1} \right)^2 \right] \right]$$

$$= \inf_{a \in \mathbb{R}} \mathbb{E}_t \left[\left(\langle \theta^\star, v \rangle - a Y_{t+1} \right)^2 \right]$$

By completing the squares, we have the following important equality

$$\inf_{a \in \mathbb{R}} \left(\mathbb{E}_t \left[\left(\langle \theta^\star, v \rangle - a Y_{t+1} \right)^2 \right] \right) = \underbrace{\mathbb{E}_t \left[\langle \theta^\star, v \rangle^2 \right]}_{I} - \underbrace{\frac{\overbrace{\mathbb{E}_t \left[\langle \theta^\star, v \rangle Y_{t+1} \right]^2}^{II}}{\underbrace{\mathbb{E}_t \left[Y_{t+1}^2 \right]}_{III}}}$$

For the first term I, we have

$$\mathbb{E}_t\left[\langle\theta^\star, v\rangle^2\right] = \mathbb{E}_t\left[v^\top\theta^\star\theta^{\star\top}v\right] = v^\top\mathbf{\Gamma}_t v$$

For the second expectation in term II, the numerator can also be computed as

$$
\begin{aligned}
\mathbb{E}_t\left[\langle\theta^\star, v\rangle Y_{t+1}\right] &= \mathbb{E}_t\left[\mathbb{E}_t\left[\langle\theta^\star, v\rangle Y_{t+1} \mid \theta^\star\right]\right] \\
&= \mathbb{E}_t\left[\langle\theta^\star, v\rangle \cdot \mathbb{E}_t\left[Y_{t+1} \mid \theta^\star\right]\right] \\
&= \mathbb{E}_t\left[\langle\theta^\star, v\rangle \langle\theta^\star, X_t\rangle\right] \\
&= \mathbb{E}_t\left[X_t^\top\theta^\star\theta^{\star\top}v\right] \\
&\overset{(a)}{=} X_t^\top\mathbb{E}_t\left[\theta^\star\theta^{\star\top}\right]v \\
&= X_t^\top\mathbf{\Gamma}_t v,
\end{aligned}
$$

where the equality (a) is by Assumption 4 that X_t is $\mathcal{F}_t$-measurable. Finally, for the denominator III,

$$
\begin{aligned}
\mathbb{E}_t\left[Y_{t+1}^2\right] &= \mathbb{E}_t\left[\mathrm{Var}_t\left(Y_{t+1} \mid \theta^\star\right) + \mathbb{E}_t\left[Y_{t+1} \mid \theta^\star\right]^2\right] \\
&= \mathbb{E}_t\left[\sigma_t^2\right] + \mathbb{E}_t\left[\langle\theta^\star, X_t\rangle^2\right] \\
&= \mathbb{E}_t\left[\sigma_t^2\right] + X_t^\top\mathbf{\Gamma}_t X_t
\end{aligned}
$$

For the case that $\mathbb{E}_t\left[\theta^\star\right] \neq 0$, we could recenter θ^* by its posterior mean. Define $\mu^\star := \theta^\star - \mathbb{E}_t\left[\theta^\star\right]$ and $Z_{t+1} := Y_{t+1} - \mathbb{E}_t\left[Y_{t+1}\right] = Y_{t+1} - \langle\mathbb{E}_t\left[\theta^*\right], X_t\rangle$.

Note that $\mathrm{Var}_t\left(\mu^\star\right) = \mathrm{Var}_t\left(\theta^\star\right) = \mathbf{\Gamma}_t$ and

$$
\begin{aligned}
\mathbb{E}_t\left[Z_{t+1} \mid \mu^\star\right] &= \mathbb{E}_t\left[Z_{t+1} \mid \theta^\star\right] = \langle\theta^*, X_t\rangle - \langle\mathbb{E}_t\left[\theta^*\right], X_t\rangle = \langle\mu^\star, X_t\rangle, \\
\mathrm{Var}_t\left(Z_{t+1} \mid \mu^\star\right) &= \mathrm{Var}_t\left(Z_{t+1} \mid \theta^\star\right) = \mathrm{Var}_t\left(Y_{t+1} - \mathbb{E}_t\left[Y_t\right] \mid \theta^\star\right) = \mathrm{Var}_t\left(Y_{t+1} \mid \theta^\star\right).
\end{aligned}
$$

We can apply the result (for the case $\mathbb{E}_t\left[\theta^\star\right] = 0$) by replacing θ^* with μ^* and replacing Y_{t+1} with Z_{t+1} and then obtain directly

$$\mathbb{E}_t\left[\mathbf{\Gamma}_{t+1}\right] = \mathbb{E}_t\left[\mathrm{Var}_{t+1}\left(\mu^\star\right)\right] \preceq \mathbf{\Gamma}_t - \frac{\mathbf{\Gamma}_t^\top X_t X_t^\top\mathbf{\Gamma}_t}{\mathbb{E}_t\left[\sigma_t^2\right] + X_t^\top\mathbf{\Gamma}_t X_t}$$

$\square$

Remark 4. *Let $\bar{\sigma}_t$ be any random variable s.t. $\bar{\sigma}_t^2 \geqslant \mathbb{E}\left[\sigma_t^2 \mid \mathcal{F}_t\right]$ almost surely. By sherman-morrison formula,*

$$\mathbb{E}\left[\mathbf{\Gamma}_{t+1} \mid \mathcal{F}_t\right]^{-1} \succeq \mathbf{\Gamma}_t^{-1} + \frac{1}{\bar{\sigma}_t^2}X_t X_t^\top.$$

Remark 5. *Recall that in the Gaussian prior and Gaussian noise case, $\mathbf{\Gamma}_{t+1}^{-1} = \mathbf{\Gamma}_t^{-1} + \frac{1}{\sigma_t^2}X_t X_t^\top$.*

Remark 6. *Lemma 11 demonstrates that the posterior covariance reduces in expectation. However, in general, the reduction does not necessarily hold almost surely.*

F.2 Proof of Lemma 7

Proof. Let $\{e_i\}_{i=1}^d$ be a set of eigenvectors of $\mathbb{E}\left[f_1\left(X_1\right)f_1\left(X_1\right)^\top\right]$ and an orthonormal basis for $\mathbb{R}^d$, i.e. $\boldsymbol{I}_{d\times d} = \sum_{i=1}^d e_i e_i^\top$. Then we have,

$$
\begin{aligned}
\mathbb{E}\left[\left(f_1\left(X_1\right)^\top f_2\left(X_2\right)\right)^2\right] &= \mathbb{E}\left[\left(\sum_{n=1}^d\left(f_1\left(X_1\right)^\top e_n\right)\left(f_2\left(X_2\right)^\top e_n\right)\right)^2\right] \\
&= \sum_{m,n}\mathbb{E}\left[\left(f_1\left(X_1\right)^\top e_m\right)\left(f_2\left(X_2\right)^\top e_m\right)\left(f_1\left(X_1\right)^\top e_n\right)\left(f_2\left(X_2\right)^\top e_n\right)\right] \\
&= \sum_{m,n}\mathbb{E}\left[\left(f_1\left(X_1\right)^\top e_m\right)\left(f_1\left(X_1\right)^\top e_n\right)\right]\mathbb{E}\left[\left(f_2\left(X_2\right)^\top e_m\right)\left(f_2\left(X_2\right)^\top e_n\right)\right] \quad (23)
\end{aligned}
$$

Since $\{e_i\}_{i=1}^d$ are orthogonal eigenvectors of $\mathbb{E}\left[f_1\left(X_1\right)f_1\left(X_1\right)^\top\right]$, we have

$$\mathbb{E}\left[\left(f_1\left(X_1\right)^\top e_m\right)\left(f_1\left(X_1\right)^\top e_n\right)\right] = e_m^\top \mathbb{E}\left[f_1\left(X_1\right)f_1\left(X_1\right)^\top\right]e_n = 0$$

for any $m \neq n$. Then we can simplify (23) to

$$\mathbb{E}\left[\left(f_1\left(X_1\right)^\top f_2\left(X_2\right)\right)^2\right] = \sum_{n=1}^d \mathbb{E}\left[\left(f_1\left(X_1\right)^\top e_n\right)^2\right]\mathbb{E}\left[\left(f_2\left(X_1\right)^\top e_n\right)^2\right]$$

Consequently, we have

$$
\begin{aligned}
\left(\frac{1}{d}\mathbb{E}\left[\left\|f_1\left(X_1\right)^\top f_2\left(X_1\right)\right\|\right]\right)^2 &= \left(\frac{1}{d}\mathbb{E}\left[\left\|\sum_{n=1}^d \left(f_1\left(X_1\right)^\top e_n\right)\left(f_2\left(X_1\right)^\top e_n\right)\right\|\right]\right)^2 \\
&\leqslant \left(\frac{1}{d}\sum_{n=1}^d \mathbb{E}\left[\left\|\left(f_1\left(X_1\right)^\top e_n\right)\left(f_2\left(X_1\right)^\top e_n\right)\right\|\right]\right)^2 \\
&\overset{(i)}{\leqslant} \frac{1}{d}\sum_{n=1}^d \mathbb{E}\left[\left\|\left(f_1\left(X_1\right)^\top e_n\right)\left(f_2\left(X_1\right)^\top e_n\right)\right\|\right]^2 \\
&= \frac{1}{d}\sum_{n=1}^d \mathbb{E}\left[\left\|f_1\left(X_1\right)^\top e_n\right\|\left\|f_2\left(X_1\right)^\top e_n\right\|\right]^2 \\
&\overset{(ii)}{\leqslant} \frac{1}{d}\sum_{n=1}^d \mathbb{E}\left[\left(f_1\left(X_1\right)^\top e_n\right)^2\right]\mathbb{E}\left[\left(f_2\left(X_1\right)^\top e_n\right)^2\right] \\
&= \frac{1}{d}\mathbb{E}\left[\left(f_1\left(X_1\right)^\top f_2\left(X_2\right)\right)^2\right]
\end{aligned}
$$

The above inequality (i) is by Jensen's inequality and above inequality (ii) is by the Cauchy-Bunyakovsky-Schwarz inequality. The main lemma is now proved.

Observe that the RHS can be further written as,

$$\mathbb{E}\left[\left(f_1\left(X_1\right)^\top f_2\left(X_2\right)\right)^2\right] = \mathrm{Tr}\left(\mathbb{E}\left[f_1(X_1)f_1(X_1)^\top\right]\mathbb{E}\left[f_2(X_2)f_2(X_2)^\top\right]\right),$$

Specifically, let $X_1 = (X,Z)$, $f_1(x,z) = x$, $f_2(x,z) = z$, and X_2 be a i.i.d copy of X_1, we have

$$\mathbb{E}\left[\left|X^\top Z\right|\right]^2 \leqslant d\,\mathrm{Tr}\left(\mathbb{E}\left[XX^\top\right]\mathbb{E}\left[ZZ^\top\right]\right).$$

$\square$

Remark 7. *The idea of using orthonomal basis in the proof is adapted from (Kalkanlı and Özgür, 2020). Our statement is stronger than previous literature since $\mathbb{E}\left[\|f_1(X_1)f_2(X_1)\|\right] \geqslant \mathbb{E}\left[f_1(X_1)f_2(X_1)\right]$ and we provide the upper bound on $\mathbb{E}\left[\|f_1(X_1)f_2(X_1)\|\right].$*

F.3 Variance relationships

Lemma 12 (Cumulative variance of value under true model $V_{M^*}^{\pi^\ell}$). *For any policy π, let the sequence $x_h = (h, s_h, a_h)$ from $h = 0$ to $h = H-1$ be generated by the interaction of policy π and MDP M. Define the observation history $O_h = (x_0, \ldots, x_h)$, we have*

$$\mathbb{V}\left(\sum_{h=0}^{H-1} R(x_h)\,\middle|\, M, \pi, O_0\right) = \mathbb{E}\left[\sum_{h=0}^{H-1} \mathbb{V}\left(V_{\pi,h+1}^M(s_{h+1})\mid M,\pi,O_h\right)\,\middle|\, M,\pi,O_0\right]$$

Proof. By the law of total variance, we can decompose $\mathbb{V}\left(\sum_{j=h}^{H-1} R(x_j) \mid M, \pi, O_h\right)$ as

$$\underbrace{\mathbb{E}\left[\mathbb{V}\left(\sum_{j=h}^{H-1} R(x_j) \mid M, \pi, O_{h+1}\right) \Big| M, \pi, O_h\right]}_{(a)} + \underbrace{\mathbb{V}\left(\mathbb{E}\left[\sum_{j=h}^{H-1} R(x_j) \mid M, \pi, O_{h+1}\right] \Big| M, \pi, O_h\right)}_{(b)}$$

Since x_h is deterministic conditioned on O_{h+1}, we rewrite (a) as

$$(a) = \mathbb{E}\left[\mathbb{V}\left(\sum_{j=h+1}^{H-1} R(x_j) \mid M, \pi, O_{h+1}\right) \Big| M, \pi, O_h\right]$$

Also, we can further rewrite (b) by the definition of $V_{\pi_\ell, h+1}^M$

$$(b) = \mathbb{V}\left(R(x_{\ell,h}) + \mathbb{E}\left[\sum_{j=h+1}^{H-1} R(x_j) \mid M, \pi, O_{h+1}\right] \Big| M, \pi, O_h\right)$$

$$= \mathbb{V}\left(\mathbb{E}\left[\sum_{j=h+1}^{H-1} R(h, s_h, a_h) \mid M, \pi, O_{h+1}\right] \Big| M, \pi, O_h\right)$$

$$= \mathbb{V}\left(\mathbb{E}\left[\sum_{j=h+1}^{H-1} R(h, s_h, \pi(s_h)) \mid M, \pi, O_{h+1}\right] \Big| M, \pi, O_h\right)$$

$$= \mathbb{V}\left(V_{\pi,h+1}^M(s_{h+1}) \mid M, \pi, O_h\right)$$

By recursive application of the above result, we obtain

$$\mathbb{V}\left(\sum_{j=0}^{H-1} R(x_j) \mid M, \pi, O_0\right) = \mathbb{E}\left[\sum_{h=0}^{H-1} \mathbb{V}\left(V_{\pi,h+1}^M(s_{h+1}) \mid M, \pi, O_h\right) \Big| M, \pi, O_0\right]$$

$\square$

Corollary 3. *Recall the definition of $\mathcal{H}_{\ell,h}$, we can apply Lemma 12 with setting $\pi = \pi_\ell$ and get,*

$$\mathbb{V}_{\ell,0}\left(\sum_{j=0}^{H-1} R(x_{\ell,j}) \mid M\right) = \mathbb{E}_{\ell,0}\left[\sum_{h=0}^{H-1} \mathbb{V}_{\ell,h}\left(V_{\pi_\ell,h+1}^M(s_{\ell,h+1}) \mid M\right) \Big| M\right]$$

Beside, we could derive the following upper bound for summation of variance

$$\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=0}^{H-1} \mathbb{V}_{\ell,h}\left(V_{\pi_\ell,h+1}^M(s_{\ell,h+1}) \mid M\right)\right] = \sum_{\ell=1}^{L} \mathbb{E}\left[\mathbb{V}_{\ell,0}\left(\sum_{j=0}^{H-1} R(x_{\ell,j}) \mid M\right)\right] \leqslant LH^2.$$

Remark 8. *The proof of Lemma 12 is by the sequential conditional version of the law of total variance, which is similar to the proof in the frequentist literature (Munos and Moore, 1999; Lattimore and Hutter, 2012; Azar et al., 2013).*

Lemma 13 (Variance difference). *For each stage h of each episode ℓ,*

$$\mathbb{V}_{\ell,h}\left(V_{\pi_\ell,h+1}^{\hat{M}_\ell}(s_{\ell,h+1}) \mid M\right) - \mathbb{V}_{\ell,h}\left(V_{\pi_\ell,h+1}^M(s_{\ell,h+1}) \mid M\right) \leqslant 2H\mathbb{E}_{\ell,h}\left[|\Delta_{\ell,h+1}(s_{\ell,h+1})| \mid M\right]$$

Proof. First, we rewrite the difference of the variance,

$$
\mathbb{V}_{\ell,h}\left(V^{\hat{M}_\ell}_{\pi_\ell,h+1}(s_{\ell,h+1})\mid M\right) - \mathbb{V}_{\ell,h}\left(V^{M}_{\pi_\ell,h+1}(s_{\ell,h+1})\mid M\right)
$$

$$
= \underbrace{\mathbb{E}_{\ell,h}\left[\left(V^{\hat{M}_\ell}_{\pi_\ell,h+1}(s_{\ell,h+1})\right)^2\mid M\right] - \mathbb{E}_{\ell,h}\left[\left(V^{M}_{\pi_\ell,h+1}(s_{\ell,h+1})\right)^2\mid M\right]}_{(a)}
$$

$$
\underbrace{- \mathbb{E}_{\ell,h}\left[V^{\hat{M}_\ell}_{\pi_\ell,h+1}(s_{\ell,h+1})\mid M\right]^2 + \mathbb{E}_{\ell,h}\left[V^{M}_{\pi_\ell,h+1}(s_{\ell,h+1})\mid M\right]^2}_{(b)}.
$$

We further bound (a) by,

$$
(a) = \mathbb{E}_{\ell,h}\left[\left(\left[V^{\hat{M}_\ell}_{\pi_\ell,h+1} - V^{M}_{\pi_\ell,h+1}\right](s_{\ell,h+1})\right)\left(\left[V^{\hat{M}_\ell}_{\pi_\ell,h+1} + V^{M}_{\pi_\ell,h+1}\right](s_{\ell,h+1})\right)\mid M\right]
$$

$$
\leqslant 2H\mathbb{E}_{\ell,h}\left[\left[V^{\hat{M}_\ell}_{\pi_\ell,h+1} - V^{M}_{\pi_\ell,h+1}\right]_+(s_{\ell,h+1})\mid M\right].
$$

Similarly, bound (b) by,

$$
(b) = \mathbb{E}_{\ell,h}\left[\left[V^{\hat{M}_\ell}_{\pi_\ell,h+1} - V^{M}_{\pi_\ell,h+1}\right](s_{\ell,h+1})\mid M\right]\mathbb{E}_{\ell,h}\left[\left[V^{\hat{M}_\ell}_{\pi_\ell,h+1} + V^{M}_{\pi_\ell,h+1}\right](s_{\ell,h+1})\mid M\right]
$$

$$
\leqslant 2H\mathbb{E}_{\ell,h}\left[\left[V^{\hat{M}_\ell}_{\pi_\ell,h+1} - V^{M}_{\pi_\ell,h+1}\right](s_{\ell,h+1})\mid M\right]_+
$$

$$
\leqslant 2H\mathbb{E}_{\ell,h}\left[\left[V^{\hat{M}_\ell}_{\pi_\ell,h+1} - V^{M}_{\pi_\ell,h+1}\right]_+(s_{\ell,h+1})\mid M\right].
$$

Putting all together,

$$
(a)+(b) = 2H\mathbb{E}_{\ell,h}\left[\left[V^{\hat{M}_\ell}_{\pi_\ell,h+1} - V^{M}_{\pi_\ell,h+1}\right]_+(s_{\ell,h+1}) + \left[V^{M}_{\pi_\ell,h+1} - V^{\hat{M}_\ell}_{\pi_\ell,h+1}\right]_+(s_{\ell,h+1})\mid M\right]
$$

$$
= 2H\mathbb{E}_{\ell,h}\left[\left|V^{\hat{M}_\ell}_{\pi_\ell,h+1}(s_{\ell,h+1}) - V^{M}_{\pi_\ell,h+1}(s_{\ell,h+1})\right|\mid M\right]
$$

$$
= 2H\mathbb{E}_{\ell,h}\left[|\Delta_{\ell,h+1}(s_{\ell,h+1})|\mid M\right].
$$

$\square$

F.4 Proof of Lemma 10

For completeness, we give a proof of potential lemma which is adpated from (Hamidi and Bayati, 2021).

Proof. First, assume that $\boldsymbol{\Sigma}$ is invertible. In this case, we have $\boldsymbol{\Sigma}'^{-1} = \boldsymbol{\Sigma}^{-1} + VV^\top$, using Sherman-Morrison formula. Due to the fact $VV^\top \preceq V^\top V\boldsymbol{I}$,

$$
f\left(\boldsymbol{\Sigma}, x + V^\top V\right) = \log\det\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\boldsymbol{\Sigma}^{-1} + x\boldsymbol{I} + V^\top V\boldsymbol{I}\right)\boldsymbol{\Sigma}^{\frac{1}{2}}\right)
$$

$$
\geqslant \log\det\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\boldsymbol{\Sigma}^{-1} + x\boldsymbol{I} + VV^\top\right)\boldsymbol{\Sigma}^{\frac{1}{2}}\right)
$$

$$
= \log\det\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\boldsymbol{\Sigma}'^{-1} + x\boldsymbol{I}\right)\boldsymbol{\Sigma}^{\frac{1}{2}}\right)
$$

We can further rewrite,

$$
\begin{aligned}
\log \det \left(\boldsymbol{\Sigma}^{\frac{1}{2}} \left(\boldsymbol{\Sigma}'^{-1} + x\boldsymbol{I} \right) \boldsymbol{\Sigma}^{\frac{1}{2}} \right) &= \log \det(\boldsymbol{\Sigma}) + \log \det \left(\boldsymbol{\Sigma}'^{-1} + x\boldsymbol{I} \right) \\
&= \log \det \left(\boldsymbol{\Sigma}' \right) - \log \det \left(\boldsymbol{I} - \frac{\boldsymbol{\Sigma}^{\frac{1}{2}} V V^{\top} \boldsymbol{\Sigma}^{\frac{1}{2}}}{1 + V^{\top} \boldsymbol{\Sigma} V} \right) + \log \det \left(\boldsymbol{\Sigma}'^{-1} + x\boldsymbol{I} \right) \\
&\overset{(c)}{=} \log \det \left(\boldsymbol{\Sigma}' \right) - \log \left(1 - \frac{V^{\top} \boldsymbol{\Sigma} V}{1 + V^{\top} \boldsymbol{\Sigma} V} \right) + \log \det \left(\boldsymbol{\Sigma}'^{-1} + x\boldsymbol{I} \right) \\
&= \log \left(1 + V^{\top} \boldsymbol{\Sigma} V \right) + \underbrace{\log \det \left(\boldsymbol{\Sigma}'^{\frac{1}{2}} \left(\boldsymbol{\Sigma}'^{-1} + x\boldsymbol{I} \right) \boldsymbol{\Sigma}'^{\frac{1}{2}} \right)}_{f(\boldsymbol{\Sigma}', x)}
\end{aligned}
$$

where (c) is from $\det \left(\boldsymbol{I} + Z Z^{\top} \right) = 1 + Z^{\top} Z$.

We remain to prove the argument for the case of non-invertible matrix $\boldsymbol{\Sigma}$. In this case, for $\epsilon > 0$, we define $\boldsymbol{\Sigma}_{\epsilon} = \boldsymbol{\Sigma} + \epsilon \boldsymbol{I}$ and

$$
\boldsymbol{\Sigma}'_{\epsilon} := \boldsymbol{\Sigma}_{\epsilon} - \frac{\boldsymbol{\Sigma}_{\epsilon} V V^{\top} \boldsymbol{\Sigma}_{\epsilon}}{1 + V^{\top} \boldsymbol{\Sigma}_{\epsilon} V}.
$$

Clearly, $\boldsymbol{\Sigma}_{\epsilon}$ is invertible. Therefore, we can apply the previous results to $\boldsymbol{\Sigma}_{\epsilon}$ to obtain

$$
\log \left(1 + V^{\top} \boldsymbol{\Sigma}_{\epsilon} V \right) + f \left(\boldsymbol{\Sigma}'_{\epsilon}, x \right) \leqslant f \left(\boldsymbol{\Sigma}_{\epsilon}, x + V^{\top} V \right)
$$

The claim then follows the continuity of the above expressions with respect to ϵ on $[0, \infty]$ □

G   Lower bound conjecture on the dependence of interactions

For the bandit and RL analysis that relies on confidence bounds, one $O(\sqrt{\log T})$ term arises from a union bound over all T interaction periods. All previous analysis for reinforcement learning[6] relies on various types of confidence bounds and thus suffers this additional terms.

A recent improved lower bound $\Omega(d\sqrt{T \log T})$ in linear bandit (Li et al., 2019) shows the fundamental difference between problems with the finite action spaces and infinite action spaces that is changing over times. From their evidence, we conjecture that the lower bound for the class of linear mixture MDPs with changing and infinite action spaces should be $\Omega(dH\sqrt{T \log T})$. Since being a model based algorithm, PSRL can naturally handle the infinite and changing action sets scenario. Therefore, our analysis for PSRL would match the lower bound up to constants if the conjecture holds.

Conjecture 1. *Suppose the number of episodes $L \geqslant \text{poly}(d, H, B)$ for some large d and H. There exists a prior distribution over Θ such that the Assumption 2 and the Assumption 3 holds with $B > 1$, such that, for any algorithm, the expected regret is lower bounded as follows:*

$$
\mathbb{E}_{\Theta} \Re \left(M_{\Theta}, L \right) \geqslant \Omega(dH\sqrt{T \log T})
$$

where $T = LH$ and $\mathbb{E}_{\Theta}$ denotes the expectation over the prior distribution over Θ and the probability distribution generated by the interconnection of the algorithm and the MDP instance.

---

[6]A concurrent work (Zhang, 2021) established the decoupling coefficient approach for analyzing randomized exploration algorithms for contextual bandits and a specific class of reinforcement learning problems with deterministic transitions.