Proceedings of Machine Learning Research vol 178:1–39, 2022 35th Annual Conference on Learning Theory

Non-Linear Reinforcement Learning in Large Action Spaces: Structural Conditions and Sample-efficiency of Posterior Sampling

Alekh Agarwal ALEKHAGARWAL@GOOGLE.COM
Google Research

Tong Zhang TOZHANG@GOOGLE.COM
Google Research and HKUST

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Provably sample-efficient Reinforcement Learning (RL) with rich observations and function approximation has witnessed tremendous recent progress, particularly when the underlying function approximators are linear. In this linear regime, computationally and statistically efficient methods exist where the potentially infinite state and action spaces can be captured through a known feature embedding, with the sample complexity scaling with the (intrinsic) dimension of these features. When the action space is finite, significantly more sophisticated results allow non-linear function approximation under appropriate structural constraints on the underlying RL problem, permitting for instance, the learning of good features instead of assuming access to them. In this work, we present the first result for non-linear function approximation which holds for general action spaces under a *linear embeddability* condition, which generalizes all linear and finite action settings. We design a novel optimistic posterior sampling strategy, TS^3 for such problems. We further show worst case sample complexity guarantees that scale with a rank parameter of the RL problem, the linear embedding dimension introduced here and standard measures of function class complexity.

Keywords: Reinforcement Learning, Exploration, Low-rank MDPs, Posterior sampling

1. Introduction

Designing sample-efficient techniques for Reinforcement Learning (RL) settings with large state and action spaces is a key question at the forefront of RL research. Typical approaches for these scenarios rely on the use of function approximation to generalize across the state and/or action spaces. When a sufficiently expressive feature embedding is available to the learner, and linear functions of these features are used for learning, a number of recent results provide computationally and statistically efficient techniques to handle continuous state, as well as action spaces with dependence only on an intrinsic dimensionality of the features (Yang and Wang, 2020; Jin et al., 2020; Zanette et al., 2020b; Agarwal et al., 2020a). However, practitioners typically rely on neural networks to parameterize the learning agent, a case not covered by the linear results. A parallel line of work (Jiang et al., 2017; Sun et al., 2019; Jin et al., 2021; Du et al., 2021) studies more general settings that allow the use of non-linear function approximation, and provides sample complexity guarantees in terms of a structural parameter called the *Bellman rank* of the RL problem. The power of these conditions is elucidated in recent representation learning results (Agarwal et al., 2020b; Uehara et al., 2021; Modi et al., 2021), that leverage low Bellman rank to learn a good feature map that captures a near-optimal policy and/or the transition dynamics. However, these methods crucially rely on having a finite number of actions, and the guarantees scale with the cardinality of the action set. With this context, our paper asks the following question:

	Structural complexity for TS^3	Other approaches
V-type Bellman rank d_1 + K actions	$d_1^2 K$	OLIVE (Jiang et al., 2017), V-type GOLF (Jin et al., 2021)
Linear MDP (effective feature dim d)	d^3	LSVI-UCB (Jin et al., 2020), Q-type GOLF (Jin et al., 2021)
Mixture of MDPs w/ Bellman rank d_1 + K actions	$d_1^2 K$	Contextual PC-IGW (Foster et al., 2021)
Rank d dynamics + ranking L out of K items (Example 1)	$d^2 KL$	??

Table 1: Settings captured by our assumptions and prior approaches for them. Our setting subsumes finite action MDPs with a small V-type Bellman rank and infinite action linear MDPs. All problems with a small Q-type Bellman rank are not covered (see Appendix A). Prior works use different methods for V-type Bellman rank with finite actions and linear MDP with infinite actions, unlike this work.

Can we design sample-efficient and non-linear RL approaches for large state and action spaces?

In this paper, we study this question in the framework of a Markov Decision Process (MDP), and focus on problems that admit a good bound on a generalization of the Bellman rank parameter introduced by Jiang et al. (2017). While the definition of Bellman rank applies to both discrete and continuous action spaces, the OLIVE algorithm of Jiang et al. (2017) applies only to discrete action spaces. The presence of a small action set facilitates uniform exploration for one time step, which lets the agent collect valuable exploration data in the vicinity of states it has already explored, allowing the discovery of successively better states. When good features are available, like in a linear MDP, a basis in the feature space serves an analogous role and indeed recent works (Foster et al., 2021) show that experimental design in the right feature space can replace uniform exploration over discrete actions.

However, this strategy fails beyond (generalized) linear settings, where there is no easy mechanism for obtaining a good exploration basis over the action set apriori. Indeed, the results of Hao et al. (2021) imply that some dependence on the size of the action space in general is unavoidable for a polynomial sample complexity in all relevant parameters (see Appendix C for further discussion). This situation motivates the investigation of additional structures between the hopeless worst-case result and the limiting small action settings. To this end, our approach is motivated by the recent work of Zhang (2021) on the Feel-Good Thompson Sampling (FGTS) algorithm, which they analyze for bandits and a class of RL problems under a linear embeddability assumption using a modified Thompson Sampling approach.

Our Contributions. With this context, our work makes the following contributions.

1. We introduce a new structural model for RL where a generalized form of Bellman rank is small, and a further linear embeddability assumption on the Bellman error is satisfied. We show that this setting subsumes prior works on both linear MDPs as well as finite action problems with a small Bellman rank. Crucially, in both linear MDPs and finite action problems, the embedding features in our definition are known apriori, while we also handle problems with an *unknown linear embedding*, which constitutes a significant generalization of the prior works. As an example, this allows us to generalize the combinatorial bandits setting to long horizon RL (Table 1).

2. We introduce a new algorithm, Two timeScale Two Sample Thompson Sampling (TS^3), which is motivated from FGTS (Zhang, 2021). The algorithm design incorporates a careful two timescale strategy (that is, a fast learning rate for the $\max$ operation and a slow learning rate for the $\min$ operation) to solve an online minimax problem for estimating Bellman residuals, and a decoupling between roll-in and roll-out policies crucial for the online minimization of these residuals. Both ideas appear novel relative to prior posterior sampling approaches in the literature. The use of nested posterior sampling to solve online minimax problems might be of independent interest.
3. We show that TS^3 solves all problems where the generalized Bellman factorization and linear embeddability conditions hold, under additional completeness and realizability assumptions. The guarantees scale polynomially in $1/\epsilon$, horizon H, Bellman rank and linear embedding parameters, as well as the function class complexity. The result generalizes both guarantees for linear MDPs in terms of feature dimension with no dependence on action cardinality, as well as Bellman rank-based and representation learning guarantees to linearly embeddable infinite action spaces. Overall, our method provides the first approach to representation learning with continuous action spaces. We summarize the settings covered in this paper in Table 1.

We remark that the convergence rate guarantees obtained here likely have room for improvement, given that they do not match the prior results for discrete actions. We leave this as an important direction for future work, and comment on the potential sources of looseness in our analysis in the later sections of the paper. We now formalize the setting before discussing our structural assumptions and our approach. Detailed discussion of the related works can be found in Appendix A.

2. Setting

We study RL in an episodic, finite horizon Contextual Markov Decision Process (MDP) that is parameterized as $(\mathcal{X}, \mathcal{A}, \mathcal{R}, \mathcal{D}, P)$, where $\mathcal{X}$ is a state space, $\mathcal{A}$ is an action space, $\mathcal{R}$ is a distribution over rewards, $\mathcal{D}$ and P are distributions over the initial context and subsequent transitions respectively. An agent observes a context $x^1 \sim \mathcal{D}$ for some fixed distribution $\mathcal{D}$.[1] At each time step $h \in \{1, \ldots, H\}$, the agent observes the state x^h, chooses an action a^h, observes $r^h \sim \mathcal{R}(\cdot \mid x_h, a_h)$ and transitions to $x^{h+1} \sim P(\cdot \mid x^h, a^h)$. We assume that x^h for any $h > 1$ always includes the context x^1 to allow arbitrary dependence of the dynamics and rewards on x^1. Following prior work (e.g. Jiang et al., 2017), we assume that $r^h \in [0,1]$ and $\sum_{h=1}^H r^h \in [0,1]$ to capture sparse-reward settings (Jiang and Agarwal, 2018). We make no assumption on the cardinality of the state and/or action spaces, allowing both to be potentially infinite. We use π to denote the agent's decision policy, which maps from $\mathcal{X} \to \Delta(\mathcal{A})$, where $\Delta(\cdot)$ represents probability distributions over a set. The goal of learning is to discover an optimal policy $\pi_\star$, which is always deterministic and conveniently defined in terms of the $Q_\star$ function (see e.g. Puterman, 2014; Bertsekas and Tsitsiklis, 1996)

$$\pi_\star^h(x^h) = \operatorname*{argmax}_{a \in \mathcal{A}} Q_\star^h(x^h, a), \quad Q_\star^h(x^h, a^h) = \underbrace{\mathbb{E}[r^h + \max_{a' \in \mathcal{A}} Q_\star^{h+1}(x^{h+1}, a') \mid x^h, a^h]}_{\mathcal{T}^h Q_\star^h(x^h, a^h)}, \tag{1}$$

where we define $Q_\star^{H+1}(x, a) = 0$ for all x, a.

In this work, we focus on value-based approaches, where the learner has access to a function class $\mathcal{F} \subseteq \{\mathcal{X} \times \mathcal{A} \to [0,1]\}$, and uses this function class to approximate the optimal value function $Q_\star$. Let us denote $[H] := \{1, 2, \ldots, H\}$. We make two common expressivity assumptions on $\mathcal{F}$.

[1] We intentionally call x^1 a context and not an initial state of the MDP as we will soon make certain structural assumptions which depend on the context, but take expectation over the states.

Assumption 1 (Realizability) *For all $h \in [H]$, we have $Q_\star^h \in \mathcal{F}$.*

There are well-known lower bounds on the sample complexity of RL (Sekhari et al., 2021), even under the structural assumptions we will make when realizability is not approximately satisfied. Let us use $\pi_f(x) = \operatorname{argmax}_{a \in \mathcal{A}} f(x,a)$ and $f(x) = \max_a f(x,a)$ for any $f \in \mathcal{F}$.

Assumption 2 (Completeness) *For any function $f \in \mathcal{F}$ and $h \in [H]$, $\exists\, g \in \mathcal{F}$ such that $g(x^h, a^h) = \mathcal{T}^h f(x^h, a^h) := \mathbb{E}[r^h + f(x^{h+1}, \pi_f(x^{h+1})) \mid x^h, a^h]$.*

The completeness assumption is not essential information-theoretically and OLIVE (Jiang et al., 2017) does not require it. However, we make this assumption to facilitate scaling to infinite action sets, which seems challenging without completeness, as we discuss later. We now present the main structural conditions we impose on the environment, and motivate them with some examples.

3. Structural Conditions

In this section, we present the two main structural conditions, namely generalized Bellman rank and linear embeddability. We define the Bellman residual of $f \in \mathcal{F}$ using another $g \in \mathcal{F}$ as:

$$\mathcal{E}^h(g, f, x^h, a^h) = g(x^h, a^h) - \mathcal{T} f(x^h, a^h). \tag{2}$$

We now state a generalized Bellman error decomposition for contextual setups.

Assumption 3 (Generalized Bellman decomposition) *We assume that for all $f, f' \in \mathcal{F}$, and $h \in [H]$, there exist (unknown) functions u^h, ψ^{h-1} and an inner product $\langle \cdot, \cdot \rangle$ such that for any starting state $x^1 \in \mathcal{X}$:*

$$\mathbb{E}_{x^h \sim \pi_{f'} | x^1} \mathcal{E}^h(f, f; x^h, \pi_f(x^h)) = \langle \psi^{h-1}(f', x^1), u^h(f, x^1) \rangle.$$

We assume that $\sup_{f \in \mathcal{F}, x^1 \in \mathcal{X}} \|u^h(f, x^1)\|_2 \leq B_1$.

Notice that both factors depend on x^1, and a similar definition is recently considered in Foster et al. (2021) to cover contextual RL setups (Abbasi-Yadkori and Neu, 2014; Modi et al., 2018). The technical treatment of contextual dependency requires a conversion of RL problems into online learning, as in Foster et al. (2021) and in this paper. Techniques used by other earlier works on decomposition of average Bellman error cannot handle such a dependency. This definition allows mixtures of MDPs with a small Bellman rank for each context x^1, without any explicit scaling with the number of contexts in our results. We now make an effective dimension assumption on ϕ^h, u^h instead of requiring them to be finite dimensional like Jiang et al. (2017).

Definition 1 (Effective Bellman rank) *Given any probability measure p on $\mathcal{F}$. Let*

$$\Sigma^h(p, x^1) = \mathbb{E}_{f' \sim p} \psi^h(f', x^1) \otimes \psi^h(f', x^1), \quad K^h(\lambda) = \sup_{p, x^1} \operatorname{trace}\left((\Sigma^h(p, x^1) + \lambda I)^{-1} \Sigma^h(p, x^1) \right),$$

where $\otimes$ is the vector outer product. $\forall \epsilon > 0$, define the effective Bellman rank of the MDP as

$$\operatorname{br}(\epsilon) = \sum_{h=1}^{H} \operatorname{br}^{h-1}(\epsilon), \quad \operatorname{br}^{h-1}(\epsilon) = \inf_{\lambda > 0} \left\{ K^{h-1}(\lambda) : \lambda K^{h-1}(\lambda) \leq \epsilon^2 \right\}.$$

When $\dim(\psi^{h-1}) = \dim(u^h) = d$, then $K^h(\lambda) \le d$ for all $\lambda > 0$ so that $\mathrm{br}(0) \le dH$. This might appear a factor of H larger than the Bellman rank, but note that we aggregate over the horizon, while taking a supremum like Jiang et al. (2017) simply incurs that factor in the sample complexity analysis instead. More generally, the following result gives the behavior of $\mathrm{br}(\epsilon)$ under typical spectral decay assumptions on Σ^h.

Proposition 2 (Rank bounds under spectral decay) *Suppose that $\|\psi(f, x^1)\| \le 1$ for all $f \in \mathcal{F}$, $x^1 \in \mathcal{X}$. Let $\lambda_i(A)$ denote the i_{th} largest eigenvalue of a psd matrix A. Suppose we have:*

- Geometric decay: *if* $\exists \alpha \in (0, 1)$*:* $\sup_{h,p,x^1} \lambda_i(\Sigma^h(p, x^1)) \le \alpha^i$, *then* $\mathrm{br}(\epsilon) \le H + 2H \cdot \frac{\log(2/(\epsilon^2(1-\alpha)))}{\log(1/\alpha)}$.
- Polynomial decay: *if* $\exists q \in (0, 1)$, $\sup_{h,p,x^1} \mathrm{trace}((\Sigma^h(p, x^1))^q) \le R_q$, *then* $\mathrm{br}(\epsilon) \le H \cdot \left(\frac{R_q}{\epsilon^2}\right)^{q/1-q}$.

We provide a proof of Proposition 2 in Appendix E. Jin et al. (2021) study a related log-determinant based effective dimension for Bellman rank, but this definition is more natural in our analysis.

While Assumption 3 and Definition 1 facilitate scaling to infinite dimensional factorizations with a low intrinsic dimension, this is still not sufficient to succeed in problems with large action spaces due to the existing lower bounds on learning under sparse transitions (see Hao et al. (2021) and Appendix C). We now introduce the second structural condition, inspired by the recent work of Zhang (2021) on contextual bandits and RL with deterministic dynamics, to handle this issue.

Assumption 4 (Linearly embeddable Bellman error) *For all $f \in \mathcal{F}$, $h \in [H]$, $x^h \in \mathcal{X}$ and $a^h \in \mathcal{A}$, there exist (unknown) functions w^h, ϕ^h and an inner product $\langle \cdot, \cdot \rangle$ such that*

$$\mathcal{E}^h(f, f; x^h, a^h) = \langle w^h(f, x^h), \phi^h(x^h, a^h) \rangle.$$

We assume that $\sup_{f \in \mathcal{F}, x^h \in \mathcal{X}} w^h(f, x^h) \le B_2$.

Note that the function w^h is allowed to depend on x^h, which makes this assumption significantly weaker than embedding Bellman errors in a fixed feature space. For instance, suppose $|\mathcal{A}| = K$. Then we can always define $\phi(x^h, a^h) = e_{a^h} \in \mathbb{R}^K$ to be the indicator of the action a^h and $w^h(f, x^h) = (f(x^h, a_1), \ldots, f(x^h, a_K))$ to satisfy Assumption 4. This shows that all finite action problems satisfy this assumption, meaning that we strictly subsume prior V-type factorizations (Jiang et al., 2017; Du et al., 2019). In linear MDPs, $w^h(f, x^h)$ only depends on f and $\phi^h(x^h, a^h)$ are the features defining the transition dynamics (Jin et al., 2020). Similar to Assumption 3, we allow the embeddings w^h, ϕ^h to have a small effective dimension.

Definition 3 (Effective Embedding Dimension) *Given any probability measure p on $\mathcal{F}$, define*

$$\tilde{\Sigma}^h(p, x) = \mathbb{E}_{f \sim p} \mathbb{E}_{a \sim \pi_f(x)} \phi^h(x, a) \otimes \phi^h(x, a), \quad \tilde{K}^h(\lambda) = \sup_{p,x} \mathrm{trace}\left((\tilde{\Sigma}^h(p, x) + \lambda I)^{-1} \tilde{\Sigma}^h(p, x) \right).$$

For any $\epsilon > 0$, define the effective embedding dimension of as

$$\mathrm{dc}(\epsilon) = \sum_{h=1}^{H} \mathrm{dc}^h(\epsilon), \quad \textit{where} \quad \mathrm{dc}^h(\epsilon) = \inf_{\lambda > 0} \left\{ \tilde{K}^h(\lambda) : \lambda \tilde{K}^h(\lambda) \le \epsilon^2 \right\}.$$

The definition of $\tilde{K}^h(\lambda)$ measures the (approximate) condition number of the covariance for the worst distribution p. If $\dim(\phi^h) = d$, then $\tilde{K}^h(\lambda) \leq d$. In this case, we get $\mathrm{dc}(0) \leq dH$. For infinite dimensional ϕ^h, we obtain bounds similar to Proposition 2.

Comparison with Bellman-Eluder dimension (Jin et al., 2021). Jin et al. (2021) present an alternative Q-type Bellman rank and its distributional generalization based on the Eluder dimension (Russo and Van Roy, 2013). However, both Eluder dimension based results (Wang et al., 2020; Feng et al., 2021) and the Q-type Bellman-Eluder dimension do not capture all finite action, non-linear contextual bandit problems with a realizable reward (see Appendix B). Our setup captures this setting, just like the V-type Bellman rank of Jiang et al. (2017), but does not include all problems with a small Q-type Bellman rank. The most prominent example of linear MDP with infinite actions covered by the GOLF algorithm of Jin et al. (2021) is also captured by our assumptions.

So far, we have explained how both the settings of a bounded Bellman rank with a finite action set, as well as linear MDPs with infinite actions satisfy both of our assumptions with good bounds on $\mathrm{br}(\epsilon)$ and $\mathrm{dc}(\epsilon)$. Next we discuss some examples beyond these where our assumptions hold and which go beyond either of these two known special cases. For the examples, we assume that the underlying MDP has low-rank transition dynamics (Jiang et al., 2017; Jin et al., 2020) so that $\mathcal{T}f(x,a) = \langle w_f, \psi(x,a)\rangle$ for all functions $f : \mathcal{X} \times \mathcal{A} \to [0,1]$ with $\psi(x,a)$ being the state-action features which factorize the dynamics. For such problems, Assumption 3 always holds (Jiang et al., 2017) and Assumption 4 is equivalent to checking that $f(x,a)$ is linearly embeddable.

Example 1 (Linear embedding of combinatorial actions) *Many recommendation settings consist of combinatorial action spaces such as lists, rankings and page layouts. While these problems are intractable in the worst-case, a line of work originating in combinatorial bandits (Cesa-Bianchi and Lugosi, 2012) assumes linear decomposition of rewards for tractability. For concreteness, let us consider a ranking scenario where the system observes some state features x depending on the current user state and any other side information, and wants to choose a ranking $\mathfrak{a}$ of L items $\alpha_1, \ldots, \alpha_L$ from a base set Ω, with $|\Omega| = K$. We observe that $|\mathcal{A}| = \binom{K}{L} \cdot L!$ in this example, which grows as $\mathcal{O}(K^L)$. Hence the sample complexity of direct exploration over rankings is $\mathcal{O}(K^L)$.*

Mirroring the setup from combinatorial bandits and its contextual generalization (Swaminathan et al., 2017), Ie et al. (2019) propose the SlateQ model where the $Q_\star$ function takes the form:

$$Q^\star(x,a) = \sum_{\alpha \in \mathfrak{a}} P(\alpha|x,\mathfrak{a}) g(x,\alpha). \tag{3}$$

Here $P(\alpha|x,\mathfrak{a})$ is the probability of a user (with features x) choosing the item α when the ranking $\mathfrak{a}$ is presented, and $g(x,\alpha)$ is an unknown value of recommending the item α in state x to the user. This assumption is motivated from typical approaches in the user-modeling literature in recommendation systems, and effectively posits that the Q-value of a ranking in a state depends on the unknown *values of the base items in that ranking, weighted by a user interaction model P.*

The user model, which encodes the likelihood of the user choosing a particular item in a ranking is often estimated separately from the RL task using a click probability model, or a cascade model (see e.g. Ie et al., 2019, for a discussion). In such scenarios, we can define a class $\mathcal{F}$ by parameterizing the function g, and obtain linear embedding of Bellman residuals whenever the MDP also has low-rank dynamics in some features ψ, using $\phi(x,\mathfrak{a}) = (P(\cdot \mid x,\mathfrak{a}), \psi(x,\mathfrak{a}))$ as the concatenation of the user model with dynamics features. Here $\psi(x,\mathfrak{a})$ intuitively encodes the information needed to describe how the state x of a user evolves across interactions, which could be low-dimensional, for instance, if there is a set of representative user types.

Algorithm 1 Two timeScale Two Sample Thompson Sampling (TS^3)

```
Require: Function class $\mathcal{F}$, prior $p_0 \in \Delta(\mathcal{F})$, learning rates $\eta, \gamma$ and optimism coefficient $\lambda$.
1: Set $S_0 = \emptyset$.
2: for $t = 1, \ldots, T$ do
3:     Observe $x_t^1 \sim \mathcal{D}$ and draw $h_t \sim \{1, \ldots, H\}$ uniformly at random.
4:     Define $q_t(g|f) = p(g|f, S_{t-1}) \propto p_0(g)\exp(-\gamma\sum_{s=1}^{t-1}\hat{\Delta}_s^{h_s}(g,f)^2)$. ▷ Inner loop TS update
5:     Define $L_t^h(f) = \eta\hat{\Delta}_t^h(f,f)^2 + \frac{\eta}{\gamma}\ln\mathbb{E}_{g\sim q_t(\cdot|f)}\left[\exp(-\gamma\hat{\Delta}_t^h(g,f)^2)\right]$. ▷ Likelihood function
6:     Define $p_t(f) = p(f|S_{t-1}) \propto p_0(f)\exp(\sum_{s=1}^{t-1}(\lambda f(x_s^1) - L_s^{h_s}(f)))$ as the posterior. ▷ Outer loop Optimistic TS update
7:     Draw $f_t, f_t' \sim p_t$ independently from the posterior. Let $\pi_t = \pi_{f_t}$ and $\pi_t' = \pi_{f_t'}$. ▷ Two independent samples $f_t, f_t'$ from posterior
8:     Play iteration $t$ using $\pi_t$ for $h = 1, \ldots, h_t - 1$ and $\pi_t'$ for $h_t$ onwards.
9:     Update $S_t = S_{t-1} \cup \{x_t^h, a_t^h, r_t^h, x_t^{h+1}\}$ for $h = h_t$.
10: end for
11: return $(\pi_1, \ldots, \pi_T)$.
```

Notably, these choices lead to a linear embedding dimension which grows as K. In many applications, the position of an item in the ranking is a dominant effect, in which case we can consider α to consist of an item-position tuple, leading to an KL-dimensional factorization. In either case, this yields an exponential saving in the sample complexity compared with direct exploration in the ranking space, mirroring prior results (Cesa-Bianchi and Lugosi, 2012; Swaminathan et al., 2017; Ie et al., 2019), which either do not consider exploration or work in simpler bandit settings. We are not aware of other existing approaches that can handle rich observation RL and combinatorial action spaces simultaneously.

In Appendix D, we give another example of using a basis expansion in the action space, where Assumption 4 holds in a natural manner. We now proceed to describe our algorithm.

4. The Two timeScale Two Sample Thompson Sampling algorithm

Having presented our structural conditions, we now present our main algorithm in this section, which is based on Thompson Sampling (Thompson, 1933) and its FGTS adaptation in Zhang (2021). To define the algorithm, we need some additional notation. At any round t of the algorithm, using the observed tuple $(x_t^h, a_t^h, r_t^h, x_t^{h+1})$, we define

$$\hat{\Delta}_t^h(g, f) = g(x_t^h, a_t^h) - r_t^h - f(x_t^{h+1}), \tag{4}$$

as a TD approximation of $\mathcal{E}(g, f, x_t^h, a_t^h)$. The algorithm is presented in Algorithm 1.

At a high-level, the algorithm performs standard Thompson Sampling updates to the posterior given the observations, with three modifications. First is that we incorporate an optimistic bias in the distribution p_t over f (Line 6), similar to the FGTS approach. The second difference arises from the challenges in estimating the Bellman error, while the third is in how we sample from the posterior to obtain the agent's policy at each time. We now explain the latter two issues in detail.

Inner loop to estimate Bellman error. Note that we would ideally define the *likelihood function* as $\mathcal{E}(f, f, x, \pi_f(x))^2$ for any x, but this requires a conditional expectation over the sampling of the next state $x' \sim P(\cdot \mid x, \pi_f(x))$. Replacing the expectation with a single random sample suffers from bias due to the double sampling issue (Antos et al., 2008), which arises from the non-linearity of squared loss inside the expectation. An ideal solution to this, following Antos et al. (2008) is to define the cumulative likelihood as:

$$\tilde{L}_t(f) = \sum_{s=1}^{t} \hat{\Delta}_s^{h_s}(f,f)^2 - \min_{g \in \mathcal{F}} \sum_{s=1}^{t} \hat{\Delta}_s^{h_s}(g,f)^2. \tag{5}$$

However, doing exact minimization over g creates instability in the online learning analysis of the distributions p_t. To ameliorate this, we instead replace the optimization over g with sampling from an appropriate distribution conditioned on f (Line 4). This distribution favors functions g which approximately minimize $\sum_{s=1}^{t} \hat{\Delta}_s^{h_s}(g,f)^2$. We then form a surrogate for the ideal likelihood of (5) for each round t in Line 5, where the second term acts as a soft minimization over g, given f.

Two samples to decouple roll-in and roll-out Using the likelihood function, it is straightforward to define a Thompson sampling distribution p_t over $f \in \mathcal{F}$. Our definition in Line 6 incorporates the optimistic term as well, giving higher weight to functions f that predict large values in the initial step. This resembles both the design of FGTS (Zhang, 2021) as well as OLIVE (Jiang et al., 2017). At this point, a typical TS approach would draw $f_t \sim p_t$ and act according to the resulting greedy policy. Doing so, however, creates a mismatch between the likelihood function we use to evaluate f_t and the guarantees we need on it for learning. This issue is most easily seen if we imagine the distribution p_t to be fixed across rounds (which is a reasonable intuition for a stable online learning algorithm). Then the likelihood at round t contains samples collected at prior rounds, when we drew $f_s \sim p$ independently of f_t for $s < t$. Thus we expect the likelihood of f to approach $\mathbb{E}_{f_s \sim p} \mathbb{E}_{(x^h, a^h) \sim \pi_{f_s}} [\mathcal{E}(f, f, x^h, a^h)^2]$. However, when we choose actions according to f_t, we require $\mathbb{E}_{x^h \sim \pi_{f_t}} [\mathcal{E}(f_t, f_t, x^h, \pi_{f_t}(x^h))^2]$ to be small.

Jiang et al. (2017) address the distributional mismatch over x^h using Bellman factorization, while the discrepancy between $a^h = \pi_{f_s}$ and $a^h = \pi_{f_t}(x^h)$ is addressed in OLIVE by a one-step uniform exploration over a finite action space, which is also adopted by subsequent works. In this paper, we replace this one-step uniform exploration by a second, independent sample from the posterior. Under the linear embedding assumption, this allows us to perform one-step exploration in infinite action spaces without any knowledge of the embedding features. One main insight of this work is to demonstrate the effectiveness of such a *two sample* strategy (Line 8). Specifically, we execute π_t for the first h_t time steps, and then complete the roll-out using π'_t with $h_t \in [H]$ chosen uniformly, and we use only the samples from step h_t in our likelihood at time t. The decoupling of the two samples is crucial to our analysis, although it will be interesting to investigate whether a single sample strategy can be analyzed using a different approach.

5. Main Results

In this section, we present our main sample complexity guarantee for TS^3. To do so, we need to introduce some measures of the complexity of our value function class $\mathcal{F}$, which we do next.

Definition 4 *For any $f \in \mathcal{F}$, we define the set $\mathcal{F}(\epsilon, f) = \left\{ g \in \mathcal{F} : \sup_{x,a,h} |\mathcal{E}^h(g, f; x, a)| \leq \epsilon \right\}$ of functions that have a small Bellman error with f for all x, a. Further assume that $\mathcal{F}$ has an L_∞ cover $f_1, \ldots, f_N \in \mathcal{F}$ for $N = N(\epsilon)$, so that $\forall f \in \mathcal{F}$, $\min_i \sup_{x,a} |f(x,a) - f_i(x,a)| \leq \epsilon$. Then we define $\kappa(\epsilon) = \sup_{f \in \mathcal{F}} - \ln p_0(\mathcal{F}(\epsilon, f))$ and $\kappa'(\epsilon) = \ln N(\epsilon)$, where $p_0 \in \Delta(\mathcal{F})$ is the prior.*

If $\mathcal{F}$ is finite with $|\mathcal{F}| = N$, then we can choose p_0 to be uniform over $\mathcal{F}$ to get $\kappa(\epsilon) = \kappa'(\epsilon) = \ln N$. We state our guarantee in terms of the regret of the learned policies with respect to the optimal policy $\pi_\star$ (1), that we define for a greedy policy π_f with some $f \in \mathcal{F}$ as:

$$\mathrm{Regret}(f, x^1) = R(\pi_\star, x^1) - R(\pi_f, x^1), \;\text{ where }\; R(\pi, x^1) = \mathbb{E}_{(x,a,r)\sim\pi|x^1} \sum_{h=1}^{H} r^h$$

The following result gives our main sample complexity bound for TS^3.

Theorem 5 *Under Assumptions 1-4, suppose we run TS^3 (Alg. 1) with parameters $\gamma = 0.1$ and $\eta \le c/(\kappa(\epsilon) + \kappa'(\epsilon) + \ln T)$ for a universal constant c. Then choosing any $\epsilon = \mathcal{O}(1/T)$, we have*

$$\mathbb{E} \sum_{t=1}^{T} \mathrm{Regret}(f_t, x_t^1) = \mathcal{O}\left(\frac{\kappa(\epsilon) + \kappa'(\epsilon)}{\lambda} + \lambda T + \tilde{\epsilon}(\lambda/\eta) T\right),$$

where $\tilde{\epsilon}(\lambda/\eta) = \inf_{\mu>0}\left[8(\lambda/\eta)\mathrm{dc}(\epsilon_2)H\mu^2 + 2\mu H\epsilon_2 B_2 + \mu^{-1}\mathrm{br}(\epsilon_1) + \epsilon_1 H B_1\right]$.

Note that the bound above is not a bound on the online regret of TS^3, since the policy we execute at round t is not π_t, but a mixture of π_t and π_t' (line 8 of Algorithm 1). However, the bound implies a PAC guarantee, when the contexts are i.i.d., as we can choose a policy $\pi_1, \ldots, \pi_T$ uniformly, and use a standard online-to-batch conversion (Cesa-Bianchi et al., 2004) to obtain a regret bound for the returned policy.

To interpret the general guarantee of Theorem 5, we now consider some special cases where the parameters can be set optimally to simplify our result. We start with the well-studied setting of a finite function class, with Bellman rank and linear embedding dimensions being finite as well.

Corollary 6 (Finite dimensional embeddings and finite $\mathcal{F}$) *Under conditions of Theorem 5, assume further that $|\mathcal{F}| = N < \infty$, $\mathrm{br}(0) \le d_1$ and $\mathrm{dc}(0) \le d_2$. Then TS^3 with p_0 as uniform on $\mathcal{F}$, $\gamma = 0.1$, $\eta = \mathcal{O}(1/\ln(NT))$ and $\lambda = T^{-3/4}H\,(d_1^2 d_2)^{-1/4}(\ln(NT))^{1/2}$ returns a policy with an average regret at most $\mathcal{O}\left(H\, d_2^{1/4}\sqrt{d_1 \ln(NT)}\, T^{-1/4}\right)$.*

The bound has a favorable dependence on all the complexity parameters, with mild scaling in the horizon, dimensions and function class complexity. However, the average regret decays at a rate of $T^{-1/4}$. In the case of a small Bellman rank and finite actions, our sample complexity is $\mathcal{O}\left(H^4(d_1 \ln(NT))^2\,|\mathcal{A}|\,\epsilon^{-4}\right)$, which is suboptimal and slower than that of OLIVE in ϵ dependence. In the setting of a linear MDP, where $d_1 = d_2 = d$, the scaling with dimension is d^3.

The main source of the suboptimality in ϵ is that our analysis relies on a decoupling argument that leverages Assumption 4 with an extra Cauchy-Schwarz inequality than in the typical analysis. The importance weighting over a uniform distribution in OLIVE does not lose an exponent in this step, and this can be extended to infinite action setting if the embedding feature $\phi^h(\cdot,\cdot)$ is known (see Section 6). However, with unknown $\phi^h(\cdot,\cdot)$, improving the analysis is an open direction for future work. Note however that all prior works that studied both finite action non-linear and infinite action (generalized) linear settings employ different algorithms for the two cases (Jin et al., 2021; Du et al., 2021; Foster et al., 2021), unlike our unified approach.

As a first example of our general result, we can instantiate the setting of Example 1, where our sample complexity only scales with $\mathcal{O}(KL)$ instead of $\mathcal{O}(K^L)$. In the setting of Example 2, we only depend on the size of the (potentially unknown) basis.

Under the infinite-dimensional examples of Proposition 2, we note that the geometric case is almost identical to the finite dimensional setting up to log factors, since the *effective rank* only scales as $\log(1/\epsilon)$, so that we can always make the additional terms coming from ϵ to be lower order. The polynomial case does yield qualitatively different results, which we show next.

Corollary 7 (Polynomial spectral decay and finite $\mathcal{F}$) *Under conditions of Theorem 5, assume further that $|\mathcal{F}| = N < \infty$, $B_1 = B_2 = B$ and $\mathrm{br}(\epsilon), \mathrm{dc}(\epsilon) \leq H\, d_q \epsilon^{-2q/(1-q)}$ for some $q \in (0, 1)$. Then TS^3 with p_0 as uniform on $\mathcal{F}$, $\gamma = 1/36$, along with appropriate settings of η and λ returns a policy with an average regret at most $\mathcal{O}\left(H(\ln(NT))^{(1-q)^2/2} d^{(3-q)(1-q)/4}\, T^{-(1-q)^2/4} B^{q(3-q)/2}\right)$.*

The result matches that of Corollary 6 for $q = 0$ corresponding to the case of finite dimensions. More generally, we allow scaling to infinite dimensional states and actions, as long as the Bellman rank and linear embedding assumptions have a low intrinsic dimension.

Our results have a qualitatively similar flavor to those of Zhang (2021), but for two important differences. Zhang (2021) do not need to account for the residual variance term in the likelihood due to the deterministic dynamics assumption, and hence they do not incur the loss in rates that we suffer. They also do not work under the low Bellman rank assumption on the problem, which significantly limits the class of problems where their guarantees hold.

We finally illustrate the benefits of our contextual formulation by studying mixtures of MDPs with a small Bellman rank.

Corollary 8 (Mixture of low-rank MDPs) *Consider a collection of M different MDPs $\{\mathcal{M}_i\}_{i=1}^M$ over the same state and action space, each with a Bellman rank at most d. Let $x^1 \sim \mathcal{D}$ where $\mathcal{D}$ is uniform over $[M]$, with the subsequent transitions happening according $\mathcal{M}_i$. Suppose $|\mathcal{F}| = N$. Then under the parameter settings of Corollary 6, TS^3 returns a policy with a regret at most $\mathcal{O}\left(H\, d_2^{1/4} \sqrt{d_1 \ln(NT)}\, T^{-1/4}\right)$.*

Comparing Corollaries 8 and 6, we observe that the mixture setting poses no extra challenge. In contrast, the mixture problem has a Bellman rank scaling with M, as we need to concatenate the respective embeddings for each $\mathcal{M}_i$, causing a naïve application of OLIVE to incur an extra M^2 term in the sample complexity. We believe the difference arises since the average Bellman error of OLIVE mixes samples across different contexts, while we use the squared Bellman error which can leverage the structure for individual transitions more effectively. Of course, this comes at the price of an additional completeness assumption.

6. Experimental design for known linear embedding features

In this section, we study a special case of our setting where the features $\phi(x, a)$ in Assumption 4 are known, and finite dimensional, with $\phi(x, a) \in \mathbb{R}^{d_2}$. For this setting, we consider an adaptation of TS^3 which replaces the two sample strategy with experimental design in the feature space. The algorithm is presented in Algorithm 2.

Algorithm 2 Two timeScale Thompson Sampling with Design (TS^2-D)

```
Require: Function class $\mathcal{F}$, prior $p_0 \in \Delta(\mathcal{F})$, learning rates $\eta, \gamma$ and optimism coefficient $\lambda$.
 1: Set $S_0 = \emptyset$.
 2: for $t = 1, \ldots, T$ do
 3:     Observe $x_t^1 \sim \mathcal{D}$ and draw $h_t \sim \{1, \ldots, H\}$ uniformly at random.
 4:     Define $q_t(g|f) = p(g|f, S_{t-1}) \propto p_0(g) \exp(-\gamma \sum_{s=1}^{t-1} \hat{\Delta}_s^{h_s}(g, f)^2)$.▷ Inner loop TS update
 5:     Define $L_t^h(f) = \eta \hat{\Delta}_t^h(f, f)^2 + \frac{\eta}{\gamma} \ln \mathbb{E}_{g \sim q_t(\cdot|f)} \left[\exp(-\gamma \hat{\Delta}_t^h(g, f)^2)\right]$.▷ Likelihood function
 6:     Define $p_t(f) = p(f|S_{t-1}) \propto p_0(f) \exp(\sum_{s=1}^{t-1} (\lambda f(x_s^1) - L_s^{h_s}(f))$ as the posterior.
            ▷ Outer loop Optimistic TS update
 7:     Draw $f_t \sim p_t$ and let $\pi_t = \pi_{f_t}$. Execute $a_t^h = \pi_t(x_t^h)$ for $h = 1, \ldots, h_t - 1$ to observe $x_t^{h_t}$.
 8:     Let $\rho_t \in \Delta(\mathcal{A})$ be a $G$-optimal design for $\phi(x_t^{h_t}, a)_{a \in \mathcal{A}}$ (Equation 6). Draw $a_{h_t}^t \sim \rho_t$ and
        observe $r_t^{h_t}$ and $x_t^{h_t+1}$.                ▷ $G$-optimal design
 9:     Update $S_t = S_{t-1} \cup \{x_t^h, a_t^h, r_t^h, x_t^{h+1}\}$ for $h = h_t$.
10: end for
11: return $(\pi_1, \ldots, \pi_T)$.
```

Before we delve into the pseudocode, recall that the G-optimal design, given a set of vectors $\{\phi(x, a)\}_{a \in \mathcal{A}}$ is a distribution $\rho(x) \in \Delta(\mathcal{A})$ over the action space, given by (see e.g. Fedorov, 2013; Kiefer and Wolfowitz, 1960).

$$\rho(x) = \operatorname*{argmin}_{\rho \in \Delta(\mathcal{A})} \underbrace{\max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_x(\rho)^{-1}}}_{g(x,\rho)} \quad \text{where} \quad \Sigma_x(\rho) = \mathbb{E}_{a \sim \rho} \phi(x, a) \phi(x, a)^\top. \tag{6}$$

If $\Sigma_x(\rho)$ is not full rank, then we can replace $\Sigma_x(\rho)^{-1}$ by the corresponding pseudo-inverse. Furthermore, by the Kiefer-Wolfowitz theorem (Kiefer and Wolfowitz, 1960), it is known that the design $\rho(x)$ satisfies $g(x, \rho(x)) = \text{rank}(\{\phi(x, a) : a \in \mathcal{A}\}) \leq d_2$. The criterion $g(\rho)$ corresponds to worst prediction variance at some action a, of an ordinary least squares estimator given samples drawn from $\rho(x)$. For intuition in the finite action setting, where $\phi(x, a) = e_a$, $\rho(x)$ corresponds to a uniform distribution over the action set and $d_2 = |\mathcal{A}|$ is the variance of importance sampling.

With this context, TS^2-D is a relatively natural adaptation of TS^3, when $\phi(x, a)$ is a known feature map. Concretely, we no longer use the two sample scheme. Instead, we only draw one function $f_t \sim p_t$ in Line 7 and execute the first $h_t - 1$ actions using the corresponding greedy policy. Having observed $x_t^{h_t}$, we now choose the action at time t using a G-optimal design in the feature space for *this particular state* in Line 8. That is, the action sampling distribution ρ_t corresponds to $\rho(x_t^{h_t})$ from Equation 6, and the design is done individually at each state. We then observe the reward and next transition for this action, which gets recorded into our dataset as before.

Theorem 9 *Under Assumptions 1-4, suppose further that* $\text{dc}(0) = d_2$ *and the features* $\phi^h(x, a)$ *are known. Suppose we run TS^2-D (Alg. 2) with* $\gamma = 0.1$ *and* $\eta \leq c/(\kappa(\epsilon) + \kappa'(\epsilon) + \ln T)$ *for some* *constant c. Then for any $\epsilon = \mathcal{O}(1/T)$, $\epsilon_1 > 0$ and $\lambda^{-1} = \sqrt{T} \min\{0.5, \mathrm{br}(\epsilon_1)d_2 H\}$, we have*

$$\mathbb{E} \sum_{t=1}^{T} \mathrm{Regret}(f_t, x_t^1) = \mathcal{O}\left(\epsilon_1 H B_1 + (\kappa(\epsilon) + \kappa'(\epsilon) + \ln T)\sqrt{\frac{\mathrm{br}(\epsilon_1) d_2 H}{T}}\right).$$

Thus, in the setting of Corollary 6, we immediately get a $1/\sqrt{T}$ rate in the above bound, leading to a $\mathcal{O}(d_1 d_2 H \ln^2(NT)/\epsilon^2)$ sample complexity. This bound is superior or comparable to that for OLIVE in all the problem parameters, except for an extra dependency on $\ln N$. This extra dependency is caused by the two timescale learning rates (discrepancy between η and γ) in the online minimax game analysis, which might be possible to improve in future work. For tabular problems with S states and A actions, where $d_1 = S$, $d_2 = A$ and $\ln N = \tilde{\mathcal{O}}(SAH)$, the bounds for TS^2-D as well as OLIVE scale as $S^3 A^2$, with an additional H^2 term in $OLIVE$ compared to the bound of TS^2-D. For linear MDPs with finite dimensional features, $d_1 = d_2 = d$ and $\ln N = \mathcal{O}(d)$, so that our bound scales as $\mathcal{O}(d^4)$, which is a factor of d worse relative to LSVI-UCB (Jin et al., 2020), and a factor of d^2 worse relatively to (Zanette et al., 2020b). One reason for the suboptimality in d is due to our choice of V-type Bellman rank instead of Q-type Bellman rank to allow non-linear scenarios. For the nicer setting of Q-type Bellman rank, an analysis of a single sample based Thompson sampling is recently carried out in Dann et al. (2021), leading to a result that matches that of (Zanette et al., 2020b) for linear MDPs. Another reason for the suboptimality is because our MDP model allows long range contextual dependency on x_t^1, which is not allowed in most prior works. If we remove this dependency by considering only non-contextual MDP models, then we can replace the online regret analysis of this paper by the uniform convergence analysis of (Dann et al., 2021). Doing so avoids the slow-fast learning rate issue in our online minimax analysis and implies a sample complexity of $\mathcal{O}(d_1 d_2 H \ln(NT)/\epsilon^2)$. However, the technique cannot be used to analyze double-posterior sampling, and thus we do not consider it in this paper. Note that the recent work of Foster et al. (2021) also analyzes design-based approaches for model based RL, but both their problem setting and algorithmic details differ significantly from ours. The result of Theorem 9 demonstrates that the suboptimality in ϵ in our more general result of Theorem 5 stems purely from the harder setting of unknown linear embedding features. We leave the development of corresponding results for spectral decay and combinatorial action settings to the reader.

In terms of analysis, the only change is that we are able to do different handling of a decoupling step using the property of optimal design in the analysis of Theorem 9, and the error terms in the linear embedding can be ignored due to the finite dimensional assumption. The rest of our arguments stay the same, and we provide a proof in Appendix I. We now give an overview of our analysis techniques.

7. Proof sketch of Theorem 5

In this section, we provide the analysis for our main result on the sample complexity of TS^3. To begin, We recall a standard result on the online regret of any value-based RL algorithm.

Proposition 10 (Jiang et al. (2017)) *Given any $f \in \mathcal{F}$ and $x^1 \in \mathcal{X}^1$, we have*

$$\mathrm{Regret}(f, x^1) = \sum_{h=1}^{H} \mathbb{E}_{(x^h, a^h)\sim \pi_f | x^1} \mathcal{E}^h(f, f, x^h, a^h) - \Delta f(x^1), \quad \textit{where} \;\; \Delta f(x^1) = f(x^1) - Q_\star^1(x^1).$$

The proposition is a consequence of a simple telescoping argument. The RHS looks related to the likelihood function in our update (line 6 in Algorithm 1), but there are a few important differences. First, we do not have access to the Bellman error, which is instead approximated through the difference of the TD term $\hat{\Delta}_t(f,f)^2$ and the residual variance measured using $\hat{\Delta}_t(g,f)^2$. If the posterior of g concentrates around the Bellman projection of f onto $\mathcal{F}$, then we can expect our likelihood to resemble $\mathcal{E}^h(f,f,x^h,a^h)^2$. The presence of a squared term instead of the linear dependence in Proposition 10 is typical to algorithms which use completeness (Jin et al., 2021). However, there are two more significant issues. On the RHS of Proposition 10, the function f whose Bellman error is measured is identical to the one whose greedy policy picks all the actions. On the other hand, in our algorithm, we control the regression loss of a function f that is different from the roll-in policy. For non-linear functions f, prior works (Jiang et al., 2017; Agarwal et al., 2020b) control this change in measure over the states using the low-rank property, while the distribution of the final action is corrected by importance weighting over the finite action set. For linear functions, change of measure over both x^h and a^h can be done using a similar use of the low-rank property, without explicit weighting over the actions (Jin et al., 2020; Du et al., 2021; Jin et al., 2021).

In our case, we adopt a slightly different approach to analyze the Bellman error term in Proposition 10. We first apply the low-rank property to *decouple* the roll-in policy from the f being evaluated. We subsequently use the linear embeddability assumption to *decouple* the action selection at step h. This part of our analysis resembles that of (Zhang, 2021) for the contextual bandit setting, which is the reason we adopt the name *decoupling coefficient* as their work, for the measure of the linear embeddability dimension. We call these two results decoupling lemmas, which can be found in Appendix G. Using the two decoupling coefficients, we can obtain the following result.

Proposition 11 (Decoupling) *For any $t \geq 1$, we have*

$$\lambda \mathbb{E}\ \mathrm{Regret}(f_t, x_t^1) \leq \mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\left[-\lambda\Delta f(x_t^1) + 0.5\eta\mathcal{E}^{h_t}(f,f,x_t^{h_t},a_t^{h_t})^2\right] + \lambda\tilde{\epsilon}(\lambda/\eta),$$

where $\tilde{\epsilon}(\lambda/\eta) = \inf_{\mu>0}\left[8(\lambda/\eta)\mathrm{dc}(\epsilon_2)H\mu^2 + 2\mu H\epsilon_2 B_2 + \mu^{-1}\mathrm{br}(\epsilon_1) + \epsilon_1 HB_1\right]$.

For optimal parameter settings, the bound scales with $\sqrt{\mathrm{dc}(\epsilon_2)\mathbb{E}\mathbb{E}_{f|S_{t-1}}\mathcal{E}(f,f;x_t^{h_t},a_t^{h_t})^2}$, and this square root is responsible for our $O(\epsilon^{-4})$ rate. In contrast with Proposition 10, Proposition 11 measures the squared Bellman error of functions f according to the states x_t^h and actions a_t^h observed during the algorithm's execution, and which we can hope to control if the TS^3 updates converge to their respective optima for both the time scales. The remainder of our analysis focuses on this online learning component, details of which are presented in Appendix H. We now give our main result to control the regret of the online learning process.

Proposition 12 (Outer loop convergence) *Assume that $\gamma = 0.1$ and $\eta \leq 0.01$. Then we have*

$$-\sum_{t=1}^{T}\mathbb{E}\mathbb{E}_{f|S_{t-1}}\lambda\Delta f(x_t^1) + \eta(1-6\eta)e^{-12\eta(1-6\eta)}\sum_{t=1}^{T}\mathbb{E}\mathbb{E}_{f|S_{t-1}}\mathcal{E}^{h_t}(f,f;x_t^{h_t},a_t^{h_t})^2$$

$$\leq \eta(1+6\eta)e^{12\eta(1+6\eta)}\sum_{t=1}^{T}\mathbb{E}\mathbb{E}_{f,g|S_{t-1}}\mathcal{E}^{h_t}(g,f;x_t^{h_t},a_t^{h_t})^2 + \lambda\epsilon T + 4\eta T\epsilon^2 + \kappa(\epsilon) + 1.5\lambda^2 T.$$

We prove Proposition 12 in Appendix H.2. The LHS of the proposition qualitatively resembles the RHS of Propositoin 11. Indeed, we subsequently set the constants so that they match. This means that the regret in Proposition 10 can be further upper bounded by the RHS of Proposition 12. The first term in the bound is intuitive. It bounds the Bellman residual of f in terms of quality of the functions $g \sim p(\cdot \mid f, S_{t-1})$ in capturing the Bellman operator applied to f. If the inner loop of TS^3 converges at a reasonable rate, we expect this error to be small by the completeness assumption. The next three terms on the RHS are a bound on the log-partition function in our outer loop updates, which are controlled by using typical potential function arguments in the analysis of multiplicative updates. The final term can be controlled by appropriate setting of the optimism parameter λ.

We now give the main bound on the convergence of our inner updates to control the first term on the RHS of Proposition 12.

Proposition 13 *Assume that $\gamma = 0.1$. Given any absolute constant $c_1 > 0$, there exists an absolute constant c_0 such that when $c_0\eta(\kappa(\epsilon) + \kappa'(\epsilon) + \ln T) \leq 0.5c_1\gamma < 0.1$, then*

$$\mathbb{E} \sum_{t=1}^{T} \mathbb{E}_{f,g|S_{t-1}} \mathcal{E}^{h_t}(g, f; x_t^{h_t}, a_t^{h_t})^2 \leq \mathcal{O}(\epsilon T + \kappa(\epsilon) + \kappa'(\epsilon) + 1) + 2c_1\gamma \sum_{t=1}^{T} \mathbb{E}\, |\mathcal{E}^{h_t}(f, f; x_t^{h_t}, a_t^{h_t})|^2.$$

The proof of Proposition 13 is rather long and technical. For the reader's convenience, we first prove a simpler result which yields a worse $\mathcal{O}(\epsilon^{-8})$ sample complexity in Appendix H.3. We then show how to improve the bound to obtain Proposition 13 in Appendix H.4.

With these results, we set both c_1 and η sufficiently small so that $(1 - 6\eta)\exp(-12(1 - 6\eta)) - 2(1 + 6\eta)c_1\gamma e^{12\eta(1+6\eta)} \geq 0.5$ along with the stated values of ϵ and γ. Plugging these into the bounds of our intermediate results and simplifying gives the conclusion of Theorem 5. Finally, we summarize the proof of Corollary 6 and defer that of Corollary 7 to Appendix F.

Proof [Proof of Corollary 6] We set the parameters as follows. Since $\epsilon_1 = \epsilon_2 = 0$, we minimize over μ to get $\mu = (d_1\eta/(d_2\lambda H))^{1/3}$. We now optimize over the choice of η, for which the leading order terms are $\eta^2 T \ln N/\lambda + T\tilde{\epsilon}(\lambda/\eta)$. Then optimizing for λ by including the $\ln N/\lambda$ term yields the stated guarantee. ■

8. Conclusion and Discussion

In this paper, we combine the framework of low Bellman rank with a linear embedding assumption over the action space to introduce a new class of rich problems which encompasses all settings with finite actions and linear function approximation, and enables new ones such as combinatorial action spaces. We show that TS^3 solves this class of problems under the usual completeness and realizability assumptions on the value function class. We believe that the identification of this linear embedding structure over actions as the key enabler of *one-step exploration* in the action space has the potential to apply to broader algorithmic approaches beyond TS^3.

The most immediate direction for future work is to improve our sample complexity results by sharpening our decoupling results. Understanding if similar results are attainable without completeness is another challenge. Finally, it would be interesting to understand what structures beyond linear embeddability afford sample-efficiency, and study applications of these ideas to continuous control problems.

References

Yasin Abbasi-Yadkori and Gergely Neu. Online learning in mdps with side information. *arXiv preprint arXiv:1406.6812*, 2014.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 2020a.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 2020b.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019.

Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Kavosh Asadi, Neev Parikh, Ronald E Parr, George D Konidaris, and Michael L Littman. Deep radial-basis value functions for continuous control. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 6696–6704, 2021.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996. ISBN 1886529108.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR, 2018.

Christoph Dann. personal communication, 2018.

Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. In *Neurips*, 2021. URL papers/neurips21-rl.pdf.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4): 633–679, 2020.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. *International Conference on Machine Learning*, 2021.

V.V. Fedorov. *Theory Of Optimal Experiments*. Probability and Mathematical Statistics. Elsevier Science, 2013. ISBN 9780323162463. URL https://books.google.com/books?id=PwUz-uXnImcC.

Fei Feng, Wotao Yin, Alekh Agarwal, and Lin Yang. Provably correct optimization and exploration with non-linear policies. In *International Conference on Machine Learning*, pages 3263–3273. PMLR, 2021.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, et al. Reinforcement learning for slate-based recommender systems: A tractable decomposition and practical methodology. *arXiv preprint arXiv:1905.12767*, 2019.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR, 2018.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 2021.

Sham M. Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *NeurIPS*, 2020.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *arXiv preprint arXiv:1902.07826*, 2019.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Zakaria Mhammedi, Dylan J Foster, Max Simchowitz, Dipendra Misra, Wen Sun, Akshay Krishnamurthy, Alexander Rakhlin, and John Langford. Learning the linear quadratic regulator from nonlinear observations. *arXiv preprint arXiv:2010.03799*, 2020.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618. PMLR, 2018.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *arXiv preprint arXiv:1306.0940*, 2013.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 2014. ISBN 9781118625873. URL https://books.google.com/books?id=VvBjBAAAQBAJ.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *arXiv preprint arXiv:1906.02870*, 2019.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*, 2017.

Ayush Sekhari, Christoph Dann, Mehryar Mohri, Yishay Mansour, and Karthik Sridharan. Agnostic reinforcement learning with low-rank mdps and rich observations. *Advances in Neural Information Processing Systems*, 34, 2021.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000, 2009.

R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 1998. ISBN 9780262303842. URL https://books.google.com/books?id=U57uDwAAQBAJ.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

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

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020b.

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 2021. to appear.

O.C. Zienkiewicz, R.L. Taylor, and J.Z. Zhu. *The Finite Element Method: Its Basis and Fundamentals*. Elsevier Science, 2005.

Appendix A. Related Work

Settings for sample-efficient RL. There is substantial recent literature (Jiang et al., 2017; Sun et al., 2019; Jin et al., 2020; Yang and Wang, 2020; Du et al., 2019, 2021; Jin et al., 2021; Foster et al., 2021) on structural conditions that enable sample-efficient RL. Of these, the Bellman rank framework and its recent generalization in the Bilinear classes model remain the broadest known frameworks for which provably sample-efficient methods are known. While Bellman rank itself makes no assumptions on the complexity of the action space (Jiang et al. (2017) show small Bellman rank for LQRs), the algorithm OLIVE developed for this setting crucially relies on importance sampling over a discrete action set. Ideas from these works have further been developed to a representation learning setting, where the transition dynamics of the MDP are linear in some *unknown features* and the agent learns this map, given a class of candidate features (Agarwal et al., 2020b; Uehara et al., 2021; Modi et al., 2021), which captures rich non-linear function approximation in the original state and action. Jin et al. (2021) developed the Bellman-Eluder dimension notion to better capture infinite action sets using a notion they call Q-type Bellman rank, while the original version of Jiang et al. (2017) is termed the V-type Bellman rank. Note, however, that the GOLF algorithm of Jin et al. (2021) for problems with a small Q-type Bellman rank, which scales to infinite action spaces, cannot be used for feature learning and does not capture all contextual bandit problems (see Section B for a lower bound), showing the limitations of this approach in terms of the non-linearity it permits. For V-type Bellman-Eluder dimension, Jin et al. (2021) also rely on uniform exploration over actions similar to OLIVE. The techniques developed here do scale to feature learning, capture contextual bandits fully and apply to infinite action scenarios satisfying the linear embedding assumption. At the same time, our assumptions do not capture all problems with a small Q-type Bellman rank, so there are problems that GOLF can handle which we do not. That said, perhaps the most prominent example for GOLF in the infinite action setting is that of linear MDPs, where the linear embedding assumption made here holds.

Continuous control. Large action spaces are the standard framing in continuous control, where the action is typically a vector in $\mathbb{R}^d$ for some control input dimension d. While there has been a number of recent results at the intersection of learning and control (see e.g. Dean et al., 2020; Mania et al., 2019; Agarwal et al., 2019; Simchowitz and Foster, 2020), a large body of work typically focuses on highly structured settings such as the Linear Quadratic Regulator (LQR), where online exploration is is straightforward due to the presence of a Gaussian noise in the dynamics. More recent results (Kakade et al., 2020; Mania et al., 2020) do combine online control and exploration, they typically focus on model-based settings and still rely on access to good features. We note that Mhammedi et al. (2020) carry out feature learning for continuous control, but their setting does not have a small Bellman rank and hence is not admissible in our conditions either.

Posteroior sampling. Posterior sampling methods for RL, motivated by Thompson sampling Thompson (1933), have been extensively developed and analyzed in terms of their expected regret under a Bayesian prior by many authors (see e.g. Osband et al., 2013; Russo et al., 2017; Osband et al., 2016) and are often popular as they offer a simple implementation heuristic through approximation by ensembles trained on random subsets of data. Worst-case analysis of TS in RL settings has also been done for both tabular (Russo, 2019; Agrawal and Jia, 2017) and linear (Zanette et al., 2020a) settings. Our work is most closely related to the recent Feel-Good Thompson Sampling strategy proposed and analyzed in (Zhang, 2021), primarily for bandits but also for RL problems with deterministic dynamics. They study problems with a similiar linear embeddability assumption, but the absence of any further structure like a small Bellman rank precludes their techniques for application to general stochastic dynamics. We also observe that FGTS retains the significant optimistic component of other exploration techniques like LSVI-UCB (Jin et al., 2020) and OLIVE (Jiang et al., 2017), which partly explains its success in worst-case settings. On the other hand, the approach has the remarkable property of working for both linear bandits and non-linear bandits with finite action sets with an identical algorithm and analysis, and we extend that benefit to RL in this work.

Minimax objectives in RL. FGTS algorithm uses a likelihood term for Thompson Sampling based on the TD error, which is well-known to have a bias in estimating the Bellman error (Antos et al., 2008; Sutton and Barto, 1998) for stochastic dynamics. The usual technique of removing the residual variance from Antos et al. (2008) creates a minimax objective, which we also use in this paper. Other minimax formulations (Dai et al., 2018) to remove this bias are also possible in general, but we find that the one from Antos et al. (2008) is the most natural under our structural assumptions. We also note that approach of keeping two timescales (Borkar, 2009) used here has been used previously in offline RL for TD learning methods (Sutton et al., 2009), but its online analysis in TS appears novel, and different from the analysis in Dann et al. (2021), which cannot handle general nonlinear feature learning considered here.

Appendix B. Lower bound for Q-type Bellman rank

We now instantiate a contextual bandit problem where the realizability assumption holds, but the Q-type Bellman rank grows linearly in the number of contexts. The construction is due to Dann (2018), but has not appeared in the literature. Note that the V-type Bellman rank of Jiang et al. (2017), which we further generalize in this work, is always 1 for a contextual bandit problem. The lower bound is demonstrated using a typical hard instance for contextual bandit problems. Let us consider a context distribution which is uniform on $[N]$, where we have N unique contexts. We have two actions $\{a_1, a_2\}$. We also have $|\mathcal{F}| = N+1$ with the following structure.

$$f^\star(x, a_1) = f_{N+1}(x, a_1) = 0, \quad \text{and} \quad f^\star(x, a_2) = f_{N+1}(x, a_2) = 0.5.$$

For $i < N+1$, we have $f_i(x, a) = f^\star(x, a)$ when $x \neq i$, and $f_i(x, a_1) = 1$, $f_i(x, a_2) = 0.5$ so that f_i makes incorrect predictions on the context i for action a_1. Notice that the design of $\mathcal{F}$ also ensures that for each context i, there is a policy $\pi_\star$ (greedy wrt $f^\star$) which picks the action a_2, while another policy π_i (greedy wrt f_i) which picks action a_1. Since this is a problem with horizon $H = 1$, the Bellman error is simply equal to $f(x, a) - f^\star(x, a)$ for any x, a. Let $\mathcal{E}^Q(f, \pi) = \mathbb{E}_{x, a\sim\pi}\mathcal{E}(f, f, x, a)$ be the Q-type Bellman error. Then, we observe that for $1 \leq i, j \leq N$:

$$\mathcal{E}^Q(f_i, \pi_j) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{a\sim\pi_j}\left[f(x, a) - f^\star(x, a)\right] = \frac{\mathbf{1}(j = i)}{N},$$

where $\mathbf{1}(\cdot)$ is an indicator function. Hence, we see that the Bellman error matrix contains an identity submatrix of size N, so that its rank is at least N.

Appendix C. Connection with sparse RL

In this section, we formalize the relationship between representation learning and RL with sparsity. Concretely, let us consider two formulations.

$$P(x' \mid x,a) = \phi^{lr}(x,a)^\top \mu^{lr}(x'), \quad \text{where} \quad (\phi^{lr},\mu^{lr}) \in \Omega^{lr}, \tag{$P1$}$$

is the standard model-based representation learning formulation for RL (Agarwal et al., 2020b; Uehara et al., 2021). Here we consider a slightly generalized setup where the feature maps (ϕ, μ) are available as pairs from a joint set Ω^{lr} instead of separate classes for ϕ and μ, which reduces to the more typical framing with separate classes for ϕ and μ when we set Ω^{lr} to be the cartesian product of their respective sets. However, all existing representation learning algorithms can handle this setup for a general Ω^{lr} without any extra difficulty. Let us denote $d^{lr} = \text{dim}(\phi^{lr})$, where we use the superscript lr to denote the low-rank problem $P1$. The second formulation is the sparse linear MDP setup of Hao et al. (2021).

$$P(x' \mid x,a) = \sum_{i\in\mathcal{I}} \phi^s(x,a)_i \mu^s(x')_i, \quad \text{where } |\mathcal{I}| = k \ll d^s = \text{dim}(\phi^s), \tag{$P2$}$$

with the feature maps ϕ^s, μ^s considered known in $P2$, but the index set $\mathcal{I}$ is unknown. We now show that the two problems are equivalent when $d^{lr} \cdot |\Omega^{lr}| = d^s$ and $k = d^{lr}$, in that any solution to $P1$ can solve $P2$ at no additional sample cost, and vice versa.

In one direction, We define the concatenated feature map for all x, a and x':

$$\begin{aligned}\phi^{all}(x,a) &= (\phi_1(x,a), \phi_2(x,a), \ldots, \phi_N(x,a)), \quad \text{where } N = |\Omega^{lr}|, \text{ and}\\ \mu^{all}(x') &= (\mu_1(x'), \mu_2(x'), \ldots, \mu_N(x')),\end{aligned}$$

where ϕ_i, μ_i refer to the i_{th} feature maps in Ω^{lr}. Clearly, the assumption $\phi^{lr}, \mu^{lr} \in \Omega$ guarantees that $P(x' \mid x,a) = \phi^{all}(x,a)^\top \mu^{all}(x')$ with $\text{dim}(\phi^{all}) = d^{lr}|\Omega^{lr}| = d^s$, by construction. However, the transitions are also sparse in the features (ϕ^{all}, μ^{all}), in that we can choose the d^{lr} coordinates of ϕ^{all}, μ^{all} which correspond to the index of (ϕ^{lr}, μ^{lr}) in Ω^{lr}. Consequently, any solution to $P2$ for $k = d^{lr}$ which uses samples that are $\mathcal{O}(\text{poly}(k \log d^s)$ can be used to solve $P1$ with a sample complexity of $\mathcal{O}(\text{poly}(d^{lr} \log(d^{lr}|\Omega^{lr}|))$, which is considered the standard goal of the representation learning problem $P1$.

In the other direction, let us say we have a solution to $P1$ with a sample complexity that is $\mathcal{O}(\text{poly}(d^{lr} \log(|\Omega^{lr}|))$. Then we given a pair of high-dimensional feature maps ϕ^s, μ^s, we define a class Ω^{lr} as follows:

$$\Omega^{lr} = \{(\phi,\mu) \ : \ \phi(x,a) = \phi^s(x,a)_{\mathcal{J}}, \mu(x') = \mu^s(x')_{\mathcal{J}}, \quad \text{where } \mathcal{J} \subseteq \{1,\ldots,d^s\} \text{ with } |\mathcal{J}| = k\}.$$

In words, we add features corresponding to every subset of size k from the original d^s features as a candidate representation to Ω^{lr}. Then $|\Omega^{lr}| = \binom{d^s}{k} = \mathcal{O}((d^s)^k)$ and each feature map in Ω^{lr} has $d^{lr} = k$. Clearly the sparse linear MDP assumption in the definition of $P2$ ($P2$) implies that Ω^{lr} contains a pair (ϕ, μ) under which the MDP is linear. Furthermore, our method for representation learning applied to the sparse linear MDP has a sample complexity of $\mathcal{O}(\text{poly}(k \log d^s))$.

The above reduction show that any obstacle to sparse linear MDP learning also results in a lower bound for representation learning. In particular, the construction of Hao et al. (2021) precludes learning sparse linear MDPs where the action set has a cardinality $\mathcal{O}(d^s)$, so that we cannot expect to carry out representation learning in arbitrarily large action spaces without further assumptions.

Appendix D. Additional example satisfying Assumption 4

Example 2 (Basis expansions in action space) *For continuous action problems, Asadi et al. (2021) study the class of deep radial basis value functions, where any $f \in \mathcal{F}$ takes the form*

$$f(x, a; \theta) = \frac{\sum_{i=1}^{N} \exp(-\beta\|a - a_i(x;\theta)\|)v_i(x;\theta)}{\sum_{i=1}^{N} \exp(-\beta\|a - a_i(x;\theta)\|)},$$

where $a_i(x;\theta)$ are a (state and θ-dependent) basis, while $v_i(x;\theta)$ are some reference values at these points. If we consider the case where the $a_i(x;\theta)$ are only dependent on the state x, but fixed across θ, then this definition satisfies Assumption 4 with $w(f, x) = v_i(x;\theta)$ and $\phi(x, a)_i = \exp(-\beta\|a-a_i(x)\|)/\sum_{i=1}^{N} \exp(-\beta\|a-a_i(x)\|)$. More generally, whenever there is a basis set of actions such that $f(x, a) = \sum_{i=1}^{N} \alpha_i(x; a)g_f(x, a_i)$ for some functions α_i and g_f, then the assumption holds. Such a basis can be obtained, for instance, by standard approximation techniques such as triangulation in an appropriate metric or other finite element methods in the action space (Zienkiewicz et al., 2005), as long as f is sufficiently smooth in a.

Appendix E. Proof of Proposition 2

We start by reviewing the bound on $\mathrm{br}(\epsilon)$ for the finite dimensional case. In this case, let $\{\lambda_i\}_{i=1}^{d}$ be the eigenvalues of $\Sigma^h(p, x^1)$ in decreasing order, for some distribution $p \in \Delta(\mathcal{F})$ and $x^1 \in \mathcal{X}$, where we recall that $\lambda_d \geq 0$ since $\Sigma^h(p, x^1)$ is a covariance matrix. Then we have

$$\mathrm{trace}((\Sigma^h(p, x^1) + \lambda I)^{-1}\Sigma^h(p, x^1)) = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda + \lambda_i} \leq \sum_{i=1}^{d} 1 \leq d.$$

Proof for geometric decay case. This follows effectively by reducing to the finite dimensional case. For any positive integer n, we have

$$\mathrm{trace}((\Sigma^h(p, x^1) + \lambda I)^{-1}\Sigma^h(p, x^1)) \leq n + \frac{\alpha^n}{\lambda(1-\alpha)}.$$

Choosing $n = \lceil n_0 \rceil$ such that $\frac{\alpha^{n_0}}{(1-\alpha)} = \epsilon^2/2$, we see that

$$\lambda K^h(\lambda) \leq n\lambda + \frac{\epsilon^2}{2},$$

so that we can set $\lambda = \epsilon^2/2n_0$. With these settings, we have

$$K^h(\lambda) \leq 1 + n_0 + \frac{2n_0\alpha^{n_0}}{\epsilon^2(1-\alpha)} = 2n_0 = 1 + 2\frac{\log\frac{2}{\epsilon^2(1-\alpha)}}{\log\frac{1}{\alpha}},$$

where the last equality follows from our setting of n_0.
Proof for polynomial decay case. In this case, the assumption on trace guarantees that

$$\sum_i \lambda_i^q \leq R_q.$$

Now since $q \in (0,1)$, we have

$$\begin{aligned}\mathrm{trace}((\Sigma^h(p,x^1)+\lambda I)^{-1}\Sigma^h(p,x^1)) &\le \sum_i \frac{\lambda_i}{\lambda+\lambda_i}\\ &\le \sum_i \left(\frac{\lambda_i}{\lambda+\lambda_i}\right)^q \qquad (x^q \ge x \text{ for } x<1 \text{ and } q\in(0,1))\\ &\le \sum_i \frac{\lambda_i^q}{\lambda^q} \le \frac{R_q^q}{\lambda^q}.\end{aligned}$$

With this, we have $\lambda K^h(\lambda) \le \lambda^{1-q} R_q^q$, so that we choose

$$\lambda_0 = \left(\epsilon^2 R_q^{-q}\right)^{1/(1-q)}.$$

With this choice, we get

$$K^h(\lambda_0) \le \frac{R_q^q}{\lambda_0^q} = \left(\epsilon^{-2} R_q\right)^{q/(1-q)}.$$

Appendix F. Proof of Corollary 7

Proof [Proof of Corollary 7] The calculations for setting the parameters in this case are a little more tedious. We work under the assumption $\mu > 1$, which is subsequently satisfied. Then optimizing the leading order terms under this assumption yields $\epsilon_1 = \left(\frac{d_q}{\mu B}\right)^{(1-q)/(1+q)}$ and $\epsilon_2 = \left(\frac{\lambda H \mu d}{\eta B}\right)^{(1-q)/(1+q)}$. Now the optimal setting of $\mu = (\eta/(\lambda H))^{(1-q)/(3-q)}$ under our assumption of $\mu \ge 1$. We now set η as in Theorem 5 and optimize to get

$$\lambda = \mathcal{O}\left(T^{-(3-q)(1+q)/4}(\ln N)^{(1-q^2+2q)/2} c_q^{-(3-q)(1+q)/4}\right),$$

where $c_q = d_q^{(1-q)/(1+q)} H^{4/((3-q)(1+q))} B^{2q/(1+q)}$. Substituting this back into Theorem 5 completes the proof. ■

Appendix G. Proofs of decoupling results

In this section, we prove Lemma 11. This is done across two smaller lemmas which provide one level of decoupling using the Bellman rank and another using the linear embedding. The first lemma's proof technique resembles the recently used "one-step back" inequalities in low-rank MDP literature (Agarwal et al., 2020b; Uehara et al., 2021). Throughout this section, we use $u^\top v$ to denote $\langle u, v\rangle$ even for infinite dimensional vectors with a slight abuse of notation to improve readability.

Lemma 14 (Bellman Rank Decoupling) *Consider any distribution p over $\mathcal{F}$. The following inequality holds for any $\epsilon, \mu_1 > 0$.*

$$\begin{aligned}&\mathbb{E}_{f\sim p}\mathbb{E}_{(x^h,a^h)\sim\pi_f|x^1}\mathcal{E}^h(f,f;x^h,a^h)\\ \le&\sqrt{\mathrm{br}^{h-1}(\epsilon)\mathbb{E}_{f'\sim p}\mathbb{E}_{x^h\sim\pi_{f'}|x^1}[\mathbb{E}_{f\sim p}\mathcal{E}^h(f,f;x^h,\pi_f(x^h))^2]}+\epsilon B_1\\ \le&\mu_1\mathbb{E}_{f'\sim p}\mathbb{E}_{x^h\sim\pi_{f'}|x^1}\mathbb{E}_{f\sim p}[\mathcal{E}^h(f,f;x^h,\pi_f(x^h))^2]+\mu_1^{-1}\mathrm{br}^{h-1}(\epsilon)+\epsilon B_1.\end{aligned}$$

Proof We have

$$
\begin{aligned}
&\mathbb{E}_{f\sim p}\mathbb{E}_{(x^h,a^h)\sim\pi_f|x^1}\mathcal{E}^h(f,f;x^h,a^h)\\
=&\mathbb{E}_{f\sim p}\mathbb{E}_{x^h\sim\pi_f|x^1}\mathcal{E}^h(f;x^h,\pi_f(x^h))\\
=&\mathbb{E}_{f\sim p}u^h(f,x^1)^\top\psi^{h-1}(f,x^1) && \text{(Assumption 3)}\\
\leq&\sqrt{K^{h-1}(\lambda)\mathbb{E}_{f\sim p}u^h(f,x^1)^\top(\Sigma^{h-1}(p,x^1)+\lambda I)u^h(f,x^1)}\\
& && (\mathbb{E}[u^Tv]\leq\sqrt{\mathbb{E}[\|u\|_M^2]\ \mathbb{E}[\|v\|_{M^{-1}}^2]}\text{ for psd }M\text{ by Cauchy-Schwarz and Definition 1})\\
\leq&\sqrt{K^{h-1}(\lambda)\mathbb{E}_{f\sim p}u^h(f,x^1)^\top(\mathbb{E}_{f'\sim p}\psi^{h-1}(f',x^1)\psi^{h-1}(f',x^1)^\top)u^h(f,x^1)}\\
&+\sqrt{\lambda K^{h-1}(\lambda)\mathbb{E}_{f\sim p}\|u^h(f,x^1)\|_2^2} && (\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}\text{ for }a,b\geq 0)\\
=&\sqrt{K^{h-1}(\lambda)\mathbb{E}_{f\sim p}\mathbb{E}_{f'\sim p}(u^h(f,x^1)^\top\psi^{h-1}(f',x^1))^2}+\sqrt{\lambda K^{h-1}(\lambda,p)\mathbb{E}_{f\sim p}\|u^h(f,x^1)\|_2^2}\\
\overset{(a)}{=}&\sqrt{K^{h-1}(\lambda)\mathbb{E}_{f\sim p}\mathbb{E}_{f'\sim p}(\mathbb{E}_{(x^h)\sim\pi_{f'}|x^1}\mathcal{E}^h(f,f;x^h,\pi_f(x^h)))^2}+\sqrt{\lambda K^{h-1}(\lambda,p)\mathbb{E}_{f\sim p}\|u^h(f,x^1)\|_2^2}\\
\overset{(b)}{\leq}&\sqrt{K^{h-1}(\lambda)\mathbb{E}_{f\sim p}\mathbb{E}_{f'\sim p}\mathbb{E}_{(x^h)\sim\pi_{f'}|x^1}\left(\mathcal{E}^h(f,f;x^h,\pi_f(x^h)))^2\right)}+\sqrt{\lambda K^{h-1}(\lambda)\mathbb{E}_{f\sim p}\|u^h(x^1,f)\|_2^2}.
\end{aligned}
$$

Here (a) holds by using Assumption 3 once more to rewrite the inner product as a Bellman error, while (b) is a consequence of Jensen's inequality. Now we recall that

$$\|u^h(f,x^1)\|_2\leq B_1.$$

By taking the largest λ so that $\lambda\sup_p K^{h-1}(\lambda,p)\leq\epsilon^2$, we obtain the desired bound. ∎

The next decoupling lemma separates the sampling of the action a^h from the function f whose Bellman error is being evaluated by using Assumption 2, which is crucial for the analysis as explained in Section 7. Note that in this particular derivation, we reduce squared Bellman error to the square root of a decoupled squared Bellman error, which loses rate. It may be possible to improve this reduction by a more careful analysis in future work.

Lemma 15 (Linear Embedding Decoupling) *We have for all $x^h\in\mathcal{X}^h$ and probability measures p on $\mathcal{F}$ and $\mu_2,\epsilon_2>0$:*

$$
\begin{aligned}
\mathbb{E}_{f\sim p}[\mathcal{E}^h(f,f;x^h,\pi_f(x^h))^2]\leq&2\sqrt{\mathrm{dc}^h(\epsilon)\mathbb{E}_{f\sim p,f'\sim p}\mathbb{E}_{a^h\sim\pi_{f'}}[\mathcal{E}^h(f,f;x^h,a^h)^2]}+2\epsilon B_2\\
\leq&2\mu_2\mathbb{E}_{f\sim p,f'\sim p}\mathbb{E}_{a^h\sim\pi_{f'}}\mathcal{E}^h(f,f;x^h,a^h)^2+2\mu_2^{-1}\mathrm{dc}^h(\epsilon)+2\epsilon B_2.
\end{aligned}
$$

Proof We have

$$
\begin{aligned}
&\mathbb{E}_{f\sim p}\left|\mathcal{E}^h(f,f;x^h,\pi_f(x^h))\right|=\mathbb{E}_{f\sim p}\left|w^h(f,x^h)^\top\phi^h(x^h,\pi_f(x^h))\right| && \text{(Assumption 4)}\\
&\leq\left[\mathbb{E}_{f\sim p}w^h(f,x^h)^\top(\tilde{\Sigma}^h+\tilde{\lambda}I)w^h(f,x^h)\right]^{1/2}\left[\mathbb{E}_{f\sim p}\phi^h(x^h,\pi_f(x^h))^\top(\tilde{\Sigma}^h+\tilde{\lambda}I)^{-1}\phi^h(x^h,\pi_f(x^h))\right]^{1/2},\\
& && \text{(Cauchy-Schwarz)}
\end{aligned}
$$

where $\tilde{\Sigma}^h$ is short for $\tilde{\Sigma}^h(p, x^h)$. Therefore we have

$$\begin{aligned}
&\mathbb{E}_{f\sim p}[\mathcal{E}^h(f, f; x^h, \pi_f(x^h))]^2 \\
\leq& 2\, \mathbb{E}_{f\sim p}[|\mathcal{E}^h(f, f; x^h, \pi_f(x^h))|] && (|\mathcal{E}^h(f, f; x^h, \pi_f(x^h))| \leq 2 \text{ since } r^h, f(x, a) \in [0, 1]) \\
\leq& 2 \sqrt{\left[\mathbb{E}_{f\sim p}(w^h(f, x^h)^\top(\tilde{\Sigma}^h + \tilde{\lambda}I)w^h(f, x^h))\right] \left[\mathbb{E}_{f\sim p}(\phi^h(x^h, \pi_f(x^h))^\top(\tilde{\Sigma}^h + \tilde{\lambda}I)^{-1}\phi^h(x^h, \pi_f(x^h)))\right]} \\
\overset{(a)}{\leq}& 2 \sqrt{\left[\mathbb{E}_{f\sim p, f'\sim p}\mathbb{E}_{a^h\sim\pi_{f'}}(w^h(f, x^h)^\top(\phi^h(x^h, a^h)\phi^h(x^h, a^h)^\top + \tilde{\lambda}I)w^h(f, x^h))\right] \tilde{K}^h(\tilde{\lambda})} \\
\overset{(b)}{=}& 2 \sqrt{\left[\mathbb{E}_{f\sim p, f'\sim p}\mathbb{E}_{a^h\sim\pi_{f'}}(\mathcal{E}^h(f, f; x^h, a^h)^2 + \tilde{\lambda}\|w^h(f, x^h)\|_2^2)\right] \tilde{K}^h(\tilde{\lambda})} \\
\leq& 2 \sqrt{\left[\mathbb{E}_{f\sim p, f'\sim p}\mathbb{E}_{a^h\sim\pi_{f'}}(\mathcal{E}^h(f, f; x^h, a^h)^2)\right] \tilde{K}^h_{q_2}(\tilde{\lambda})} + 2\sqrt{\tilde{\lambda}\tilde{K}^h(\tilde{\lambda})\mathbb{E}_{f\sim p}\|w^h(f, x^h)\|_2^2}. \\
& && (\sqrt{a+b} \leq \sqrt{a} + \sqrt{b})
\end{aligned}$$

Here (a) holds due to the definition of $\tilde{K}^h$ (Definition 3). Note that

$$\|w^h(f, x^h)\|_2 \leq B_2.$$

By taking the largest $\tilde{\lambda}$ so that $\tilde{\lambda} \sup_p \tilde{K}^{h-1}(\tilde{\lambda}, p) \leq \epsilon^2$, we obtain the first desired bound. The second bound follows by Cauchy-Schwarz inequality. ■

Proof of Proposition 11. Using the two lemmas above, it is easy to establish Proposition 11.
Proof [Proof of Proposition 11] Since x_t^1 is randomly drawn from $\mathcal{D}$, it is independent of f_t. Therefore we have

$$\begin{aligned}
&\lambda\mathbb{E}\ \text{Regret}(f_t, x_t^1) + \lambda\mathbb{E}\mathbb{E}_{f|S_{t-1}}\ \Delta f^1(x_t^1) = \lambda\mathbb{E}\ \text{Regret}(f_t, x_t^1) + \lambda\mathbb{E}\ \Delta f_t^1(x_t^1) \\
=& \lambda\sum_{h=1}^{H} \mathbb{E}\ \mathbb{E}_{(x^h, a^h)\sim\pi_{f_t}|x_t^1}\mathcal{E}^h(f_t, f_t, x^h, a^h) = \lambda H\mathbb{E}\ \mathcal{E}^{h_t}(f_t, f_t, x_t^{h_t}, a_t^{h_t}) \\
\leq& \lambda\left[\epsilon_1 HB_1 + \mu_1^{-1}\text{br}(\epsilon_1) + H\mu_1\mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\mathcal{E}^{h_t}(f, f, x_t^{h_t}, \pi_f(x_t^{h_t}))^2\right] \\
\leq& \lambda\left[2\mu_1\mu_2 H\mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\mathcal{E}^{h_t}(f, f; x_t^{h_t}, a_t^{h_t})^2 + 2\mu_1\mu_2^{-1}\text{dc}(\epsilon_2) + 2\epsilon_2\mu_1 HB_2 + \epsilon_1 HB_1 + \mu_1^{-1}\text{br}(\epsilon_1)\right].
\end{aligned}$$

The first equality used the independence of x_t^1 and f_t. The second equality used Proposition 10. The third equality used the fact that h_t is uniformly drawn from $[H]$. The first inequality used Lemma 14. The second inequality used Lemma 15. Note that the last two steps crucially use that the function f_t used to draw $x_t^{h_t}$, f_t' to draw $a_t^{h_t}$ and $f \sim p$ being evaluated are all mutually independent. Taking $\mu_1 = \mu$ and $\mu_2 = \eta/(4\lambda\mu_1 H)$, we obtain the desired result. ■

Appendix H. Convergence of the online learning in TS^3

In this section, we study the convergence of the online learning updates for the f and g functions. To do so, it is helpful to define some additional notation.

Define $\sigma = 2$, and let

$$\hat{\epsilon}_t^h(f) = \mathcal{T}^h f(x_t^h, a_t^h) - [r_t^h + f^{h+1}(x_t^{h+1})],$$

be the noise in the Bellman residuals. We observe that $|\hat{\epsilon}_t^h(f)| \leq \sigma$ for all $t \in [T]$, $h \in [H]$ and $f \in \mathcal{F}$ under our normalization assumptions. For convenience, we use σ to denote this upper bound on the $\hat{\epsilon}_t^h(f)$ to avoid carrying constants. We also have the following observations about the noise.

For each $h \geq 1$, we also define:

$$\begin{aligned}
\hat{\delta}_t^h(g, f) &= \hat{\Delta}_t^h(g, f)^2 - \hat{\epsilon}_t^h(f)^2, \\
\hat{\delta}_t^h(f) &= \mathbb{E}_{g \sim p(g|f, S_{t-1})} \hat{\delta}_t^h(g, f), \qquad \text{and} \\
\hat{Z}_t^h(f) &= -\frac{1}{\gamma} \ln \mathbb{E}_{g \sim p(g|f, S_{t-1})} \exp(-\gamma \hat{\delta}_t^h(g, f)).
\end{aligned} \tag{7}$$

Here $\hat{\delta}_t^h(g, f)$ measures how well g captures the Bellman residual of f and $\hat{\delta}_t^h(f)$ measures the quality of our posterior distribution over g in doing so. Finally $\hat{Z}_t^h(f)$ is a log-partition function for the posterior. We further define the log-partition function for the posterior over f:

$$Z(S_t) = -\ln \mathbb{E}_{f \sim p_0} \exp\left(\sum_{s=1}^{t} \lambda \Delta f(x_s^1) - \eta \sum_{s=1}^{t} [\hat{\delta}_s^{h_s}(f, f) - \hat{Z}_s^{h_s}(f)] \right),$$

where

$$\Delta f(x^1) = f(x^1) - Q_\star^1(x^1).$$

We also introduce the following definition

$$\hat{Z}_t = \hat{Z}_t^{h_t} = -\frac{1}{\eta} \ln \mathbb{E}_{f|S_{t-1}} \exp\left(-\eta [\hat{\delta}_t^{h_t}(f, f) - \hat{Z}_t^{h_t}(f)] \right),$$

which is the normalization factor of $p(f|S_t)/p(f|S_{t-1})$.

Let $\psi(z) = (e^z - z - 1)/z^2$. It is known that $\psi(z)$ is an increasing function of z. We organize the rest of this section as follows. We begin with some basic properties of Bellman residuals, then analyze the convergence of outer and inner updates respectively.

H.1. Properties of Bellman residuals

Lemma 16 *For any f that may depend on S_{t-1}:*

$$\mathbb{E}_{x_t^{h+1}, r_t^h | x_t^h, a_t^h, h_t = h} \hat{\epsilon}_t^h(f) = 0,$$

and for any constant b_t independent of $\hat{\epsilon}_t^h(f)$:

$$\mathbb{E}_{x_t^{h+1}, r_t^h | x_t^h, a_t^h, h_t = h} \exp(b_t \hat{\epsilon}_t^h(f)) \leq \exp(b_t^2 \sigma^2 / 2).$$

The first equality follows since the conditional expectation only acts over the MDP rewards and dynamics, which are conditionally independent of any f that even depends on S_{t-1}. The second bound is a consequence of the sub-Gaussian bound for bounded random variables in the proof of Hoeffding's inequality.

Lemma 17 *Given any g, f. We have*

$$-\sigma^2 \leq \hat{\delta}_t^h(g,f) = \mathcal{E}^h(g,f;x_t^h,a_t^h)^2 + 2\hat{\epsilon}_t^h(f)\mathcal{E}^h(g,f;x_t^h,a_t^h) \leq 1+2\sigma.$$

Therefore

$$\max\left(|\hat{\delta}_t^h(f,f) - \hat{Z}_t^h(f)|, |\hat{\delta}_t^h(f,f) - \hat{\delta}_t^h(f)|\right) \leq (1+\sigma)^2.$$

Moreover for any $t \in [T]$, $h \in [H]$, $x \in \mathcal{X}$ and $a \in \mathcal{A}$, we have

$$\mathbb{E}[\hat{\delta}_t^h(g,f) \mid x_t^h, a_t^h] = \mathcal{E}^h(g,f;x_t^h,a_t^h)^2, \quad \mathbb{E}[\hat{\delta}_t^h(g,f)^2 \mid x_t^h, a_t^h] \leq 5\sigma^2\mathcal{E}^h(g,f;x_t^h,a_t^h)^2,$$

and for any $c > 0$,

$$\begin{aligned}
\mathbb{E}[\exp(c\hat{\delta}_t^h(f)) \mid x_t^h, a_t^h] &\leq \exp(c(1+2c\sigma^2)\mathbb{E}_{g|f,S_{t-1}}\mathcal{E}(g,f;x_t^h,a_t^h)^2), \\
\mathbb{E}[\exp(c\hat{\delta}_t^h(g,f) \mid x_t^h, a_t^h] &\leq \exp(c(1+2c\sigma^2)\mathcal{E}(g,f;x_t^h,a_t^h)^2), \\
\mathbb{E}[\exp(-c\hat{\delta}_t^h(g,f) \mid x_t^h, a_t^h] &\leq \exp(-c(1-2c\sigma^2)\mathcal{E}(g,f;x_t^h,a_t^h)^2).
\end{aligned}$$

Proof We note that

$$\hat{\delta}_t^h(g,f) = \mathcal{E}^h(g,f;x_t^h,a_t^h)^2 + 2\hat{\epsilon}_t^h(f)\mathcal{E}^h(g,f;x_t^h,a_t^h).$$

This implies the first result. We note that Lemma 16 implies that

$$\mathcal{E}^h(g,f;x,a) = \mathbb{E}_{r_t^h, x_t^{h+1}|x_t^h=x, a_t^h=a}\hat{\delta}_t^h(g,f).$$

This implies the second result. Moreover, Lemma 16 implies that

$$\begin{aligned}
\mathbb{E}_{r_t^h, x_t^{h+1}|x_t^h=x, a_t^h=a}\hat{\delta}_t^h(g,f)^2 =& \mathcal{E}^h(g,f;x,a)^4 + 4\mathbb{E}_{r_t^h, x_t^{h+1}|x_t^h=x, a_t^h=a}\hat{\epsilon}_t^h(g,f)^2\mathcal{E}^h(g,f;x,a)^2 \\
\leq& (\sigma^2 + 4\sigma^2)\mathcal{E}^h(g,f;x,a)^2,
\end{aligned}$$

where the final inequality uses $|\mathcal{E}^h(g,f;x,a)^2| \leq \sigma$, since $g(x,a) \in [0,1]$ and $r + f(x', \pi_f(x')) \in [0,2]$. This implies the third result. The last three inequalities follow by using the first part of the lemma and then using the sub-Gaussian exponential bound from Lemma 16 with $b_t = c\mathcal{E}(g,f;x_t^h,a_t^h)$ and $b_t = -c\mathcal{E}(g,f;x_t^h,a_t^h)$ respectively for the two bounds. ■

H.2. Convergence of the outer updates

In this section, we build the necessary results to prove Lemma 12. We begin with the following bound for $\hat{Z}_t^h(f)$.

Lemma 18 *We have*

$$\hat{\delta}_t^h(f) - \gamma\psi(\gamma\sigma^2)\mathbb{E}_{g|f,S_{t-1}}\hat{\delta}_t^h(g,f)^2 \leq \hat{Z}_t^h(f) \leq \hat{\delta}_t^h(f).$$

Assume that

$$\alpha' = \exp(\gamma\sigma^2(1-2\gamma\sigma^2))/(1-2\gamma\sigma^2) > 0.$$

Then for any f that may depend on S_{t-1}:

$$\mathbb{E}_{x_t, a_t, r_t|h_t=h}\, \mathbb{E}_{g|f,S_{t-1}}\mathcal{E}^h(g,f;x_t^h,a_t^h)^2 \leq \alpha'\mathbb{E}_{x_t,a_t,r_t|h_t=h}\, \hat{Z}_t^h(f).$$

Proof We have

$$\begin{aligned}-\gamma\hat{Z}_t^h(f) =& \ln\ \mathbb{E}_{g\sim p(g|f,S_{t-1})}\exp(-\gamma\hat{\delta}_t^h(g,f))\\ \leq& \mathbb{E}_{g\sim p(g|f,S_{t-1})}\exp(-\gamma\hat{\delta}_t^h(g,f))-1 && (\ln z\leq z-1)\\ =&\mathbb{E}_{g\sim p(g|f,S_{t-1})}\left[-\gamma\hat{\delta}_t^h(g,f)+\gamma^2\psi(-\gamma\hat{\delta}_t^h(g,f))\hat{\delta}_t^h(g,f)^2\right] && (\psi(z)\text{ is increasing in }z)\\ \leq&\mathbb{E}_{g\sim p(g|f,S_{t-1})}\left[-\gamma\hat{\delta}_t^h(g,f)+\gamma^2\psi(\gamma\sigma^2)\hat{\delta}_t^h(g,f)^2\right]. && (\text{Lemma 17})\end{aligned}$$

This can be simplified to obtain the first half of the first inequality. The second half of the first inequality follows directly from Jensen's inequality.

Now conditioned on $h_t = h$, we have

$$\begin{aligned}&-\gamma\mathbb{E}_{x_t,a_t,r_t}\ \hat{Z}_t^h(f)\\ \leq&\mathbb{E}_{x_t,a_t,r_t}\ln\ \mathbb{E}_{g\sim p(g|f,S_{t-1})}\mathbb{E}_{x_t^{h+1},r_t^h|x_t^h,a_t^h}e^{-\gamma\hat{\delta}_t^h(g,f)} && (\text{Jensen's inequality})\\ \leq&\mathbb{E}_{x_t,a_t,r_t}\ln\ \mathbb{E}_{g\sim p(g|f,S_{t-1})}e^{-\gamma(1-2\gamma\sigma^2)\mathcal{E}^h(g,f,x_t,a_t)^2} && (\text{Lemma 17})\\ \leq&\mathbb{E}_{x_t,a_t,r_t}\ \left[\mathbb{E}_{g\sim p(g|f,S_{t-1})}e^{-\gamma(1-2\gamma\sigma^2)\mathcal{E}^h(g,f,x_t,a_t)^2}-1\right] && (\ln z\leq z-1)\\ \leq&-\mathbb{E}_{x_t,a_t,r_t}\ \mathbb{E}_{g\sim p(g|f,S_{t-1})}e^{-\gamma\sigma^2(1-2\gamma\sigma^2)}\gamma(1-2\gamma\sigma^2)\mathcal{E}^h(g,f,x_t,a_t)^2.\end{aligned}$$

Here the last inequality used $e^{-z}-1\leq -e^{-z'}z$ for $0\leq z\leq z'$. This implies the desired bound. ■

Lemma 19 *We have*

$$\hat{Z}_t\leq\mathbb{E}_{f|S_{t-1}}[\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f)],$$

and

$$|\hat{Z}_t|\leq(1+\sigma)^2.$$

Proof The first inequality follows from Jensen's inequality. The second inequality follows from Lemma 17. ■

We also require a bound on the log-partition function.

Lemma 20 *If $2\gamma\sigma^2<1$, then for all $t\leq T$:*

$$\mathbb{E}\ Z(S_t)\leq\lambda\epsilon T+4\eta T\epsilon^2+\kappa(\epsilon).$$

Proof For any probability distribution p on $\mathcal{F}$, we have

$$\begin{aligned}\mathbb{E}_{S_t}\ Z(S_t)\leq&\mathbb{E}_{f\sim p}\mathbb{E}_{S_t}\left(\sum_{s=1}^t-\lambda\Delta f(x_s^1)+\eta\sum_{s=1}^t[\hat{\delta}_s^{h_s}(f,f)-\hat{Z}_s^{h_s}(f)]\right)+\mathbb{E}_{f\sim p}\ln\frac{p(f)}{p_0(f)}\\ \leq&\mathbb{E}_{f\sim p}\mathbb{E}_{S_t}\left(\sum_{s=1}^t-\lambda\Delta f(x_s^1)+\eta\sum_{s=1}^t\hat{\delta}_s^{h_s}(f,f)\right)+\mathbb{E}_{f\sim p}\ln\frac{p(f)}{p_0(f)}.\end{aligned}$$

The first inequality used the fact that $p(f|S_t)$ is the minimizer of the right hand side over p. The second inequality used the second inequality of Lemma 18 and $2\gamma\sigma^2 < 1$. Using the first half of the third displayed inequality of Lemma 17, we further obtain

$$\mathbb{E}_{S_t} Z(S_t) \le \mathbb{E}_{f\sim p}\left[\mathbb{E}_{S_t}\left(\sum_{s=1}^{t} -\lambda\Delta f(x_s^1) + \eta\sum_{s=1}^{t} \mathcal{E}^{h_s}(f,f,x_s,a_s)^2\right) + \mathbb{E}_{f\sim p}\ln\frac{p(f)}{p_0(f)}\right].$$

We now recall our definition of the set $\mathcal{F}(\epsilon, f)$ for any $f \in \mathcal{F}$ from Definition 4 as the set of all functions which capture the Bellman error of f up to an error ϵ. Then we have $\forall f \in \mathcal{F}(\epsilon, Q_\star) = \prod_h \mathcal{F}(\epsilon, Q_\star^h)$

$$|\Delta f(x_s^1)| \le \epsilon, \qquad \mathcal{E}^h(f,f,x_t,a_t) \le 2\epsilon.$$

We can now take

$$p(f) = \frac{p_0(f) I(f \in \mathcal{F}(\epsilon, Q_\star))}{p_0(\mathcal{F}(\epsilon, Q_\star))}.$$

It implies the desired bound. ■

We are now ready to prove Proposition 12.

Proof of Proposition 12

Proof Let us define $\alpha = 6\eta\sigma^2 < 1$ Lemma 18 implies that

$$\hat{Z}_t^{h_t}(f) \le \hat{\delta}_t^{h_t}(f).$$

Therefore

$$\begin{aligned}
&\mathbb{E}[Z(S_{t-1}) - Z(S_t)] = \mathbb{E}\left[\ln \mathbb{E}_{f|S_{t-1}} \exp\left(\lambda\Delta f(x_t^1) - \eta[\hat{\delta}_t^{h_t}(f,f) - \hat{Z}_t^{h_t}(f)]\right)\right] \\
\le&\mathbb{E}\left[\ln \mathbb{E}_{f|S_{t-1}} \exp\left(\lambda\Delta f(x_t^1) - \eta[\hat{\delta}_t^{h_t}(f,f) - \hat{\delta}_t^{h_t}(f)]\right)\right] \qquad \text{(Lemma 18)} \\
\le&\frac{1}{3}\left[\ln \mathbb{E}\mathbb{E}_{f|S_{t-1}} \exp\left(3\lambda\Delta f(x_t^1)\right) + \ln \mathbb{E}\mathbb{E}_{f|S_{t-1}} \exp\left(-3\eta\hat{\delta}_t^{h_t}(f,f)\right) + \ln \mathbb{E}\mathbb{E}_{f|S_{t-1}} \exp\left(3\hat{\delta}_t^{h_t}(f)\right)\right],
\end{aligned}$$

where the last inequality follows from Jensen's inequality to show that $\mathbb{E}[XYZ] \le \sqrt[3]{\mathbb{E}[X^3]\mathbb{E}[Y^3]\mathbb{E}[Z^3]}$ for non-negative random variables X, Y, Z. Now we further simplify each term by using the last two bounds in Lemma 17, by taking a conditional expectation with respect to x_t^{h+1}, r_t^h, conditioned on x_t^h, a_t^h to obtain

$$\begin{aligned}
&\mathbb{E}[Z(S_{t-1}) - Z(S_t)] \\
\le&\frac{1}{3}\mathbb{E}\mathbb{E}_{f|S_{t-1}}\left(3\lambda\Delta f(x_t^1) + 9\lambda^2/2\right) + \frac{1}{3}\ln \mathbb{E}\mathbb{E}_{f|S_{t-1}} \exp\left(-(3\eta - 18\eta^2\sigma^2)\mathcal{E}^{h_t}(f,f,x_t,a_t)^2\right) \\
&+ \frac{1}{3}\ln \mathbb{E}\mathbb{E}_{f|S_{t-1}} \exp\left((3\eta + 18\eta^2\sigma^2)\mathbb{E}_{g|f,S_{t-1}}\mathcal{E}^{h_t}(g,f,x_t,a_t)^2\right) \\
\le&\frac{1}{3}\mathbb{E}\mathbb{E}_{f|S_{t-1}}\left(3\lambda\Delta f(x_t^1) + 9\lambda^2\right) - \eta(1-\alpha)e^{-3\eta\sigma^2(1-\alpha)}\mathbb{E}\mathbb{E}_{f|S_{t-1}}\mathcal{E}^h(f,f,x_t,a_t)^2 \\
&+ \eta(1+\alpha)e^{3\eta\sigma^2(1+\alpha)}\mathbb{E}\mathbb{E}_{f|S_{t-1}}\mathbb{E}_{g|f,S_{t-1}}\mathcal{E}^h(g,f,x_t,a_t)^2.
\end{aligned}$$

The last inequality used $\alpha = 6\eta\sigma^2 < 1$ and

$$\ln \mathbb{E}_\xi \exp(-z(\xi)) \leq \mathbb{E}_\xi \exp(-z(\xi)) - 1 \leq -\exp(-\max_\xi z(\xi))\mathbb{E}_\xi z(\xi),$$

again using $e^{-z} - 1 \leq -ze^{-z'}$ for $0 \leq z \leq z'$. Similarly, we also get $\ln \mathbb{E}_\xi \exp(z(\xi)) \leq \exp(\max_\xi z(\xi))\mathbb{E}_\xi z(\xi)$ for $z(\xi) \geq 0$. By summing over $t = 1$ to $t = T$, and note that $Z(S_0) = 0$, we obtain the desired bound. ■

H.3. Simplified convergence analysis of the inner updates

In this section, we establish the following result, which is simpler to prove than Proposition 13.

Proposition 21 (Inner loop convergence) *Suppose $\gamma < 1/2$. Then we have for all $\epsilon > 0$:*

$$\frac{1}{2}\mathbb{E}\mathbb{E}_{f,g|S_{t-1}}\mathcal{E}^{h_t}(g, f; x_t^{h_t}, a_t^{h_t})^2 \leq \epsilon T(\epsilon + 3)(6 + 99\eta T) + 219\eta T + \frac{1 + 35\eta T}{\gamma}(\kappa(\epsilon) + \kappa'(\epsilon)).$$

We first show how this immediately yields an $\mathcal{O}(T^{-1/8})$ sample complexity of our algorithm, before proving the statement through a series of lemmas. Though we eventually supercede this analysis with a sharper one, several intermediate results will be reused and we believe that the simpler analysis of this proposition illustrates the main ideas of solving the nested minimax setup.

We now state a form of Theorem 5, using Proposition 21.

Theorem 22 *Under Assumptions 1-4, suppose we run TS^3 (Algorithm 1) with some parameters $\gamma \leq 1/36$ and $\eta \leq 0.01$. Then choosing any $\epsilon \leq 0.6/T$, we have*

$$\mathbb{E}\sum_{t=1}^{T}\text{Regret}(f_t, x_t^1) = \mathcal{O}\left(\frac{\eta}{\lambda}(1 + \eta T)(\kappa(\epsilon) + \kappa'(\epsilon)) + \lambda T + \frac{\kappa(\epsilon)}{\lambda} + \tilde{\epsilon}(\lambda/\eta)T\right),$$

where $\tilde{\epsilon}(\lambda/\eta) = \inf_{\mu>0}\left[8(\lambda/\eta)\text{dc}(\epsilon_2)H\mu^2 + 2\mu H\epsilon_2 B_2 + \mu^{-1}\text{br}(\epsilon_1) + \epsilon_1 HB_1\right]$.

To get this result, we set $\eta < 0.01$ so that $(1 - 6\eta)\exp(-12\eta(1 - 6\eta)) \geq 0.5$ along with the stated values of ϵ and γ. Plugging these into the bounds of Propositions 10, 11, 12 and 21, and simplifying gives the result of Theorem 22. Further assuming the conditions of Corollary 6, we can choose $\mu = \left(d_1\eta/(d_2\lambda H^2)\right)^{1/3}$, $\eta = 1/\sqrt{T}$, $\gamma = 1/36$ and $\lambda = T^{-7/8}(d_1^2 d_2 H^2)^{-1/4}(\ln N)^{3/4}$ to get a sample complexity of $\mathcal{O}\left(H(\ln N)^{1/4}(d_1^2 d_2 H)^{1/4}T^{-1/8}\right)$.

We begin the proof of Proposition 21 with a result that carries out a potential function analysis for the inner updates. We recall our definition of $q_t(g|f) = p(g|f, S_t)$ in line 4 of Algorithm 1, which will be repeatedly used in this section.

Lemma 23 *Assume that $2\gamma\sigma^2 \leq 1$. Let*

$$\tilde{p}(g|f) = \frac{p_0(g)I(g \in \mathcal{F}(\epsilon, f))}{p_0(\mathcal{F}(\epsilon, f))}, \quad \text{and} \quad \hat{H}_t(f) = \mathbb{E}_{g\sim\tilde{p}(\cdot|f)}\ln\frac{\tilde{p}(g|f)}{p(g|f, S_t)}.$$

Then for all t:

$$\mathbb{E}_{S_t}\sup_f \hat{H}_t(f) \leq \ln \mathbb{E}_{S_t}\sup_f \exp(\hat{H}_t(f)) \leq \kappa(\epsilon) + \kappa'(\epsilon) + 4\gamma\epsilon(\epsilon + 1 + \sigma)t.$$

Proof Let

$$A_t(g,f) = \gamma \sum_{s=1}^{t} \hat{\delta}_s^{h_s}(g,f).$$

Then for each $g \in \mathcal{F}(\epsilon, f)$, we have:

$$-2\gamma t\epsilon\sigma \leq A_t(g,f) \leq \gamma(\epsilon^2 + 2\epsilon\sigma)t.$$

Let $f_1, \ldots, f_N$ be a cover of $\mathcal{F}$ so that for any $f \in \mathcal{F}$, $\exists j$ such that $|f^h(x) - f_j^h(x)| \leq \epsilon$ for all x. We know that $\ln N \leq \kappa'(\epsilon)$ (Definition 4). This implies that for all S_t, $\exists j \in [N]$, such that for all g:

$$-A_t(g,f) \leq -A_t(g,f_j) + \gamma(\epsilon^2 + 2\epsilon(1+\sigma))t.$$

It follows that with $p_0(g) = \prod_{h=1}^{H} p_0^h(g^h)$, we obtain

$$\begin{aligned}
&\ln \mathbb{E}_{S_t} \sup_{f\in\mathcal{F}} \mathbb{E}_{g\sim p_0} \exp\left(-A_t^h(g,f)\right)\\
\leq&\gamma\epsilon t(\epsilon+2+2\sigma) + \ln \mathbb{E}_{S_t} \sup_{j} \mathbb{E}_{g\sim p_0} \exp\left(-\gamma \sum_{s=1}^{t} \hat{\delta}_s^{h_s}(g,f_j)\right)\\
\leq&\gamma\epsilon t(\epsilon+2+2\sigma) + \ln \sum_{j} \mathbb{E}_{g\sim p_0} \mathbb{E}_{S_t} \exp\left(-\gamma \sum_{s=1}^{t} \hat{\delta}_s^{h_s}(g,f_j)\right)\\
\overset{(a)}{\leq}&\gamma\epsilon t(\epsilon+2+2\sigma) + \ln \sum_{j} \mathbb{E}_{g\sim p_0} \mathbb{E}_{S_t} \exp\left(-\gamma(1-2\gamma\sigma^2)\mathcal{E}^h(g,f_j,x_t^{h_t},a_t^{h_t})^2 - \gamma \sum_{s=1}^{t-1} \hat{\delta}_s^{h_s}(g,f_j)\right)\\
\leq&\gamma\epsilon t(\epsilon+2+2\sigma) + \ln \sum_{j} \mathbb{E}_{g\sim p_0} \mathbb{E}_{S_t} \exp\left(-\gamma \sum_{s=1}^{t-1} \hat{\delta}_s^{h_s}(g,f_j)\right)\\
\leq&\cdots\\
\leq&\gamma\epsilon t(\epsilon+2+2\sigma) + \ln N.
\end{aligned}$$

The first inequality used covering property. Inequality (a) uses the last bound in Lemma 17. We thus have

$$\begin{aligned}
&\mathbb{E}_{S_t} \sup_{f\in\mathcal{F}} \hat{H}_t(f) \leq \ln \mathbb{E}_{S_t} \sup_{f\in\mathcal{F}} \exp(\hat{H}_t(f))\\
=&\ln \mathbb{E}_{S_t} \sup_{f\in\mathcal{F}} \exp\left(\mathbb{E}_{g\sim\tilde{p}(\cdot|f)} \ln \frac{\tilde{p}(g|f)}{p(g|f,S_t)}\right)\\
=&\ln \mathbb{E}_{S_t} \sup_{f\in\mathcal{F}} \exp\left[\mathbb{E}_{g\sim\tilde{p}(\cdot|f)} \ln \frac{p_0(g)}{p_0(g)\exp(-A_t(g,f))} + \ln \frac{1}{p_0(\mathcal{F}(\epsilon,f))} + \ln \mathbb{E}_{g\sim p_0} \exp\left(-A_t(g,f)\right)\right]\\
\leq&\gamma(\epsilon^2+2\epsilon\sigma)t + \kappa(\epsilon) + 2\gamma\epsilon t(\epsilon+2+2\sigma) + \kappa'(\epsilon).
\end{aligned}$$

The first inequality used Jensen's inequality. This implies the result. ■

We give an upper bound on the log-partition function $\hat{Z}_t^h(f)$. Note that this is the part of our analysis which relies on η being smaller than γ, and leads to a loss in rates. It is possible to sharpen this analysis through a more careful self-bounding argument, which we carry out in the next section. For now, we continue with a simpler argument.

Lemma 24 *Let*

$$\alpha'' = 2(1+\sigma)^2 \exp(2\eta(1+\sigma)^2).$$

Then

$$\mathbb{E}\,\sum_{t=1}^{T} \mathbb{E}_{f|S_{t-1}} \hat{Z}_t^{h_t}(f) \leq \epsilon T(\epsilon+1+\sigma)(6+4\eta\alpha'' T) + \eta T(1+\sigma)^2\alpha'' + (\gamma^{-1} + \eta\alpha'' T/\gamma)(\kappa(\epsilon)+\kappa'(\epsilon)).$$

Proof We know that

$$p(f|S_t) = p(f|S_{t-1}) \exp(-\eta[(\hat{\delta}_t^{h_t}(f,f) - \hat{Z}_t^{h_t}(f)) - \hat{Z}_t]).$$

Since

$$|\hat{\delta}_t^{h}(f,f) - \hat{Z}_t^{h}(f)) - \hat{Z}_t| \leq 2(1+\sigma)^2,$$

by Lemma 17, we obtain by using $|\exp(z)-1| \leq |z|\exp(|z|)$ with $z = \hat{\delta}_t^h(f,f) - \hat{Z}_t^h(f)) - \hat{Z}_t$

$$\left| \exp(-\eta[(\hat{\delta}_t^{h_t}(f,f) - \hat{Z}_t^{h_t}(f)) - \hat{Z}_t]) - 1 \right| \leq \eta\alpha''. \tag{8}$$

Let $p_t = p(f|S_t)$. We have

$$\begin{aligned}
&\mathbb{E}_{f\sim p_t}\hat{H}_t(f) - \mathbb{E}_{f\sim p_{t-1}}\hat{H}_{t-1}(f)\\
=&\mathbb{E}_{f\sim p_t - p_{t-1}}\mathbb{E}_{g\sim\tilde{p}(\cdot|f)} \ln\frac{\tilde{p}(g|f)}{p(g|f,S_{t-1})} - \mathbb{E}_{f\sim p_t - p_{t-1}}\mathbb{E}_{g\sim\tilde{p}(\cdot|f)} \ln\frac{p(g|f,S_t)}{p(g|f,S_{t-1})}\\
&- \mathbb{E}_{f\sim p_{t-1}}\mathbb{E}_{g\sim\tilde{p}(\cdot|f)} \ln\frac{p(g|f,S_t)}{p(g|f,S_{t-1})}.
\end{aligned}$$

Now we observe that

$$p(g|f,S_t) = p(g|f,S_{t-1})\exp(-\gamma(\hat{\delta}_t^{h_t}(g,f) - \hat{Z}_t^{h_t}(f))),$$

which allows us to further rewrite

$$\begin{aligned}
&\mathbb{E}_{f\sim p_t}\hat{H}_t(f) - \mathbb{E}_{f\sim p_{t-1}}\hat{H}_{t-1}(f)\\
=&\mathbb{E}_{f\sim p_{t-1}}\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f) - \hat{Z}_t^{h_t}(f)) - \hat{Z}_t^{h_t}]} - 1\right]\mathbb{E}_{g\sim\tilde{p}(\cdot|f)} \ln\frac{\tilde{p}(g|f)}{p(g|f,S_{t-1})}\\
&- \gamma\mathbb{E}_{f\sim p_{t-1}}\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f) - \hat{Z}_t^{h_t}(f)) - \hat{Z}_t^{h_t}]} - 1\right]\mathbb{E}_{g\sim\tilde{p}(\cdot|f)}[-\hat{\delta}_t^{h_t}(g,f) + \hat{Z}_t^{h_t}(f)]\\
&- \gamma\mathbb{E}_{f\sim p_{t-1}}\mathbb{E}_{g\sim\tilde{p}(\cdot|f)}[-\hat{\delta}_t^{h_t}(g,f) + \hat{Z}_t^{h_t}(f)]\\
\leq&\eta\alpha''\mathbb{E}_{f\sim p_{t-1}}\hat{H}_{t-1}(f) + \eta\gamma(1+\sigma)^2\alpha'' - \gamma\mathbb{E}_{f|S_{t-1}}\mathbb{E}_{g\sim\tilde{p}(\cdot|f)}[-\hat{\delta}_t^{h_t}(g,f) + \hat{Z}_t^{h_t}(f)].
\end{aligned}$$

The last inequality used $|\hat{\delta}_t^{h_t}(g,f) + \hat{Z}_t^{h_t}(f)| \leq (1+\sigma)^2$, along with our earlier inequality (8) and the observation that $\mathbb{E}_{g\sim\tilde{p}(\cdot|f)} \ln\frac{\tilde{p}(g|f)}{p(g|f,S_{t-1})}$ is a KL divergence, and hence non-negative. By rearranging the terms, we obtain

$$\begin{aligned}
\mathbb{E}_{f\sim p_{t-1}}\hat{Z}_t^{h_t}(f) \leq&\mathbb{E}_{f\sim p_{t-1}}\mathbb{E}_{g\sim\tilde{p}(\cdot|f)}\hat{\delta}_t^{h_t}(g,f) + \eta(1+\sigma)^2\alpha''\\
&+ \frac{1}{\gamma}[(1+\eta\alpha'')\mathbb{E}_{f\sim p_{t-1}}\hat{H}_{t-1}(f) - \mathbb{E}_{f\sim p_t}\hat{H}_t(f)].
\end{aligned}$$

Note that for $g \in \mathcal{F}(\epsilon, f)$, we have $\hat{\delta}_t^{h_t}(g, f) \leq \epsilon^2 + 2\epsilon\sigma$. By summing over t, we obtain

$$\mathbb{E} \sum_{t=1}^{T} \mathbb{E}_{f|S_{t-1}} \hat{Z}_t^{h_t}(f) \leq \epsilon(\epsilon + 2\sigma)T + \eta T(1+\sigma)^2\alpha'' + \frac{1}{\gamma}(1 + \eta\alpha'' T)\mathbb{E} \sup_{t \leq T, f \in \mathcal{F}} \hat{H}_t(f).$$

We can now obtain the desired bound using Lemma 23. ■

We are now ready to prove Proposition 21.

Proof of Proposition 21

Proof We have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\mathbb{E}_{f,g|S_{t-1}} \mathcal{E}^{h_t}(g, f; x_t^{h_t}, a_t^{h_t})^2 \\
\leq &\alpha' \sum_{t=1}^{T} \mathbb{E}\mathbb{E}_{f|S_{t-1}} \hat{Z}_t^{h_t}(f) \\
\leq &\alpha' \left[\epsilon T(\epsilon + 1 + \sigma)(6 + 4\eta\alpha'' T) + \eta T(1+\sigma)^2\alpha'' + (\gamma^{-1} + \eta\alpha'' T/\gamma)(\kappa(\epsilon) + \kappa'(\epsilon))\right] \\
\leq &2 \left[\epsilon T(\epsilon + 3)(6 + 99\eta T) + 219\eta T + (1/\gamma)(1 + 35\eta T)(\kappa(\epsilon) + \kappa'(\epsilon))\right]
\end{aligned}$$

The first inequality used the last inequality of Lemma 18. The second inequality used Lemma 24. The last inequality used our assumptions on the various parameters, which imply that $\alpha < 0.25$, $\alpha' < 2$, and $\alpha'' < 2.7(1+\sigma)^2$. ■

H.4. Proof of Proposition 13

In the following, we derive a refinement of Lemma 24, which allows us to prove Proposition 13. In order to avoid complex constant calculations, we will use the $\mathcal{O}(\cdot)$ notation that hides absolute constants. Here we take $\sigma = \mathcal{O}(1), \eta = \mathcal{O}(1), \gamma = \mathcal{O}(1)$, and $\epsilon = \mathcal{O}(1)$. Consequently, we also have that $\alpha = \mathcal{O}(1)$ and $\alpha' = \mathcal{O}(\gamma) = \mathcal{O}(1)$. We also repeatedly use that for $b = \mathcal{O}(1), \exp(b) \leq 1 + \theta b$ for $\theta \leq b$ by the intermediate value theorem, so that $\exp(b) - 1 = \mathcal{O}(b)$ when $b = \mathcal{O}(1)$. In particular, for the function $\psi(z)$ defined at the start of this section, we have

$$z^2\psi(z) = e^z - z - 1 = \mathcal{O}(z^2), \quad \text{when } z = \mathcal{O}(1). \tag{9}$$

We have the following high probability bound for the entropy considered in Lemma 23.

Lemma 25 *In the setting of Lemma 23, for each t, event A_t, defined below, holds with probability at least $1 - 1/T^2$ over S_t:*

$$A_t = \left\{ \sup_{f \in \mathcal{F}} \hat{H}_t(f) \leq \kappa(\epsilon) + \kappa'(\epsilon) + 4\gamma\epsilon(\epsilon + 1 + \sigma)t + 2\ln T \right\}.$$

Proof From Lemma 23, we obtain for each $t \le T$:

$$\ln \mathbb{E}_{S_t} \sup_{f \in \mathcal{F}} \exp(\hat{H}_t(f)) \le \kappa(\epsilon) + \kappa'(\epsilon) + 4\gamma\epsilon(\epsilon + 1 + \sigma)t.$$

Using Markov's inequality, we obtain for each $t \le T$, with probability $1 - 1/T^2$,

$$\begin{aligned} \sup_f \exp(\hat{H}_t(f)) \le & \exp\left(\kappa(\epsilon) + \kappa'(\epsilon) + 4\gamma\epsilon(\epsilon + 1 + \sigma)t\right) T^2 \\ \le & \exp\left(\kappa(\epsilon) + \kappa'(\epsilon) + 4\gamma\epsilon(\epsilon + 1 + \sigma)t + 2\ln T\right). \end{aligned}$$

This leads to the bound. ■

We also have the following uniform bound for the entropy considered in Lemma 25.

Lemma 26 *Under the Assumption of Lemma 23, for all $t \le T$ and f:*

$$\hat{H}_t(f) = \mathcal{O}(\kappa(\epsilon) + T).$$

Proof We note that in the proof of Lemma 23, we can simply bound

$$|A_t(g, f)| = \mathcal{O}(T).$$

We thus have

$$\begin{aligned} \hat{H}_t(f) =& \mathbb{E}_{g \sim \tilde{p}(\cdot|f)} \ln \frac{\tilde{p}(g|f)}{p(g|f, S_t)} \\ =& \mathbb{E}_{g \sim \tilde{p}(\cdot|f)} \ln \frac{p_0(g)}{p_0(g)\exp(-A_t(g, f))} + \ln \frac{1}{p_0(\mathcal{F}(\epsilon, f))} + \ln \mathbb{E}_{g \sim p_0} \exp\left(-A_t(g, f)\right) \\ =& \kappa(\epsilon) + \mathcal{O}(T). \end{aligned}$$

This implies the result. ■

Lemma 27 *We have*

$$\mathbb{E}_{x_t, a_t, r_t | h_t = h} |\hat{Z}_t^h(f)|^2 = \mathcal{O}(\mathbb{E}_{x_t, a_t, r_t | h_t = h} \hat{Z}_t^h(f)).$$

Proof From Lemma 18, along with (9), we obtain

$$\begin{aligned} |\hat{Z}_t^h(f)| =& \mathcal{O}(|\mathbb{E}_{g|f, S_{t-1}} \hat{\delta}_t^h(g, f)| + \gamma \mathbb{E}_{g|f, S_{t-1}} \hat{\delta}_t^h(g, f)^2) \\ =& \mathcal{O}(\mathbb{E}_{g|f, S_{t-1}} |\hat{\delta}_t^h(g, f)|). && \text{(since } \gamma = \mathcal{O}(1) \text{ by assumption)} \end{aligned}$$

It follows that

$$\begin{aligned} \mathbb{E}_{x_t, a_t, r_t | h_t = h} |\hat{Z}_t^h(f)|^2 =& \mathbb{E}_{x_t, a_t, r_t | h_t = h} \mathcal{O}(\mathbb{E}_{g|f, S_{t-1}} |\hat{\delta}_t^h(g, f)|^2) \\ =& \mathbb{E}_{x_t, a_t, r_t | h_t = h} \mathcal{O}(\mathbb{E}_{g|f, S_{t-1}} \mathcal{E}^h(g, f; x_t^h, a_t^h)^2) && \text{(Lemma 17)} \\ =& \mathbb{E}_{x_t, a_t, r_t | h_t = h} \mathcal{O}(\mathbb{E}_{g|f, S_{t-1}} \hat{Z}_t^h(f))). && \text{(Lemma 18)} \end{aligned}$$

This proves the desired result. ■

Lemma 28 *We have*

$$\mathbb{E}_{x_t,a_t,r_t|h_t=h}|\hat{Z}_t^{h_t}|^2 =\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}\mathcal{O}(|\mathcal{E}^h(f,f;x_t^h,a_t^h)|^2+\hat{Z}_t^h(f)).$$

Proof We have from the definition of $\hat{Z}_t^h$:

$$|\hat{Z}_t^{h_t}| \leq\mathbb{E}_{f|S_{t-1}}\mathcal{O}(|\hat{\delta}_t^{h_t}(f,f)|+|\hat{Z}_t^{h_t}(f)|),$$

where we used $\ln z \leq z-1$ and then conventions above (9) to simplify $\exp(\cdot)-1$.

Therefore we obtain

$$\begin{aligned}\mathbb{E}_{x_t,a_t,r_t|h_t=h}|\hat{Z}_t^{h_t}|^2 =&\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}\mathcal{O}(|\hat{\delta}_t^{h_t}(f,f)|^2+|\hat{Z}_t^{h_t}(f)|^2)\\ =&\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}O(|\mathcal{E}^h(f,f;x_t^h,a_t^h)|^2+\hat{Z}_t^h(f)),\end{aligned}$$

where the second inequality used the second half of the last displayed equation of Lemma 17, and the result of Lemma 27. ■

Lemma 29 *There exists an absolute constant c such that when $\eta \leq c$, then for all $f \in \mathcal{F}$:*

$$\begin{aligned}&\mathbb{E}_{x_t,a_t,r_t|h_t=h}\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]}-1\right]\\ \leq&\eta\mathcal{O}\left(\mathbb{E}_{x_t,a_t,r_t|h_t=h}\hat{Z}_t^h(f)+\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f'|S_{t-1}}\mathcal{E}^h(f',f',x_t^h,a_t^h)^2\right).\end{aligned}$$

Proof We have

$$\begin{aligned}&\mathbb{E}_{x_t,a_t,r_t|h_t=h}\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]}-1\right]\\ \leq&\mathbb{E}_{x_t,a_t,r_t|h_t=h}\left[e^{-\eta\hat{\delta}_t^{h_t}(f,f)+\eta\hat{\delta}_t^{h_t}(f)+\eta\mathbb{E}_{f'|S_{t-1}}[\hat{\delta}_t^{h_t}(f',f')-\hat{Z}_t^{h_t}(f')]}-1\right] &&\text{(Lemmas 18 and 19)}\\ \leq&0.25\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f'|S_{t-1}}\left[e^{-4\eta\hat{\delta}_t^{h_t}(f,f)}+e^{4\eta\hat{\delta}_t^{h_t}(f)}+e^{4\eta\hat{\delta}_t^{h_t}(f',f')}+e^{-4\eta\hat{Z}_t^{h_t}(f')}-4\right].\\ &&&\text{(Cauchy-Schwarz)}\end{aligned}$$

We now bound each of the four terms in turn. Note that conditioned on $h_t=h$, we have

$$\mathbb{E}_{x_t,a_t,r_t}e^{-4\eta\hat{\delta}_t^h(f,f)}-1 \leq \mathbb{E}_{x_t,a_t,r_t}\exp\left(-4\eta(1-2\eta\sigma^2)\mathcal{E}^h(f,f,x_t^h,a_t^h)^2\right)-1 \leq 0, \quad \text{(Lemma 17)}$$

and

$$\begin{aligned}\mathbb{E}_{x_t,a_t,r_t}e^{4\eta\hat{\delta}_t^h(f)}-1 \leq&\mathbb{E}_{x_t,a_t,r_t}\exp\left(4\eta(1+2\eta\sigma^2)\mathbb{E}_{g|f,S_{t-1}}\mathcal{E}^h(g,f,x_t^h,a_t^h)^2\right)-1 &&\text{(Lemma 17)}\\ \leq&\eta\mathcal{O}\left(\mathbb{E}_{x_t,a_t,r_t}\mathbb{E}_{g|f,S_{t-1}}\mathcal{E}^h(g,f,x_t^h,a_t^h)^2\right)\\ \leq&\eta\mathcal{O}\left(\hat{Z}_t^h(f)\right). &&\text{(Lemma 18)}\end{aligned}$$

Here the second inequality follows from the intermediate value theorem, since $c = 4\eta(1+2\eta\sigma^2)\mathbb{E}_{g|f,S_{t-1}}\mathcal{E}^h(g,f,x_t^h,a_t^h)^2 = \mathcal{O}(1)$, so that $e^c \leq 1+\mathcal{O}(c)$. In the last inequality, we use $\alpha' = \mathcal{O}(1)$ in Lemma 18, since $\gamma = \mathcal{O}(1)$.

$$\begin{aligned}
\mathbb{E}_{x_t,a_t,r_t} e^{4\eta\hat{\delta}_t^h(f',f')} - 1 \leq & \mathbb{E}_{x_t,a_t,r_t} \exp\left(4\eta(1+2\eta\sigma^2)\mathcal{E}^h(f',f',x_t^h,a_t^h)^2\right) - 1 && \text{(Lemma 17)}\\
\leq & \eta\mathcal{O}\left(\mathbb{E}_{x_t,a_t,r_t}\mathcal{E}^h(f',f',x_t^h,a_t^h)^2\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{x_t,a_t,r_t} e^{-4\eta\hat{Z}_t^h(f')} - 1 = & \mathbb{E}_{x_t,a_t,r_t}\left[\psi(-4\eta\hat{Z}_t^h(f'))(4\eta\hat{Z}_t^h(f'))^2 - 4\eta\hat{Z}_t^h(f')\right]\\
\leq & \mathbb{E}_{x_t,a_t,r_t}\left(\mathcal{O}(\eta^2\hat{Z}_t^h(f')^2) - 4\eta\hat{Z}_t^h(f')\right) && (\psi(z) = \mathcal{O}(1))\\
\leq & \mathbb{E}_{x_t,a_t,r_t}\left(\mathcal{O}(\eta^2\hat{Z}_t^h(f')) - 4\eta\hat{Z}_t^h(f')\right) && \text{(Lemma 27)}\\
\leq & 0,
\end{aligned}$$

where the last inequality assumed that we choose η small enough so that $\mathcal{O}(\eta^2) \leq 4\eta$, and $\mathbb{E}_{x_t,a_t,r_t}\hat{Z}_t^h(f') \geq 0$, which follows from Lemma 18. ■

Lemma 30 *For each $t \leq T$, we have*

$$\begin{aligned}
& \mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]} - 1\right]\hat{H}_{t-1}(f)\\
= & \eta\mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\ \mathcal{O}\left((\hat{Z}_t^{h_t}(f) + \mathcal{E}^{h_t}(f,f,x_t^{h_t},a_t^{h_t})^2)(\kappa(\epsilon)+\kappa'(\epsilon)+\ln T)\right) + \mathcal{O}(\eta(\kappa(\epsilon)+1)/T).
\end{aligned}$$

Proof Let the indicator $I(A_t)$ denotes that the event of Lemma 25 holds. We have

$$\begin{aligned}
& \mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]} - 1\right]\hat{H}_{t-1}(f)\\
= & \mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\left[\mathbb{E}_{x_t,a_t,r_t}\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]} - 1\right]\right]\hat{H}_{t-1}(f)\\
& && \text{(Since } x_t, a_t, r_t \text{ are independent of } S_{t-1})\\
= & \eta\mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\mathcal{O}\left(\mathbb{E}_{x_t,a_t,r_t}\hat{Z}_t^{h_t}(f) + \mathbb{E}_{x_t,a_t,r_t}\mathbb{E}_{f'|S_{t-1}}\mathcal{E}^{h_t}(f',f',x_t^{h_t},a_t^{h_t})^2\right)\hat{H}_{t-1}(f) && \text{(Lemma 29)}\\
= & \eta\mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\mathcal{O}\left(\mathbb{E}_{x_t,a_t,r_t}\hat{Z}_t^{h_t}(f) + \mathbb{E}_{x_t,a_t,r_t}\mathbb{E}_{f'|S_{t-1}}\mathcal{E}^{h_t}(f',f',x_t^{h_t},a_t^{h_t})^2\right)\hat{H}_{t-1}(f)I(A_t)\\
& + \eta\mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\mathcal{O}\left(\mathbb{E}_{x_t,a_t,r_t}\hat{Z}_t^{h_t}(f) + \mathbb{E}_{x_t,a_t,r_t}\mathbb{E}_{f'|S_{t-1}}\mathcal{E}^{h_t}(f',f',x_t^{h_t},a_t^{h_t})^2\right)\hat{H}_{t-1}(f)(1-I(A_t))\\
= & \eta\mathbb{E}\ \mathcal{O}\left(\mathbb{E}_{f|S_{t-1}}\hat{Z}_t^{h_t}(f) + \mathbb{E}_{f'|S_{t-1}}\mathcal{E}^{h_t}(f',f',x_t^{h_t},a_t^{h_t})^2\right)\mathcal{O}(\kappa(\epsilon)+\kappa'(\epsilon)+\ln T)\\
& && \text{(Definition of } A_t \text{ in Lemma 25 and } \mathbb{E}_{x_t,a_t,r_t}[\hat{Z}_t^{h_t}(f)] \geq 0)\\
& + \eta\mathcal{O}(\kappa(\epsilon)+T)\mathbb{E}(1-I(A_t)) && \text{(Lemma 26)}\\
= & \eta\mathbb{E}\ \mathcal{O}\left(\mathbb{E}_{f|S_{t-1}}\hat{Z}_t^{h_t}(f) + \mathbb{E}_{f'|S_{t-1}}\mathcal{E}^{h_t}(f',f',x_t^{h_t},a_t^{h_t})^2\right)\mathcal{O}(\kappa(\epsilon)+\kappa'(\epsilon)+\ln T)\\
& + \eta\mathcal{O}((\kappa(\epsilon)+T)/T^2). && \text{(Probability of } A_t \text{ in Lemma 25)}
\end{aligned}$$

This implies the desired bound. ■

Lemma 31 *We have*

$$\begin{aligned}&\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}\left|\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]}-1\right]\mathbb{E}_{g\sim\tilde{p}(\cdot|f)}[-\hat{\delta}_t^{h_t}(g,f)+\hat{Z}_t^{h_t}(f)]\right|\\=&\eta\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}\mathcal{O}\left(\epsilon+|\mathcal{E}^{h_t}(f,f;x_t^h,a_t^h)|^2+\hat{Z}_t^h(f)\right).\end{aligned}$$

Proof We have

$$|\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}|=\mathcal{O}(1),$$

and $\forall g\in\mathcal{F}(\epsilon,f)$, we have $\hat{\delta}_t^{h_t}(g,f)=\mathcal{O}(\epsilon)$. Therefore by the definition of $\tilde{p}(\cdot|f)$ in Lemma 23, we have

$$\left|[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]}-1]\mathbb{E}_{g\sim\tilde{p}(\cdot|f)}\hat{\delta}_t^{h_t}(g,f)\right|=\eta\mathcal{O}(\epsilon).$$

Moreover,

$$\begin{aligned}&\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}\left|[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]}-1]\hat{Z}_t^{h_t}(f)\right|\\=&\eta\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}O\left([|\hat{\delta}_t^{h_t}(f,f)|+|\hat{Z}_t^{h_t}(f))|+|\hat{Z}_t^{h_t}|]|\hat{Z}_t^{h_t}(f)|\right)\\=&\eta\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}O\left(|\hat{\delta}_t^{h_t}(f,f)|^2+|\hat{Z}_t^{h_t}(f))|^2+|\hat{Z}_t^{h_t}|^2\right) && \text{(Cauchy-Schwarz)}\\=&\eta\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}O\left(|\mathcal{E}_t^h(f,f;x_t^h,a_t^h)|^2+|\hat{Z}_t^h(f))|^2+|\hat{Z}_t^h|^2\right) && \text{(Lemma 17)}\\=&\eta\mathbb{E}_{x_t,a_t,r_t|h_t=h}\mathbb{E}_{f|S_{t-1}}O\left(|\mathcal{E}_t^h(f,f;x_t^h,a_t^h)|^2+\hat{Z}_t^h(f))\right)\end{aligned}$$

where the second to the last equation was obtained by taking conditional expectation conditioned on (x_t^h,a_t^h) with respect to the transition and reward, followed by Lemma 17. The last equation used Lemma 27 and Lemma 28. ■

We are now ready to prove Proposition 13.

Proof of Proposition 13

Proof From the proof of Lemma 24, we obtain

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}_{f\sim p_t}\hat{H}_t(f) - \mathbb{E}_{f\sim p_{t-1}}\hat{H}_{t-1}(f)] \\
=&\mathbb{E}\ \mathbb{E}_{f\sim p_{t-1}}\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]}-1\right]\hat{H}_{t-1}(f) \\
&-\gamma\mathbb{E}\ \mathbb{E}_{f\sim p_{t-1}}\left[e^{-\eta[(\hat{\delta}_t^{h_t}(f,f)-\hat{Z}_t^{h_t}(f))-\hat{Z}_t^{h_t}]}-1\right]\mathbb{E}_{g\sim\tilde{p}(\cdot|f)}[-\hat{\delta}_t^{h_t}(g,f)+\hat{Z}_t^{h_t}(f)] \\
&-\gamma\mathbb{E}\ \mathbb{E}_{f\sim p_{t-1}}\mathbb{E}_{g\sim\tilde{p}(\cdot|f)}[-\hat{\delta}_t^{h_t}(g,f)+\hat{Z}_t^{h_t}(f)] \\
\leq&\eta c_0\mathbb{E}\ \mathbb{E}_{f\sim p_{t-1}}\left(|\mathcal{E}^{h_t}(f,f;x_t^{h_t},a_t^{h_t})|^2+\hat{Z}_t^{h_t}(f)\right)(\kappa(\epsilon)+\kappa'(\epsilon)+\ln T)+\eta c_0(\kappa(\epsilon)+1)/T \\
&&\text{(Lemma 30)} \\
&+\gamma\eta c_0\mathbb{E}\mathbb{E}_{f\sim p_{t-1}}\left(|\mathcal{E}^{h_t}(f,f;x_t^{h_t},a_t^{h_t})|^2+\hat{Z}_t^{h_t}(f)\right) &\text{(Lemma 31)} \\
&-\gamma\mathbb{E}\ \mathbb{E}_{f\sim p_{t-1}}\hat{Z}_t^{h_t}(f)+\mathcal{O}(\epsilon) &(\forall g\in\mathcal{F}(\epsilon,f),\text{ we have }\hat{\delta}_t^{h_t}(g,f)=\mathcal{O}(\epsilon)) \\
\leq&c_1\gamma\mathbb{E}\ \mathbb{E}_{f\sim p_{t-1}}|\mathcal{E}^{h_t}(f,f;x_t^{h_t},a_t^{h_t})|^2-0.5\gamma\mathbb{E}\ \mathbb{E}_{f\sim p_{t-1}}\hat{Z}_t^{h_t}(f)+\mathcal{O}(\epsilon+\eta(\kappa(\epsilon)+1)/T).
\end{aligned}$$

The last inequality used the condition on η in the statement of the proposition. Now by summing over $t = 1$ to $t = T$, and applying Lemma 23 to bound $\mathbb{E}\hat{H}_0(f)$, we obtain

$$\gamma\mathbb{E}\ \mathbb{E}_{f\sim p_{t-1}}\hat{Z}_t^{h_t}(f) \leq \mathcal{O}(\epsilon T+\kappa(\epsilon)+\kappa'(\epsilon)+1)+2c_1\gamma\sum_{t=1}^{T}\mathbb{E}\ |\mathcal{E}^{h_t}(f,f;x_t^{h_t},a_t^{h_t})|^2.$$

The proof is now completed by recalling Lemma 18. ■

Appendix I. Proof of Theorem 9

In this section, we provide a proof for Theorem 9. We focus on the differences from the proof of Theorem 5, which are limited to our decoupling analysis. In particular, in the setting of Theorem 9, we have the following improved analog for Lemma 15.

Lemma 32 (Design based decoupling) *Under conditions of Theorem 9, we have for all $x^h \in \mathcal{X}^h$ and $f \in \mathcal{F}$:*

$$\mathcal{E}^h(f,f;x^h,\pi_f(x^h))^2 \leq d_2\mathcal{E}^h(f,f,x^h,\rho^h(x^h))^2,$$

where $\rho^h(x^h)$ is the G-optimal design (6) at the state x^h.

Proof Let $\Sigma_\star(x^h)$ be the covariance $\mathbb{E}_{a^h\sim\rho(x^h)}[\phi^h(x^h,a^h)\phi^h(x^h,a^h)^\top]$. The definition of G-optimal design implies that

$$\sup_{a^h\in\mathcal{A}}\|\phi^h(x^h,a^h)\|^2_{\Sigma_\star(x^h)^{-1}} \leq d_2. \tag{10}$$

By Assumption 4, we have

$$\begin{aligned}
\mathcal{E}^h(f,f;x^h,\pi_f(x^h))^2 =& \left\langle w^h(f,x^h), \phi^h(x^h,a)\right\rangle^2 \\
\le& \|w^h(f,x^h)\|^2_{\Sigma_\star(x^h)} \|\phi^h(x^h,a^h)\|^2_{\Sigma_\star(x^h)^{-1}} \\
\le& d_2 \mathbb{E}_{a^h\sim\rho^h(x^h)} \left\langle w^h(f,x^h), \phi^h(x^h,a^h)\right\rangle^2 \\
=& d_2 \mathcal{E}^h(f,f,x^h,\rho^h(x^h))^2.
\end{aligned}$$

The first inequality used Cauchy-Schwarz. The second inquality used (10). ■

Combining this with the Bellman rank decoupling in Lemma 14 and Proposition 10, we get

$$\begin{aligned}
&\lambda\mathbb{E}\ \mathrm{Regret}(f_t,x_t^1) + \lambda\mathbb{E}\mathbb{E}_{f|S_{t-1}}\ \Delta f^1(x_t^1) = \lambda H\mathbb{E}\ \mathcal{E}^{h_t}(f_t,f_t,x_t^{h_t},a_t^{h_t}) \\
\le&\lambda\left[\epsilon_1 HB_1 + \mu_1^{-1}\mathrm{br}(\epsilon_1) + H\mu_1\mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\mathcal{E}^{h_t}(f,f,x_t^{h_t},\pi_f(x_t^{h_t}))^2\right] \\
\le&\lambda\left[\epsilon_1 HB_1 + \mu_1^{-1}\mathrm{br}(\epsilon_1) + H\mu_1 d_2\mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\mathcal{E}^{h_t}(f,f,x_t^{h_t},\rho^{h_t}(x_t^{h_t}))^2\right] \\
=&\lambda\left[\epsilon_1 HB_1 + \mu_1^{-1}\mathrm{br}(\epsilon_1) + H\mu_1 d_2\mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\mathcal{E}^{h_t}(f,f,x_t^{h_t},a_t^{h_t})^2\right]
\end{aligned}$$

Further choosing $\lambda H\mu_1 d_2 = 0.5\eta$, we get the following analog of Proposition 11.

$$\begin{aligned}
\lambda\mathbb{E}\ \mathrm{Regret}(f_t,x_t^1) \le& \mathbb{E}\ \mathbb{E}_{f|S_{t-1}}\left[-\lambda\Delta f(x_t^1) + 0.5\eta\mathcal{E}^{h_t}(f,f,x_t^{h_t},a_t^{h_t})^2\right] \\
&+\lambda\left(\epsilon_1 HB_1 + \frac{2\lambda\mathrm{br}(\epsilon_1)H_1 d_2}{\eta}\right).
\end{aligned}$$

Now we combine this result with Proposition 12 and Proposition 13 (both results still hold without modification, because the proofs did not rely on how a_t^h is generated given x_t^h), leading to

$$\begin{aligned}
\sum_{t=1}^T \lambda\mathbb{E}\ \mathrm{Regret}(f_t,x_t^1) \le& \mathcal{O}\left(\lambda\epsilon T + 4\eta T\epsilon^2 + \kappa(\epsilon) + 1.5\lambda^2 T\right) \\
&+\mathcal{O}(\epsilon T + \kappa(\epsilon) + \kappa'(\epsilon) + 1) \\
&+\lambda T\left(\epsilon_1 HB_1 + \frac{2\lambda\mathrm{br}(\epsilon_1)H_1 d_2}{\eta}\right).
\end{aligned}$$

This implies the result of Theorem 9.