Bilinear Classes: A Structural Framework for Provable Generalization in RL

Simon S. Du [1] **Sham M. Kakade** [1] **Jason D. Lee** [2] **Shachar Lovett** [3] **Gaurav Mahajan** [3] **Wen Sun** [4] **Ruosong Wang** [5]

Abstract

This work introduces Bilinear Classes, a new structural framework, which permit generalization in reinforcement learning in a wide variety of settings through the use of function approximation. The framework incorporates nearly all existing models in which a polynomial sample complexity is achievable, and, notably, also includes new models, such as the Linear Q^*/V^* model in which both the optimal Q-function and the optimal V-function are linear in some known feature space. Our main result provides an RL algorithm which has polynomial sample complexity for Bilinear Classes; notably, this sample complexity is stated in terms of a reduction to the generalization error of an underlying supervised learning sub-problem. These bounds nearly match the best known sample complexity bounds for existing models. Furthermore, this framework also extends to the infinite dimensional (RKHS) setting: for the the Linear Q^*/V^* model, linear MDPs, and linear mixture MDPs, we provide sample complexities that have no explicit dependence on the explicit feature dimension (which could be infinite), but instead depends only on information theoretic quantities.

1. Introduction

Tackling large state-action spaces is a central challenge in reinforcement learning (RL). Here, function approximation and supervised learning schemes are often employed for generalization across large state-action spaces. While there have been a number of successful applications (Mnih et al., 2013; Kober et al., 2013; Silver et al., 2017; Wu et al., 2017). there is also a realization that practical RL approaches are quite sample inefficient.

Theoretically, there is a growing body of results showing how sample efficiency is possible in RL for particular model classes (often with restrictions on the model dynamics though in some cases on the class of value functions), e.g. State Aggregation (Li, 2009; Dong et al., 2020), Linear MDPs (Yang and Wang, 2019; Jin et al., 2020), Linear Mixture MDPs (Modi et al., 2020a; Ayoub et al., 2020), Reactive POMDPs (Krishnamurthy et al., 2016), Block MDPs (Du et al., 2019a), FLAMBE (Agarwal et al., 2020b), Reactive PSRs (Littman et al., 2001), Linear Bellman Complete (Munos, 2005; Zanette et al., 2020).

More generally, there are also a few lines of work which propose more general frameworks, consisting of *structural conditions* which permit sample efficient RL; these include the low-rankness structure (e.g. the Bellman rank (Jiang et al., 2017) and Witness rank (Sun et al., 2019)) or under a complete condition (Munos, 2005; Zanette et al., 2020). The goal in these latter works is to develop a unified theory of generalization in RL, analogous to more classical notions of statistical complexity (e.g. VC-theory and Rademacher complexity) relevant for supervised learning. These latter frameworks are not contained in each other (see Table 1), and, furthermore, there are a number of natural RL models that cannot be incorporated into each of these frameworks (see Table 2).

Motivated by this latter line of work, we aim to understand if there are simple and natural structural conditions which capture the learnability in a general class of RL models.

Our Contributions. This work[1] provides a simple structural condition on the hypothesis class (which may be either model-based or value-based), where the Bellman error has a particular bilinear form, under which sample efficient learning is possible; we refer such a framework as a Bilinear Class. This structural assumption can be seen as generalizing the Bellman rank (Jiang et al., 2017); furthermore, it not only contains existing frameworks, it also covers a number

[1]University of Washington [2]Princeton University [3]University of California, San Diego [4]Cornell University [5]Carnegie Mellon University. Correspondence to: Gaurav Mahajan <gmahajan@eng.ucsd.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]Full version appears as [arXiv:2103.10897].

Framework	B-Rank	B-Complete	W-Rank	Bilinear Class (this work)
B-Rank	✓	✗	✓	✓
B-Complete	✗	✓	✗	✓
W-Rank	✗	✗	✓	✓
Bilinear Class (this work)	✗	✗	✗	✓

Table 1. Relations between frameworks. ✓: the column framework contains the row framework. ✗: the column framework does not contains the row framework. B-Rank: Bellman Rank (Jiang et al., 2017), which is defined in terms of the roll-in distribution and the function approximation class for Q^*. B-Complete: Bellman Complete (Munos, 2005) (Zanette et al. (2020) proposed a sample efficient algorithm), which assumes the function class is closed under the Bellman operator. W-Rank: Witness Rank (Sun et al., 2019): a model-based analogue of Bellman Rank. Bilinear Class: our proposed framework.

	B-Rank	B-Complete	W-Rank	Bilinear Class (this work)
Tabular MDP	✓	✓	✓	✓
Reactive POMDP (Krishnamurthy et al., 2016)	✓	✗	✓	✓
Block MDP (Du et al., 2019a)	✓	✗	✓	✓
Flambe / Feature Selection (Agarwal et al., 2020b)	✓	✗	✓	✓
Reactive PSR (Littman and Sutton, 2002)	✓	✗	✓	✓
Linear Bellman Complete (Munos, 2005)	✗	✓	✗	✓
Linear MDPs (Yang and Wang, 2019; Jin et al., 2020)	✓!	✓	✓!	✓
Linear Mixture Model (Modi et al., 2020b)	✗	✗	✗	✓
Linear Quadratic Regulator	✗	✓	✗	✓
Kernelized Nonlinear Regulator (Kakade et al., 2020)	✗	✗	✓	✓
Factored MDP (Kearns and Koller, 1999)	✗	✗	✗	✓
$Q^\star$ "irrelevant" State Aggregation (Li, 2009)	✓	✗	✗	✓
Linear $Q^\star$/$V^\star$ (this work)	✗	✗	✗	✓
RKHS Linear MDP (this work)	✗	✗	✗	✓
RKHS Linear Mixture MDP (this work)	✗	✗	✗	✓
Low Occupancy Complexity (this work)	✗	✗	✗	✓
$Q^\star$ State-action Aggregation (Dong et al., 2020)	✗	✗	✗	✗
Deterministic linear $Q^\star$ (Wen and Van Roy, 2013)	✗	✗	✗	✗
Linear $Q^\star$ (Weisz et al., 2020)	Sample efficiency is not possible			

Table 2. Whether a framework includes a model that permits a sample efficient algorithm. ✓ means the framework includes the model, ✗ means not, and ✓! means the sample complexity using that framework needs to scale with the number of action (which is not necessary). "Sample efficient is not possible" means the sample complexity needs to scale exponentially with at least one problem parameter. See Section 3.3, Appendix A and full version of the paper (link) for detailed descriptions of the models.

of new settings that are not easily incorporated in previous frameworks (see Tables 1 and 2).

Our main result presents an optimization-based algorithm, BiLin-UCB, which provably enjoys a polynomial sample complexity guarantee for Bilinear Classes (cf. Theorem 4.2). Although our framework is more general than existing ones, our proof is substantially simpler – we give a unified analysis based on the elliptical potential lemma, developed for the theory of linear bandits (Dani et al., 2008; Srinivas et al., 2009).

Furthermore, as a point of emphasis, our results are non-parametric in nature (stated in terms of an information gain quantity (Srinivas et al., 2009)), as opposed to finite dimensional as in prior work. From a technical point of view, it is not evident how to extend prior approaches to this non-parametric setting. Notably, the non-parametric regime is particularly relevant to RL due to that, in RL, performance bounds do *not* degrade gracefully with approximation error or model mis-specification (e.g. see Du et al. (2020a) for discussion of these issues); the relevance of the non-parametric regime is that it may provide additional flexibility to avoid the catastrophic quality degradation due to approximation error or model mis-specification.

A few further notable contributions are:

- *Definition of Bilinear Class:* Our key conceptual contribution is the definition of the Bilinear Class, which isolates two key critical properties. The first property is that the Bellman error can be upper bounded by a bilinear form depending on the hypothesis. The second property is that the corresponding bilinear form for all hypothesis in the hypothesis class can be estimated with the same dataset. Analogous to supervised learning, this allows for efficient data reuse to estimate the Bellman error for all hypothesis simultaneously and eliminate those with high error.
- *A reduction to supervised learning:* One appealing aspect of this framework is that the our main sample

complexity result for RL is quantified via a reduction to the generalization error of a supervised learning problem, where we have a far better understanding of the latter. This is particularly important due to that we make no explicit assumptions on the hypothesis class $\mathcal{H}$ itself, thus allowing for neural hypothesis classes in some cases (the Bilinear Class posits an *implicit* relationship between $\mathcal{H}$ and the underlying MDP $\mathcal{M}$).

- *New models:* We show our Bilinear Class framework incorporates new natural models, that are not easily incorporated into existing frameworks, e.g. linear Q^*/V^*, Low Occupancy Complexity, along with (infinite-dimensional) RKHS versions of linear MDPs and linear mixture MDPs. The linear Q^*/V^* result is particularly notable due to a recent and remarkable lower bound which showed that if we only assume Q^* is linear in some given set of features, then sample efficient learning is information theoretically not possible (Weisz et al., 2020). In perhaps a surprising contrast, our works shows that if we assume that both $Q^\star$ and $V^\star$ are linear in some given features then sample efficient learning is in fact possible.

- *Non-parametric rates:* Our work is applicable to the non-parametric setting, where we develop new analysis tools to handle a number of technical challenges. This is notable as non-parametric rates for RL are few and far between. Our results are stated in terms of the *critical information gain* which can viewed as an analogous quantity to the *critical radius*, a quantity which is used to obtain sharp rates in non-parametric statistical settings (Wainwright, 2019).

Organization Section 2 introduce some technical background and notation. Section 3 introduces our Bilinear Class framework, where we instantiate it on the several RL models, and Section 4 describes our algorithm and provides our main theoretical results.

2. Setting

We denote an episodic finite horizon, non-stationary MDP with horizon H, by $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, r, H, \{P_h\}_{h=0}^{H-1}, s_0\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the expected reward function with the corresponding random variable $R(s, a)$, $P_h : \mathcal{S} \times \mathcal{A} \mapsto \triangle(\mathcal{S})$ (where $\triangle(\mathcal{S})$ denotes the probability simplex over $\mathcal{S}$) is the transition kernel for all h, $H \in \mathbb{Z}_+$ is the planning horizon and s_0 is a fixed initial state[2]. For ease of exposition, we use the notation o_h for "observed transition info at timestep h" i.e. $o_h = (r_h, s_h, a_h, s_{h+1})$ where r_h is the observed reward $r_h = R(s_h, a_h)$ and s_h, a_h, s_{h+1} is the observed state transition at timestep h.

A deterministic, stationary policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ specifies a decision-making strategy in which the agent chooses actions adaptively based on the current state, i.e. $a_h \sim \pi(s_h)$. We denote a non-stationary policy $\pi = \{\pi_0, \dots, \pi_{H-1}\}$ as a sequence of stationary policies where $\pi_h : \mathcal{S} \mapsto \mathcal{A}$.

Given a policy π and a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the Q-function at time step h is defined as

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{h'=h}^{H-1} R(s_{h'}, a_{h'}) \mid s_h = s, a_h = a, \pi\right],$$

and, similarly, a value function time step h of a given state s under a policy π is defined as

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^{H-1} R(s_{h'}, a_{h'}) \mid s_h = s, \pi\right],$$

where both expectations are with respect to $s_0, a_0, \dots s_{H-1}, a_{H-1} \sim d^\pi$. We use $Q_h^\star$ and $V_h^\star$ to denote the Q and V-functions of the optimal policy.

Sample Efficient Algorithms. Throughout the paper, we will consider an algorithm as sample-efficient, if it uses number of trajectories polynomial in the problem horizon H, inherent dimension d, accuracy parameter $1/\epsilon$ and polylogarithmic in the number of candidate value-functions.

Notation. For any two vectors x, y, we denote $[x, y]$ as the vector that concatenates x, y, i.e., $[x, y] := [x^\top, y^\top]^\top$. For any set S, we write $\triangle(S)$ to denote the probability simplex. We often use $U(S)$ as the uniform distribution over set S. We will let $\mathcal{V}$ denote a Hilbert space (which we assume is either finite dimensional or separable).

We let $[H]$ denote the set $\{0, \dots H-1\}$. We slightly abuse notation (overloading d^π with its marginal distributions), where $s_h \sim d^\pi$, $(s_h, a_h) \sim d^\pi$, $(r_h, s_h, a_h, s_{h+1}) \sim d^\pi$ and most frequently $o_h \sim d^\pi$ denotes the marginal distributions at timestep h. We also use the shorthand notation $s_0, a_0, \dots s_{H-1}, a_{H-1} \sim \pi$, $s_h, a_h \sim \pi$ for $s_0, a_0, \dots s_{H-1}, a_{H-1} \sim d^\pi$, $s_h, a_h \sim d^\pi$.

3. Bilinear Classes

Before, we define our structural framework – Bilinear Class, we first define our hypothesis class.

Hypothesis Classes. We assume access to a hypothesis class $\mathcal{H} = \mathcal{H}_0 \times \ldots \times \mathcal{H}_{H-1}$, which can be abstract sets that permit for both *model-based and value-based* hypotheses. The only restriction we make is that for all $f \in \mathcal{H}$, we have an associated state-action value function $Q_{h,f}$ and a value

[2] Our results generalizes to any fixed initial state distribution

function $V_{h,f}$. Furthermore, we assume the hypothesis class is constrained so that $V_{h,f}(s) = \max_a Q_{h,f}(s,a)$ for all $f \in \mathcal{H}$, $h \in [H]$, and $s \in \mathcal{S}$, which is always possible as we can remove hypothesis for which this is not true. We let $\pi_{h,f}$ be the greedy policy with respect to $Q_{h,f}$, i.e., $\pi_{h,f}(s) = \text{argmax}_{a\in\mathcal{A}} Q_{h,f}(s,a)$, and π_f as the sequence of time-dependent policies $\{\pi_{h,f}\}_{h=0}^{H-1}$.

3.1. Warmup: Bellman rank, the Q and V versions.

As a motivation for our structural framework, we next discuss Bellman rank framework considered in (Jiang et al., 2017). In this case, the hypothesis class $\mathcal{H}_h$ contains Q value functions, i.e.,

$$\mathcal{H}_h \subset \{Q_h \mid Q_h \text{ is a function from } \mathcal{S}\times\mathcal{A} \mapsto [0,H]\}\,.$$

In this case, for any hypothesis $f := (Q_0, Q_1, \dots, Q_{H-1}) \in \mathcal{H}$, we take the associated state-action value function $Q_{h,f} = Q_h$ and the associated state value $V_{h,f}$ function to be greedy with respect to the $Q_{h,f}$ function i.e. $V_{h,f}(\cdot) = \max_{a\in\mathcal{A}} Q_{h,f}(\cdot, a)$.

Definition 3.1 (**V-Bellman Rank**). *A MDP has a V-*Bellman rank *of dimension d if for all $h \in [H]$, there exist functions $W_h : \mathcal{H} \to \mathbb{R}^d$ and $X_h : \mathcal{H} \to \mathbb{R}^d$, such that for all $f, g \in \mathcal{H}$:*

$$\begin{aligned}&\mathbb{E}\big[V_{h,g}(s_h) - r(s_h,a_h) - V_{h+1,g}(s_{h+1})\big]\\ &\quad= \langle W_h(g) - W_h(f^\star), X_h(f)\rangle.\end{aligned}$$

where $a_{0:h-1} \sim d^{\pi_f}$, $a_h = \pi_g(s_h)$ and $s_{h+1} \sim P(s_h, a_h)$

Even though (Jiang et al., 2017) only considered V-Bellman Rank, as a natural extension of this definition, we can also consider the Q-Bellman Rank.

Definition 3.2 (**Q-Bellman Rank**). *For a given MDP $\mathcal{M}$, we say that our state-action value hypothesis class $\mathcal{H}$ has a Q-*Bellman rank *of dimension d if for all $h \in [H]$, there exist functions $W_h : \mathcal{H} \to \mathbb{R}^d$ and $X_h : \mathcal{H} \to \mathbb{R}^d$, such that for all $f, g \in \mathcal{H}$*

$$\begin{aligned}&\mathbb{E}\big[Q_{h,g}(s_h,a_h) - r(s_h,a_h) - V_{h+1,g}(s_{h+1})\big]\\ &\quad= \langle W_h(g) - W_h(f^\star), X_h(f)\rangle.\end{aligned}$$

where $a_{0:h} \sim d^{\pi_f}$ and $s_{h+1} \sim P(s_h, a_h)$

Let us interpret how the two definitions differ in the usage of functions $V_{h,f}$ vs $Q_{h,f}$ (along with the usage of the "estimation" policies $a_{0:h} \sim \pi_f$ vs $a_{0:h-1} \sim \pi_f$ and $a_h \sim \pi_g$). Recall that the Bellman equations can be written in terms of the value functions or the state-action values; here, the intuition is that the former definition corresponds to enforcing Bellman consistency of the value functions while the latter definition corresponds to enforcing Bellman consistency of the state-action value functions. Our more general structural framework, Bilinear Classes, will cover both these definitions for infinite dimensional hypothesis class (note that (Jiang et al., 2017) only considered finite dimensional hypothesis class).

3.2. Bilinear Classes

We now introduce a new structural framework – the Bilinear Class.

Realizability. We say that $\mathcal{H}$ is *realizable* for an MDP $\mathcal{M}$ if, for all $h \in [H]$, there exists a hypothesis $f^\star \in \mathcal{H}$ such that $Q^\star_h(s,a) = Q_{h,f^\star}(s,a)$, where $Q^\star_h$ is the optimal state-action value at time step h in the ground truth MDP $\mathcal{M}$. For instance, for the model-based perspective, the realizability assumption is implied if the ground truth transition P belongs to our hypothesis class $\mathcal{H}$.

Now we are ready to introduce the Bilinear Class.

Definition 3.3 (Bilinear Class). *Consider an MDP $\mathcal{M}$, a hypothesis class $\mathcal{H}$, a discrepancy function $\ell_f : (\mathbb{R}\times\mathcal{S}\times\mathcal{A}\times\mathcal{S})\times\mathcal{H} \to \mathbb{R}$ (defined for each $f \in \mathcal{H}$), and a set of estimation policies $\Pi_{\text{est}} = \{\pi_{est}(f) : f \in \mathcal{H}\}$. We say $(\mathcal{H}, \ell_f, \Pi_{\text{est}}, \mathcal{M})$ is (*implicitly*) a* Bilinear Class *if $\mathcal{H}$ is realizable in $\mathcal{M}$ and if there exist functions $W_h : \mathcal{H} \to \mathcal{V}$ and $X_h : \mathcal{H} \to \mathcal{V}$ for some Hilbert space $\mathcal{V}$, such that the following two properties hold for all $f \in \mathcal{H}$ and $h \in [H]$:*

1. *We have:*

$$\begin{aligned}&\big|\mathbb{E}_{a_{0:h}\sim\pi_f}\big[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1})\big]\big|\\ &\le |\langle W_h(f) - W_h(f^\star), X_h(f)\rangle| \qquad (1)\end{aligned}$$

2. *The policy $\pi_{est}(f)$ and discrepancy measure $\ell_f(o_h, g)$ can be used for estimation in the following sense: for any $g \in \mathcal{H}$, we have that (here $o_h = (r_h, s_h, a_h, s_{h+1})$ is the "observed transition info")*

$$\begin{aligned}&\big|\mathbb{E}_{a_{0:h-1}\sim\pi_f}\mathbb{E}_{a_h\sim\pi_{est}(f)}\big[\ell_f(o_h,g)\big]\big|\\ &= |\langle W_h(g) - W_h(f^\star), X_h(f)\rangle|\,. \qquad (2)\end{aligned}$$

   Typically, $\pi_{est}(f)$ will be either the uniform distribution on $\mathcal{A}$ or π_f itself; in the latter case, we refer to the estimation strategy as being on-policy.

We also define $\mathcal{X}_h := \{X_h(f) \colon f \in \mathcal{H}\}$ and $\mathcal{X} := \{\mathcal{X}_h : h \in [H]\}$.

We emphasize the above definition only assumes the existence of W and X functions. Particularly, our algorithm only uses the discrepancy function ℓ_f, and does not need to know W or X. A typical example of discrepancy function $\ell_f(o_h, g)$ would be the bellman error

$Q_{h,g}(s_h, a_h) - r_h - V_{h+1,g}(s_{h+1})$, but we would often need to use a different discrepancy function see for e.g. Linear Mixture Models (Section 3.3.1).

We now provide some intuition for definition of Bilinear Class. The first part of the definition (Equation (1)) basically relates the Bellman error for hypothesis f (and hence sub-optimality) to the sum of bilinear forms $|\langle W_h(f) - W_h(f^\star), X_h(f)\rangle|$ (see for example proof of Lemma C.5). Crucially, the second part of the definition (Equation (2)), allows us to "reuse" data from hypothesis f to estimate the bilinear form $|\langle W_h(g) - W_h(f^\star), X_h(f)\rangle|$ for *all* hypothesis g in our hypothesis class! This is reminiscent of uniform convergence guarantees in supervised learning, where data can be reused to simultaneously estimate the loss for all hypothesis and eliminate those with high loss.

3.2.1. Finite Bellman rank $\implies$ Bilinear Class

Here we show our framework naturally generalizes the Bellman rank framework (Section 3.1). For Q-bellman rank case, we define the discrepancy function ℓ_f for observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$ as:

$$\ell_f(o_h, g) = Q_{h,g}(s_h, a_h) - r_h - V_{h+1,g}(s_{h+1}).$$

Lemma 3.1 (Finite Q-Bellman Rank $\implies$ Bilinear Class). *For given MDP $\mathcal{M}$, suppose our hypothesis class $\mathcal{H}$ has a Q-Bellman rank of dimension d. Then, for on-policy estimation policies $\pi_{est} = \pi_f$, and the discrepancy function ℓ_f defined above, $(\mathcal{H}, \ell_f, \Pi_{\text{est}}, \mathcal{M})$ is (*implicitly*) a* Bilinear Class.

In the V-Bellman rank setting, we define the discrepancy function $\ell_f(o_h, g)$ for observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$ as:

$$\frac{\mathbf{1}\{a_h = \pi_g(s_h)\}}{1/|\mathcal{A}|}\left(V_{h,g}(s_h) - r_h - V_{h+1,g}(s_{h+1})\right).$$

Lemma 3.2 (Finite V-Bellman Rank $\implies$ Bilinear Class). *For given MDP $\mathcal{M}$, suppose our hypothesis class $\mathcal{H}$ has a V-Bellman rank of dimension d. Then, for uniform estimation policies $\pi_{est} = U(\mathcal{A})$, and the discrepancy function ℓ_f defined above, $(\mathcal{H}, \ell_f, \Pi_{\text{est}}, \mathcal{M})$ is (*implicitly*) a* Bilinear Class.

3.3. Examples

We now provide examples of Bilinear Classes: two known models (Linear Bellman Complete and Linear Mixture Models) and two new models that we propose (Linear $Q^\star/V^\star$ and Low Occupancy Complexity). We return to these examples to give non-parametric sample complexities in Section 4.2.

3.3.1. Linear Mixture MDP.

First, we show our definition naturally captures model-based hypothesis class.

Definition 3.4 (Linear Mixture Model). *We say that a MDP $\mathcal{M}$ is a* Linear Mixture Model *if there exists (known) features $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathcal{V}$ and $\psi : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{V}$; and (unknown) $\theta^\star \in \mathcal{V}$ for some Hilbert space $\mathcal{V}$ such that for all $h \in [H]$ and $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$*

$$P_h(s' \mid s, a) = \langle \theta_h^\star,\ \phi(s, a, s')\rangle \text{ and } r(s, a) = \langle \theta_h^\star,\ \psi(s, a)\rangle.$$

We denote hypothesis in our hypothesis class $\mathcal{H}$ as tuples $(\theta_0, \ldots \theta_{H-1})$, where $\theta_h \in \mathcal{V}$. Recall that given a model $f \in \mathcal{H}$ (i.e. f is the time-dependent transitions, i.e., $f_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$), we denote $V_{h,f}$ as the optimal value function under model f and corresponding reward function (in this case defined by ψ). Specifically, for any hypothesis $g = \{\theta_0, \ldots, \theta_{H-1}\} \in \mathcal{H}$, $V_{h,g}$ and $Q_{h,g}$ satisfy the following Bellman optimality equation:

$$Q_{h,g}(s_h, a_h) = \theta_h^\top \left(\psi(s_h, a_h) + \sum_{\bar{s} \in \mathcal{S}} \phi(s_h, a_h, \bar{s}) V_{h+1,g}(\bar{s})\right)$$

Note that in this example, discrepancy function will explicitly depend on f. For hypothesis $g = \{\theta_0, \ldots, \theta_{H-1}\} \in \mathcal{H}$ and observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$, we define

$$\ell_f(o_h, g) = \theta_h^\top \left(\psi(s_h, a_h) + \sum_{\bar{s} \in \mathcal{S}} \phi(s_h, a_h, \bar{s}) V_{h+1,f}(\bar{s})\right) - \left(V_{h+1,f}(s_{h+1}) + r_h\right).$$

Lemma 3.3 (Linear Mixture Model $\implies$ Bilinear Class). *Consider a MDP $\mathcal{M}$ which is a Linear Mixture Model. Then, for the hypothesis class $\mathcal{H}$, discrepancy function ℓ_f defined above and on-policy estimation policies $\pi_{est}(f) = \pi_f$, $(\mathcal{H}, \ell_f, \Pi_{\text{est}}, \mathcal{M})$ is (*implicitly*) a* Bilinear Class.

3.3.2. Linear $Q^\star/V^\star$ (new model)

We introduce a new model: *linear $Q^\star/V^\star$* where we assume both the optimal $Q^\star$ and $V^\star$ are linear functions in features that lie in (possibly infinite dimensional) Hilbert space.

Definition 3.5 (Linear $Q^\star/V^\star$). *We say that a MDP $\mathcal{M}$ is a* linear $Q^\star/V^\star$ *model if there exist (known) features $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{V}_1$, $\psi : \mathcal{S} \mapsto \mathcal{V}_2$ and (unknown) $(w^\star, \theta^\star) \in \mathcal{V}_1 \times \mathcal{V}_2$ for some Hilbert spaces $\mathcal{V}_1, \mathcal{V}_2$ such that for all $h \in [H]$ and for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$Q_h^\star(s, a) = \langle w_h^\star,\ \phi(s, a)\rangle \quad \text{and} \quad V_h^\star(s') = \langle \theta_h^\star,\ \psi(s')\rangle.$$

Here, our hypothesis class $\mathcal{H} = \mathcal{H}_0 \times \ldots, \mathcal{H}_{H-1}$ is a set of linear functions i.e. for all $h \in [H]$, the set $\mathcal{H}_h$ is defined as:

$$\Big\{(w, \theta) \in \mathcal{V}_1 \times \mathcal{V}_2 \colon \max_{a \in \mathcal{A}} w^\top \phi(s,a) = \theta^\top \psi(s)\,,\ \forall s \in \mathcal{S}\Big\}.$$

We define the following discrepancy function ℓ_f (in this case the discrepancy function does not depend on f), for hypothesis $g = \{(w_h, \theta_h)\}_{h=0}^{H-1}$ and observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$:

$$\begin{aligned}\ell_f(o_h, g) &= Q_{h,g}(s_h, a_h) - r_h - V_{h+1,g}(s_{h+1}) \\ &= w_h^\top \phi(s_h, a_h) - r_h - \theta_{h+1}^\top \psi(s_{h+1})\,.\end{aligned}$$

Lemma 3.4 (Linear $Q^\star/V^\star \implies$ Bilinear Class). *Consider a MDP $\mathcal{M}$ which is a linear $Q^\star/V^\star$ model. Then, for the hypothesis class $\mathcal{H}$, the discrepancy function ℓ_f defined above and on-policy estimation policies $\pi_{est}(f) = \pi_f$, $(\mathcal{H}, \ell_f, \Pi_{\mathrm{est}}, \mathcal{M})$ is (*implicitly*) a* Bilinear Class.

3.3.3. Bellman Complete and Linear MDPs

We now consider Bellman Complete which captures the linear MDP model (see the full paper for more detail on linear MDP model). Here, our hypothesis class $\mathcal{H}$ is set of linear functions with respect to some (known) feature $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{V}$, where $\mathcal{V}$ is a Hilbert space. We denote hypothesis in our hypothesis class $\mathcal{H}$ as tuples $(\theta_0, \ldots \theta_{H-1})$, where $\theta_h \in \mathcal{V}$.

Definition 3.6 (Linear Bellman Complete). *We say our hypothesis class $\mathcal{H}$ is* Linear Bellman Complete *with respect to $\mathcal{M}$ if $\mathcal{H}$ is realizable and there exists $\mathcal{T}_h : \mathcal{V} \to \mathcal{V}$ such that for all $(\theta_0, \ldots \theta_{H-1}) \in \mathcal{H}$ and $h \in [H]$,*

$$\mathcal{T}_h(\theta_{h+1})^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a' \in \mathcal{A}} \theta_{h+1}^\top \phi(s', a').$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

We define the following discrepancy function ℓ_f (in this case the discrepancy function does not depend on f), for hypothesis $g = (\theta_0, \ldots, \theta_{H-1})$ and observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$:

$$\begin{aligned}\ell_f(o_h, g) &= Q_{h,g}(s_h, a_h) - r_h - V_{h+1,g}(s_{h+1}) \\ &= \theta_h^\top \phi(s_h, a_h) - r_h - \max_{a' \in \mathcal{A}} \theta_{h+1}^\top \phi(s_{h+1}, a')\,.\end{aligned}$$

Lemma 3.5 (Linear Bellman Complete $\implies$ Bilinear Class). *Consider an MDP $\mathcal{M}$ and hypothesis class $\mathcal{H}$ such that $\mathcal{H}$ is Linear Bellman Complete with respect to $\mathcal{M}$. Then, for on-policy estimation policies $\pi_{est}(f) = \pi_f$ and the discrepancy function ℓ_f defined above, $(\mathcal{H}, \ell_f, \Pi_{\mathrm{est}}, \mathcal{M})$ is (*implicitly*) a* Bilinear Class.

3.3.4. Low Occupancy Complexity (new model).

We introduce another new model: *Low Occupancy Complexity*.

Definition 3.7 (Low Occupancy Complexity). *We say that a MDP $\mathcal{M}$ and hypothesis class $\mathcal{H}$ has* low occupancy complexity *with respect to a (possibly unknown) feature mapping $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathcal{V}$ (where $\mathcal{V}$ is a Hilbert space) if $\mathcal{H}$ is realizable and there exists a (possibly unknown) $\beta_h : \mathcal{H} \mapsto \mathcal{V}$ for $h \in [H]$ such that for all $f \in \mathcal{H}$ and $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$ we have that:*

$$d^{\pi_f}(s_h, a_h) = \langle \beta_h(f), \phi_h(s_h, a_h) \rangle.$$

It is important to emphasize that for this hypothesis class, we are only assuming realizability, but it is otherwise arbitrary (e.g. it could be a neural state-action value class) and the algorithm does not need to know the features ϕ_h nor β_h. It is straight forward to see that such a class is Bilinear Class with discrepancy function ℓ_f defined for hypothesis $g \in \mathcal{H}$ and observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$ as,

$$\ell_f(o_h, g) = Q_{h,g}(s_h, a_h) - r_h - V_{h+1,g}(s_{h+1})$$

Lemma 3.6 (Low Occupancy Complexity $\implies$ Bilinear Class). *Consider a MDP $\mathcal{M}$ and hypothesis class $\mathcal{H}$ which has low occupancy complexity. Then, for the the discrepancy function ℓ_f defined above and on-policy estimation policies $\pi_{est}(f) = \pi_f$, $(\mathcal{H}, \ell_f, \Pi_{\mathrm{est}}, \mathcal{M})$ is (*implicitly*) a* Bilinear Class.

Note that as such the hypothesis class $\mathcal{H}$ could be arbitrary and unlike other models where we assume linearity, here it could be a neural state-action value class. Our model can also capture the setting where the state-only occupancy has low complexity, i.e., $d^{\pi_f}(s_h) = \beta_h(f)\mu_h(s_h)$, for some $\mu_h : \mathcal{S} \to \mathcal{V}$. In this case, we will use $\pi_{est} = U(\mathcal{A})$.

4. The Algorithm and Theory

Our algorithm, BiLin-UCB, is described in Algorithm 1, which takes three parameters as inputs, the number of iterations T, the trajectory batch size m per iteration and a confidence radius R. The key component of the algorithm is a constrained optimization in Line 1. For each time step h, we use all previously collected data to form a single constraint using ℓ_f. The constraint refines the original version space $\mathcal{H}$ to be a restricted version space containing only hypothesis that are consistent with the current batch data. We then perform an optimistic optimization: we search for a feasible hypothesis g that achieves the maximum total reward $V_g(s_0)$.

There are two ways to collect batch samples. For the case where $\pi_{est} = \pi_{f_t}$, then for data collection in Line 1, we can

Algorithm 1 BiLin-UCB

1: **Input**: number of iterations T, estimator function ℓ, batch size m, confidence radius R
2: **for** iteration $t = 0, 1, 2, \ldots, T-1$ **do**
3: Set f_t as the solution of the following program:

$$\underset{g\in\mathcal{H}}{\text{argmax}}\, V_g(s_0) \text{ subject to}$$
$$\sum_{i=0}^{t-1}(\mathcal{L}_{\mathcal{D}_{i;h},f_i}(g))^2 \le R^2 \quad \forall h \in [H]$$

4: For all $h \in [H]$, create batch datasets $\mathcal{D}_{t;h} = \{(r_h^i, s_h^i, a_h^i, s_{h+1}^i)\}_{i=0}^{m-1}$ sampled from distribution induced by $a_{0:h-1} \sim d^{\pi_{f_t}}$ and $a_h \sim \pi_{est}$.
5: **end for**
6: **return** $\max_{t\in[T]} V^{\pi_{f_t}}$.

generate m length-H trajectories by executing π_{f_t} starting from s_0. For the general case (e.g. consider setting π_{est} to be a uniform distribution over $\mathcal{A}$), we gather the data for each $h \in [H]$ independently. For $h \in [H]$, we first roll-in with π_{f_t} to generate s_h; then execute $a_h \sim \pi_{est}$; and then continue to generate $s_{h+1} \sim P_h(\cdot|s_h, a_h)$ and $r_h \sim R(\cdot|s_h, a_h)$. Repeating this process for all h, we need Hm trajectories to form the batch datasets $\{\mathcal{D}_{t;h}\}_{h=0}^{H-1}$.

4.1. Main Theory: Generalization in Bilinear Classes

We now present our main result. We first define some notations. We denote the expectation of the function $\ell_f(\cdot, g)$ under distribution μ over $\mathbb{R}\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ by

$$\mathcal{L}_{\mu,f}(g) = \mathbb{E}_{o\sim\mu}[\ell_f(o, g)]$$

For a set $\mathcal{D} \subset \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, we will also use $\mathcal{D}$ to represent the uniform distribution over this set.

Assumption 4.1 (Ability to Generalize). *We assume there exists functions $\varepsilon_{gen}(m, \mathcal{H})$ and $conf(\delta)$ such that for any distribution μ over $\mathbb{R}\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and for any $\delta \in (0, 1/2)$, with probability of at least $1-\delta$ over choice of an i.i.d. sample $\mathcal{D} \sim \mu^m$ of size m,*

$$\sup_{g\in\mathcal{H}} |\mathcal{L}_{\mathcal{D},f}(g) - \mathcal{L}_{\mu,f}(g)| \le \varepsilon_{gen}(m, \mathcal{H}) \cdot conf(\delta)$$

Remark 4.1. *It is helpful to separate the dependence of generalization error on failure probability δ and number of samples m in order to state Theorem 4.2 concisely. $\varepsilon_{gen}(m, \mathcal{H})$ is related to uniform convergence and measures the generalization error of hypothesis class $\mathcal{H}$ and for the hypothesis classes discussed in this paper, $\varepsilon_{gen}(m, \mathcal{H}) \to 0$ as $m \to \infty$. One example is when $\pi_{est} = \pi_f$, and $\mathcal{H}$ is a discrete function class, then we have $\varepsilon_{gen}(m, \mathcal{H}) = O\left(\sqrt{(1 + \ln(|\mathcal{H}|))/m}.\right)$. In Appendix F, we also discuss uniform convergence via a novel covering argument for infinite dimensional RKHS.*

Recall the definitions $\mathcal{X}_h := \{X_h(f) : f \in \mathcal{H}\}$ and $\mathcal{X} := \{\mathcal{X}_h : h \in [H]\}$. We first present our main theorem for the finite dimensional case i.e. when $\mathcal{X}_h \subset \mathbb{R}^d$ for all timesteps h.

Theorem 4.1. *(Finite-dimensional case) Suppose $(\mathcal{H}, \ell, \Pi_{\text{est}}, \mathcal{M})$ is a Bilinear Class with $\mathcal{X}_h \subset \mathbb{R}^d$ for all timesteps h and Assumption 4.1 holds. Assume $\sup_{f\in\mathcal{H},h\in[H]} \|W_h(f)\|_2 \le B_W$ and $\sup_{f\in\mathcal{H},h\in[H]} \|X_h(f)\|_2 \le B_X$. Fix $\delta \in (0, 1/3)$ and batch sample size m and define:*

$$\widetilde{d}_m = H\left\lceil 3d \ln\left(1 + \frac{3B_X^2 B_W^2}{\varepsilon_{gen}^2(m, \mathcal{H})}\right)\right\rceil.$$

Set the parameters as: number of iterations $T = \widetilde{d}_m$ and confidence radius $R = \sqrt{T}\varepsilon_{gen}(m, \mathcal{H}) \cdot conf(\delta/(TH))$. With probability at least $1-\delta$, Algorithm 1 uses at most mHT trajectories and returns a hypothesis f such that:

$$V^\star(s_0) - V^{\pi_f}(s_0) \le 3H\varepsilon_{gen}(m, \mathcal{H}) \cdot \left(1 + \sqrt{\widetilde{d}_m} \cdot conf\left(\frac{\delta}{\widetilde{d}_m H}\right)\right).$$

As discussed in the Remark 4.1, $\varepsilon_{\text{gen}}(m, \mathcal{H})$ and $\text{conf}(\delta)$ measure the uniform convergence of discrepancy functions ℓ_f for the hypothesis class $\mathcal{H}$. Therefore, if $\varepsilon_{\text{gen}}(m, \mathcal{H})$ decays at least as fast as $m^{-\alpha}$ for any constant α, we will get efficient reinforcement learning. In fact, we will see in our examples (Section 4.2), that this is true for all known models where efficient reinforcement learning is possible. One such example is finite hypothesis classes where we immediately get the following sample complexity bound showing only a *logarithmic* dependence on the size of the hypothesis space.

Corollary 4.1. *(Finite-dimensional, Finite Hypothesis Case) Suppose $(\mathcal{H}, \ell, \Pi_{\text{est}}, \mathcal{M})$ is a Bilinear Class with $\mathcal{X}_h \subset \mathbb{R}^d$ for all timesteps h, $|\mathcal{H}| > 1$ and Assumption 4.1 holds. Assume $\sup_{f\in\mathcal{H},h\in[H]} \|W_h(f)\|_2 \le B_W$ and $\sup_{f\in\mathcal{H},h\in[H]} \|X_h(f)\|_2 \le B_X$ for some $B_X, B_W \ge 1$. Assume the discrepancy function ℓ_f is bounded i.e. $\sup_{f\in\mathcal{H}} |\ell_f(\cdot)| \le H + 1$. Fix $\delta \in (0, 1/3)$ and $\epsilon \in (0, 1)$. Then there exists absolute constants c_1, c_2, c_3, c_4 such that setting the parameters: batch sample size*

$$m = \frac{c_1 \nu d H^5 \ln(dH^2) \ln(|\mathcal{H}|) \ln(1/\delta)}{\epsilon^2},$$

number of iterations $T = c_2 dH \ln\left(B_X B_W m\right)$ and confidence radius $R = c_3\sqrt{T} \cdot H\sqrt{\ln(|\mathcal{H}|)/m} \cdot \ln(TH/\delta)$, with probability at least $1-\delta$, Algorithm 1 returns a hypothesis f such that $V^\star(s_0) - V^{\pi_f}(s_0) \le \epsilon$ using at most

$$\frac{c_4 \nu^2 d^2 H^7 \ln(dH^2) \ln(|\mathcal{H}|) \ln(1/\delta)}{\epsilon^2}$$

trajectories where

$$\nu = \ln\Big(\frac{dHB_XB_W \ln(|\mathcal{H}|)\ln(1/\delta)}{\epsilon}\Big).$$

The proof for this corollary follows from bounds on $\varepsilon_{\text{gen}}(m,\mathcal{H})$ and $\text{conf}(\delta)$ using Hoeffding's inequality (Lemma G.1). We present the complete proof in Appendix D.

Our next results will be non-parametric in nature and therefore it is helpful to introduce the *maximum information gain* (Srinivas et al., 2009), which captures an important notion of the effective dimension of a set. Let $\mathcal{X} \subset \mathcal{V}$, where $\mathcal{V}$ is a Hilbert space. For $\lambda > 0$ and integer $n > 0$, the *maximum information gain* $\gamma_n(\lambda;\mathcal{X})$ is defined as:

$$\gamma_n(\lambda;\mathcal{X}) := \max_{x_0\ldots x_{n-1}\in\mathcal{X}} \ln\det\left(\mathrm{I} + \frac{1}{\lambda}\sum_{t=0}^{n-1} x_t x_t^\top\right). \quad (3)$$

If $\mathcal{X}$ is of the form $\mathcal{X} = \{\mathcal{X}_h : h \in [H]\}$, we use the notation

$$\gamma_n(\lambda;\mathcal{X}) := \sum_{h\in[H]} \gamma_n(\lambda;\mathcal{X}_h)\,. \quad (4)$$

Define *critical information gain*, denoted by $\widetilde{\gamma}(\lambda;\mathcal{X})$, as the smallest integer $k > 0$ s.t. $k \ge \gamma_k(\lambda;\mathcal{X})$, i.e.

$$\widetilde{\gamma}(\lambda;\mathcal{X}) := \min_{k\ge\gamma_k(\lambda;\mathcal{X})} k, \quad (5)$$

(where k is an integer). Note that such a $\widetilde{\gamma}(\lambda;\mathcal{X})$ exists provided that the information gain $\gamma_n(\lambda;\mathcal{X})$ has a sufficiently mild growth condition in both n and $1/\lambda$. The *critical information gain* can viewed as an analogous quantity to the *critical radius*, a quantity which arises in non-parametric statistics (Wainwright, 2019).

Remark 4.2. *For finite dimension setting where $\mathcal{X} \subset \mathbb{R}^d$ and $\|x\| \le B_X$ for any $x \in \mathcal{X}$, we have: $\gamma_n(\lambda;\mathcal{X}) \le d\ln\left(1 + nB_X^2/d\lambda\right)$ and $\tilde{\gamma}(\lambda;\mathcal{X}) \le 3d\ln\left(1+3B_X^2/\lambda\right)$ (see Lemma G.3 for a proof). Note that $1/\lambda$, n, and the norm bound B_X only appear inside the log. Furthermore, it is possible that $\gamma_n(\lambda;\mathcal{X})$ is much smaller than the dimension of $\mathcal{X}$ (or $\mathcal{V}$), when the eigenspectrum of the covariance matrices concentrates in a low-dimension subspace. In fact when $\mathcal{X}$ belongs to some infinite dimensional RKHS, $\gamma_n(\lambda;\mathcal{X})$ could still be small (Srinivas et al., 2009).*

We now present our main theorem. Recall the definitions $\mathcal{X}_h := \{X_h(f)\colon f \in \mathcal{H}\}$ and $\mathcal{X} := \{\mathcal{X}_h : h \in [H]\}$.

Theorem 4.2. *(RKHS case) Suppose $(\mathcal{H}, \ell, \Pi_{\text{est}}, \mathcal{M})$ is a Bilinear Class and Assumption 4.1 holds. Assume $\sup_{f\in\mathcal{H},h\in[H]}\|W_h(f)\|_2 \le B_W$. Fix $\delta \in (0, 1/3)$, batch sample size m, and define:*

$$\widetilde{d}_m = \widetilde{\gamma}\Big(\varepsilon^2_{gen}(m,\mathcal{H})/B_W^2;\mathcal{X}\Big).$$

Set the parameters as: number of iterations $T = \widetilde{d}_m$ and confidence radius $R = \sqrt{\widetilde{d}_m}\varepsilon_{gen}(m,\mathcal{H}) \cdot conf(\delta/(\widetilde{d}_m H))$. With probability at least $1-\delta$, Algorithm 1 uses at most $mH\widetilde{d}_m$ trajectories and returns a hypothesis f such that:

$$V^\star(s_0) - V^{\pi_f}(s_0) \le 3H\varepsilon_{gen}(m,\mathcal{H}) \cdot \Big(1 + \sqrt{\widetilde{d}_m} \cdot conf(\frac{\delta}{\widetilde{d}_m H})\Big)\,.$$

Next, we provide an elementary and detailed proof for our main theorem using an elliptical potential argument.

4.2. Corollaries for Particular Models

In this section, we apply our main theorem to special models: linear $Q^\star/V^\star$, bellman complete, linear mixture model, and low occupancy complexity model. While linear bellman complete and linear mixture model have been studied, our results extends to infinite dimensional RKHS setting. Due to space constraints, we present the finite dimensional results in this section and defer the infinite dimensional results to Appendix D.

4.2.1. LINEAR $Q^\star/V^\star$

In this subsection, we provide the sample complexity result for the linear $Q^\star/V^\star$ model (Definition 3.5). To state our results for linear $Q^\star/V^\star$, we define the following sets:

$$\Phi = \Big\{\phi(s,a)\colon (s,a) \in \mathcal{S}\times\mathcal{A}\Big\},\ \Psi = \Big\{\psi(s')\colon s' \in \mathcal{S}\Big\}.$$

and define the concatenation set[3]

$$\Phi \circ \Psi = \Big\{[x,y] : x \in \Phi, y \in \Psi\Big\}$$

Corollary 4.2 (Finite Dimensional Linear $Q^\star/V^\star$). *Suppose MDP $\mathcal{M}$ is a linear $Q^\star/V^\star$ model with $\Phi\circ\Psi \subset \mathbb{R}^d$. Assume $\sup_{(w,\theta)\in\mathcal{H}_h,h\in[H]}\|[w,\theta]\|_2 \le B_W$ and $\sup_{x\in\Phi\circ\Psi}\|x\|_2 \le B_X$ for some $B_X, B_W \ge 1$. Fix $\delta \in (0,1/3)$ and $\epsilon \in (0,H)$. There exists an appropriate setting of batch sample size m, number of iteration T and confidence radius R such that with probability at least $1-\delta$, Algorithm 1 returns a hypothesis f such that $V^\star(s_0) - V^{\pi_f}(s_0) \le \epsilon$ using at most*

$$c_1 \frac{d^3H^6\ln(1/\delta)}{\epsilon^2} \cdot \Big(\ln\big(\frac{d^3H^7B_X^2B_W^2\ln(1/\delta)}{\epsilon^2}\big)\Big)^5$$

trajectories for some absolute constant c_1.

4.2.2. BELLMAN COMPLETE.

In this subsection, we provide the sample complexity result for the Linear Bellman Complete model (Definition 3.6). To state our results, we define

$$\Phi = \{\phi(s,a) : s,a \in \mathcal{S}\times\mathcal{A}\}\,.$$

[3] For infinite dimensional Φ and Ψ, we consider the natural inner product space where $\langle[x_1,y_1],[x_2,y_2]\rangle = \langle x_1,x_2\rangle + \langle y_1,y_2\rangle$.

We first provide the result for the finite dimensional case i.e. when $\Phi \subset \mathcal{V} \subset \mathbb{R}^d$.

Corollary 4.3 (Finite Dimensional Linear Bellman Complete). *Suppose $\mathcal{H}$ is* Bellman Complete *with respect to MDP $\mathcal{M}$ for some Hilbert space $\mathcal{V} \subset \mathbb{R}^d$. Assume $\sup_{\theta \in \mathcal{H}_h, h\in[H]} \|\theta\|_2 \le B_W$ and $\sup_{x\in\Phi}\|x\|_2 \le B_X$ for some $B_X, B_W \ge 1$. Fix $\delta \in (0, 1/3)$ and $\epsilon \in (0, H)$. There exists an appropriate setting of batch sample size m, number of iteration T and confidence radius R such that with probability at least $1-\delta$, Algorithm 1 returns a hypothesis f such that $V^\star(s_0) - V^{\pi_f}(s_0) \le \epsilon$ using at most*

$$c_1 \frac{d^3 H^6 \ln(1/\delta)}{\epsilon^2} \cdot \Big(\ln\big(\frac{d^3 H^7 B_X^2 B_W^2 \ln(1/\delta)}{\epsilon^2}\big)\Big)^5$$

trajectories for some absolute constant c_1.

In comparison, (Jin et al., 2020) has sample complexity $\widetilde{O}(d^3H^3/\epsilon^2\log(1/\delta))$ and (Zanette et al., 2020) has $\widetilde{O}(d^2H^3/\epsilon^2\log(1/\delta))$. We prove the sample complexity result (Corollary D.2) for the RKHS case in Appendix D. Note that RKHS Linear MDP is a special instance of RKHS Bellman Complete. Prior works that studied RKHS Linear MDP either achieves worse rate (Agarwal et al., 2020a) or further assumes finite covering dimension of the space of all possible upper confidence bound Q functions which are algorithm dependent quantities (Yang et al., 2020).

4.2.3. Linear mixture model

In this subsection, we provide the sample complexity result for the Linear Mixture model (Definition 3.4). To present our sample complexity results, we define:

$$\Phi_h = \Big\{\psi(s,a) + \sum_{s'\in\mathcal{S}} \phi(s,a,s')V_{f;h+1}(s') \\ : (s,a)\in\mathcal{S}\times\mathcal{A}, f\in\mathcal{H}\Big\}.$$

We first provide the result for the finite dimensional case i.e. when $\Phi_h \subset \mathcal{V} \subset \mathbb{R}^d$ for all $h \in [H]$.

Corollary 4.4 (Finite Dimensional Linear Mixture Model). *Suppose MDP $\mathcal{M}$ is a* linear Mixture Model *for some Hilbert space $\mathcal{V} \subset \mathbb{R}^d$. Assume $\sup_{\theta\in\mathcal{H}_h,h\in[H]}\|\theta\|_2 \le B_W$ and $\sup_{x\in\Phi_h, h\in[H]}\|x\|_2 \le B_X$ for some $B_X, B_W \ge 1$. Fix $\delta \in (0,1/3)$ and $\epsilon \in (0,H)$. There exists an appropriate setting of batch sample size m, number of iteration T and confidence radius R such that with probability at least $1-\delta$, Algorithm 1 returns a hypothesis f such that $V^\star(s_0) - V^{\pi_f}(s_0) \le \epsilon$ using at most*

$$c_1 \frac{d^3 H^6 \ln(1/\delta)}{\epsilon^2} \cdot \Big(\ln\big(\frac{d^3 H^7 B_X^2 B_W^2 \ln(1/\delta)}{\epsilon^2}\big)\Big)^5$$

trajectories for some absolute constant c_1.

In comparison, (Modi et al., 2020a) has sample complexity $\widetilde{O}(d^2H^2/\epsilon^2\log(1/\delta))$. We will present the sample complexity result (Corollary D.3) for the infinite dimensional RKHS case in Appendix D.

4.2.4. Low Occupancy Complexity

In this section, we will prove sample complexity bounds for low occupancy complexity model (Definition 3.7).

Corollary 4.5 (Low Occupancy Complexity). *Suppose $\mathcal{H}$ has* low occupancy complexity. *Assume $\sup_{f\in\mathcal{H}_h,h\in[H]}\|W_h(f)\|_2 \le B_W$. Fix $\delta \in (0,1/3)$, batch sample size m, and define:*

$$\widetilde{d}_m(\mathcal{X}) = \widetilde{\gamma}\Big(\frac{8H^2\big(1+\ln(|\mathcal{H}|)\big)}{mB_W^2};\mathcal{X}\Big).$$

Set $T = \widetilde{d}_m(\mathcal{X})$ and $R = (2\sqrt{2}H/\sqrt{m}) \cdot \sqrt{\widetilde{d}_m(\mathcal{X})} \cdot \sqrt{1+\ln\big(|\mathcal{H}|\big)} \cdot \sqrt{\ln\big(\widetilde{d}_m(\mathcal{X})H\big) + \ln\big(1/\delta\big)}$. With probability greater than $1-\delta$, Algorithm 1 uses at most $mH\widetilde{d}_m(\mathcal{X})$ trajectories and returns a hypothesis f such that $V^\star(s_0) - V^{\pi_f}(s_0)$ is at most:

$$12\sqrt{2}H^2 \frac{\sqrt{\widetilde{d}_m(\mathcal{X})}\cdot\sqrt{1+\ln\big(|\mathcal{H}|\big)}}{\sqrt{m}} \cdot v,$$

where $v = \sqrt{\ln\big(\widetilde{d}_m(\mathcal{X})H\big) + \ln\big(1/\delta\big)}$.

4.2.5. Finite Bellman Rank

In this section, we will prove sample complexity bounds for MDPs with finite Bellman Rank (Section 3.1).

Corollary 4.6 (Bellman Rank). *For a given MDP $\mathcal{M}$, suppose a hypothesis class $\mathcal{H}$ has* Bellman rank *d. Assume $\sup_{f\in\mathcal{H}_h,h\in[H]}\|W_h(f)\|_2 \le B_W$ and $\sup_{f\in\mathcal{H},h\in[H]}\|X_h(f)\| \le B_X$ for some $B_W, B_X \ge 1$. Fix $\delta\in(0,1/3)$ and $\epsilon \in (0,H)$. There exists an appropriate setting of batch sample size m, number of iteration T and confidence radius R such that with probability at least $1-\delta$, Algorithm 1 returns a hypothesis f such that $V^\star(s_0) - V^{\pi_f}(s_0) \le \epsilon$ using at most*

$$\widetilde{O}\left(\frac{d^2H^7|\mathcal{A}|(1+\ln(|\mathcal{H}|))}{\epsilon^2}\ln^3\Big(\frac{B_W^2B_X^2(1+\ln(|\mathcal{H}|))}{\delta}\Big)\right)$$

trajectories.

Note that in comparison, (Jiang et al., 2016) has sample complexity $\widetilde{O}(d^2H^5|\mathcal{A}|/\epsilon^2\log(1/\delta))$. We present the proof in Appendix D.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. In *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. *arXiv preprint arXiv:2006.10814*, 2020b.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. *arXiv:2006.01107*, 2020.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Provably efficient reinforcement learning with aggregated states, 2020.

Simon S Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, 2019a.

Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, 2019b.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020a.

Simon S Du, Jason D Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic Q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity. In *Advances in Neural Information Processing Systems*, 2020b.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *International Conference on Machine Learning*, 2015.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low bellman rank are pac-learnable, 2016.

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*, 2017.

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

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pages 740–747, 1999.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1848–1856, 2016.

Lihong Li. *A Unifying Framework for Computational Reinforcement Learning Theory*. PhD thesis, USA, 2009. AAI3386797.

Michael L Littman and Richard S Sutton. Predictive representations of state. In *Advances in Neural Information Processing Systems*, 2002.

Michael L Littman, Richard S Sutton, and Satinder P Singh. Predictive representations of state. In *NIPS*, volume 14, page 30, 2001.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *Conference on Artificial Intelligence and Statistics*, 2020a.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020b.

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

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*, 2019.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Gellert Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions, 2020.

Zheng Wen and Benjamin Van Roy. Efficient exploration and value function generalization in deterministic systems. In *Advances in Neural Information Processing Systems*, 2013.

Cathy Wu, Kanaad Parvate, Nishant Kheterpal, Leah Dickstein, Ankur Mehta, Eugene Vinitsky, and Alexandre M Bayen. Framework for control and deep reinforcement learning in traffic. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2017.

Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, 2019.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *arXiv preprint arXiv:2011.04622*, 2020.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error, 2020.

A. Related Work: Frameworks and Models

Relations Among Frameworks. We first review existing frameworks and the relations among them. See Table 1 for a summary.

Jiang et al. (2017) defines a notion, Bellman Rank (B-Rank in Tables), in terms of the roll-in distribution and the function approximation class for Q^*, and give an algorithm with a polynomial sample complexity in terms of the Bellman Rank. They also showed a class of models, including tabular MDP, LQR, Reactive POMDP (Krishnamurthy et al., 2016), and Reactive PSR (Littman and Sutton, 2002) admit a low Bellman Rank, and thus they can be solved efficiently. Some recently proposed models, such as Block MDP (Du et al., 2019a), linear MDP (Yang and Wang, 2019; Jin et al., 2020) can also be shown to have a low Bellman rank. One caveat is that their algorithm requires a finite number of actions, so cannot be directly applied to (infinite-action) linear MDP and LQR. Subsequently, Sun et al. (2019) proposed a new framework, Witness Rank (W-Rank in tables), which generalizes Bellman Rank to model-based setting.

Bellman Complete (B-Complete in tables) is a framework of another style, which assumes that the class used for approximating the Q-function is closed under the Bellman operator. As shown in Table 1, neither the low-rank-style framework (Bellman Rank and Witness Rank) nor the complete-style framework (B-Complete) contains the other (See e.g., (Zanette et al., 2020)).

Reinforcement Learning Models. Now we discuss existing RL models. A summary on whether a model can be incorporated into a framework is provided in Table 2.

Tabular MDP is the most basic model, which has a finite number of states and actions, and all frameworks incorporate this model. When the state-action space is large, different RL models have been proposed to study when one can generalize across the state-action pairs.

Reactive POMDP (Krishnamurthy et al., 2016) assumes there is a small number of hidden states and the Q^*-function belongs to a pre-specified function class. Block MDP (Du et al., 2019a) also assumes there is a small number of hidden states and further assumes the hidden states are decodable. Reactive PSR (Littman et al., 2001) considers partial observable systems whose parameters are grounded in observable quantities. FLAMBE (Agarwal et al., 2020b) considers the feature selection and removes the assumption of known feature in linear MDP. These models all admit a low-rank structure, and thus can be incorporated into the Bellman Rank or Witness Rank and our Bilinear Classes.

The Linear Bellman Complete model (Munos, 2005) uses linear functions to approximate the Q-function, and assumes the linear function class is closed under the Bellman operator. Zanette et al. (2020) presented a statistically efficient algorithm for this model. This model does not have a low Bellman Rank or Witness Rank but can be incorporated into the Bellman Complete framework and ours.

Linear MDP (Yang and Wang, 2019; Jin et al., 2020) assumes the transition probability and the reward are linear in given features. This model not only admits a low-rank structure, but also satisfies the complete condition. Therefore, this model belongs in all frameworks. However, when the number of action is infinite, the algorithms for Bellman Rank and Witness Rank are not applicable because their sample complexity scales with the number of actions. Linear mixture MDP (Modi et al., 2020a; Ayoub et al., 2020) assumes the transition probability is a linear mixture of some base models. This model cannot be included in Bellman Rank, Witness Rank, or Bellman Complete, but our Bilinear Classes includes this model.

LQR is a fundamental model for continuous control that can be efficiently solvable (Dean et al., 2019). While LQR has a low Bellman Rank and low Witness Rank, since the algorithms for Bellman Rank and Witness Rank scale with the number of actions and LQR's action set is uncountable, these two frameworks cannot incorporate LQR.

There is a line of work on state-action aggregation. Q^* "irrelevance" state aggregation assumes one can aggregate states to a meta-state if these states share the same Q^* value, and the number of meta-states is small (Li, 2009; Jiang et al., 2015). Q^* state-action aggregation aggregates state-action pairs to a meta-state-action pair if these pairs have the same Q^*-value (Dong et al., 2020; Li, 2009).

Lastly, when only assuming Q^* is linear, there exists an exponential lower bound (Weisz et al., 2020), but with the additional assumption that the MDP is (nearly) deterministic and has large sub-optimality gap, there exists sample efficient algorithms (Wen and Van Roy, 2013; Du et al., 2019b; 2020b).

B. Proofs for Section 3

B.1. Bellman Rank

Proof of Lemma 3.2. Note that for $g = f$, we have that for observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$

$$\mathbb{E}_{s_h \sim d^{\pi_f}} \mathbb{E}_{a_h \sim U(\mathcal{A})} \left[\ell(o_h, f)\right] = \mathbb{E}_{s_h, a_h, s_{h+1} \sim d^{\pi_f}} \left[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})\right]$$

Therefore, to prove that this is a Bilinear Class, we will show that a stronger "equality" version of Equation (2) holds (which will also prove Equation (1) holds). Observe that for any h,

$$\begin{aligned}
&\mathbb{E}_{s_h \sim d^{\pi_f}} \mathbb{E}_{a_h \sim U(\mathcal{A})} \left[\ell_f(o_h, g)\right] \\
&= \mathbb{E}_{s_h \sim d^{\pi_f}} \left[Q_{h,g}(s_h, \pi_g(s_h)) - r(s_h, \pi_g(s_h)) - \mathbb{E}\left[V_{h+1,g}(s_{h+1}) | s_h, \pi_g(s_h)\right]\right] \\
&= \langle W_h(g) - W_h(f^\star), X_h(f) \rangle
\end{aligned}$$

This completes the proof. □

B.2. Linear Mixture MDP.

Recall that for any hypothesis $g = \{\theta_0, \ldots, \theta_{H-1}\} \in \mathcal{H}$, $V_{h,g}$ and $Q_{h,g}$ satisfy the following Bellman optimality equation:

$$Q_{h,g}(s_h, a_h) = \theta_h^\top \left(\psi(s_h, a_h) + \sum_{\bar{s} \in \mathcal{S}} \phi(s_h, a_h, \bar{s}) V_{h+1,g}(\bar{s}) \right) \tag{6}$$

Proof of Lemma 3.3. Observe that for $g = f$, using Equation (6), for observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$,

$$\ell_f(o_h, f) = Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \,.$$

and therefore

$$\mathbb{E}_{o_h \sim d^{\pi_f}} \Big[\ell_f(o_h, f)\Big] = \mathbb{E}_{a_{0:h} \sim \pi_f} \Big[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})\Big] \,.$$

We consider on-policy estimation $\pi_{est} = \pi_f$. To prove that linear mixture MDP is a Bilinear Class, we only need to show that an "equality" version of Equation (2) holds (which implies Equation (1) holds by the frame above). For $g = \{\theta_0, \ldots, \theta_{H-1}\} \in \mathcal{H}$, observe:

$$\begin{aligned}
&\mathbb{E}_{o_h \sim d^{\pi_f}} \Big[\ell_f\left(o_h, g\right)\Big] \\
&= \mathbb{E}_{s_h, a_h \sim d^{\pi_f}} \left[\theta_h^\top \left(\psi(s_h, a_h) + \sum_{\bar{s} \in \mathcal{S}} \phi(s_h, a_h, \bar{s}) V_{h+1,f}(\bar{s})\right) - \mathbb{E}_{s_{h+1} \sim P_h(s_h, a_h)} \Big[V_{h+1,f}(s_{h+1}) + r_h\Big]\right] . \\
&= \mathbb{E}_{s_h, a_h \sim d^{\pi_f}} \left[(\theta_h - \theta_h^\star)^\top \left(\psi(s_h, a_h) + \sum_{\bar{s} \in \mathcal{S}} \phi(s_h, a_h, \bar{s}) V_{h+1,f}(\bar{s})\right)\right] \\
&= \langle W_h(g) - W_h(f^\star), X_h(f) \rangle
\end{aligned}$$

where we defined the W_h, X_h functions as follows:

$$\begin{aligned}
W_h(g) &= \theta_h, \\
X_h(f) &= \mathbb{E}_{s_h, a_h \sim d^{\pi_f}} \left[\psi(s_h, a_h) + \sum_{\bar{s} \in \mathcal{S}} \phi(s_h, a_h, \bar{s}) V_{h+1,f}(\bar{s})\right] .
\end{aligned}$$

This concludes that Linear Mixture Model also forms a Bilinear Class. □

B.3. Linear $Q^\star/V^\star$ (new model)

Proof of Lemma 3.4. Note that we will show that a stronger "equality" version of Equation (2) holds, which will also prove Equation (1) holds since for observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$,

$$\mathbb{E}_{o_h \sim d^{\pi_f}}\left[\ell_f(o_h, f)\right] = \mathbb{E}_{a_{0:h} \sim \pi_f}\left[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})\right].$$

Observe that for any h

$$\begin{aligned}
&\mathbb{E}_{o_h \sim d^{\pi_f}}\left[\ell(o_h, g)\right] \\
&= \mathbb{E}_{s_h, a_h, s_{h+1} \sim d^{\pi_f}}\left[w_h^\top \phi(s_h, a_h) - \theta_{h+1}^\top \psi(s_{h+1}) - Q_h^\star(s_h, a_h) + V_{h+1}^\star(s_{h+1})\right] \\
&= \langle W_h(g) - W_h(f^\star), X_h(f)\rangle
\end{aligned}$$

where

$$\begin{aligned}
W_h(g) &= [w_h, \theta_{h+1}], \\
X_h(f) &= \mathbb{E}_{s_h, a_h \sim d^{\pi_f}, s_{h+1} \sim P_h(s_h, a_h)}\left[\phi(s_h, a_h), \psi(s_{h+1})\right].
\end{aligned}$$

This concludes the proof. □

B.4. Bellman Complete and Linear MDPs

Proof of Lemma 3.5. Note that in this case, we will show that a stronger version of Equation (2) holds i.e with equality instead of $\leq$ inequality, which will also prove Equation (1) holds since for observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$,

$$\mathbb{E}_{o_h \sim d^{\pi_f}}\left[\ell_f(o_h, f)\right] = \mathbb{E}_{a_{0:h} \sim \pi_f}\left[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})\right].$$

Observe that for any h

$$\begin{aligned}
\mathbb{E}_{o_h \sim d^{\pi_f}}\left[\ell(o_h, g)\right] &= \mathbb{E}_{s_h, a_h \sim d^{\pi_f}}\left[\theta_h^\top \phi(s_h, a_h) - \mathcal{T}_h(\theta_{h+1})^\top \phi(s_h, a_h)\right] \\
&= \langle W_h(g) - W_h(f^\star), X_h(f)\rangle
\end{aligned}$$

where

$$\begin{aligned}
W_h(g) &= \theta_h - \mathcal{T}_h(\theta_{h+1}) \\
X_h(f) &= \mathbb{E}_{s_h, a_h \sim d^{\pi_f}}[\phi(s_h, a_h)].
\end{aligned}$$

Observe that $W_h(f^\star) = 0$ for all h. □

B.5. Low Occupancy Complexity (new model).

Proof of Lemma 3.6. To see why this is a Bilinear Class, as in previous proofs, we will show that an "equality" version of Equation (2) holds, which will also prove Equation (1) holds since

$$\mathbb{E}_{o_h \sim d^{\pi_f}}\left[\ell_f(o_h, f)\right] = \mathbb{E}_{a_{0:h} \sim \pi_f}\left[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})\right].$$

Observe that for any h (here observed transition info $o_h = (r_h, s_h, a_h, s_{h+1})$):

$$\begin{aligned}
&\mathbb{E}_{o_h \sim d^{\pi_f}}\left[\ell_f(o_h, g)\right] \\
&= \sum_{(s_h, a_h) \in \mathcal{S} \times \mathcal{A}} d^{\pi_f}(s_h, a_h)\big(Q_{h,g}(s_h, a_h) - r(s_h, a_h) - \mathbb{E}[V_{h+1,g}(s_{h+1}) | s_h, a_h]\big) \\
&= \Big\langle \beta_h(f), \sum_{(s_h, a_h) \in \mathcal{S} \times \mathcal{A}} \phi_h(s_h, a_h)\big(Q_{h,g}(s_h, a_h) - r(s_h, a_h) - \mathbb{E}[V_{h+1,g}(s_{h+1}) | s_h, a_h]\big)\Big\rangle \\
&= \langle W_h(g) - W_h(f^\star), X_h(f)\rangle
\end{aligned}$$

where the notation $\mathbb{E}[V(s_{h+1})|s_h, a_h]$ is shorthand for $\mathbb{E}_{s_{h+1}\sim P_h(s_h,a_h)}[V(s_{h+1})]$ and we defined the W_h, X_h functions as follows:

$$X_h(f) := \beta_h(f),$$
$$W_h(g) := \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \phi_h(s,a)\big(Q_{h,g}(s,a) - r(s,a) - \mathbb{E}_{s'\sim P_h(s,a)}[V_{h+1,g}(s')]\big).$$

Note that $W_h(f^\star) = 0$. This completes the proof. □

Note that as such the hypothesis class $\mathcal{H}$ could be arbitrary and unlike other models where we assume linearity, here it could be a neural state-action value class. Our model can also capture the setting where the state-only occupancy has low complexity, i.e., $d^{\pi_f}(s_h) = \beta_h(f)\mu_h(s_h)$, for some $\mu_h : \mathcal{S} \to \mathcal{V}$. In this case, we will use $\pi_{est} = U(\mathcal{A})$.

C. Proof of Theorem 4.1 and Theorem 4.2

In this section, we prove our main theorems – Theorem 4.1 and Theorem 4.2.

Notation To simplify notation, we denote by $\mu_{t;h}$ the distribution induced over $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ by $a_{0:h-1} \sim d^{\pi_{f_t}}$ and $a_h \sim \pi_{est}$; $\mathcal{D}_{t;h}$ the batch dataset collected from distribution $\mu_{t;h}$; ε_{gen} the *generalization error* $\varepsilon_{\text{gen}}(m, \mathcal{H}) \cdot \text{conf}(\delta/(TH))$. Also, recall that for any distribution μ over $\mathbb{R}\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and hypothesis $f, g \in \mathcal{H}$

$$\mathcal{L}_{\mu,f}(g) = \mathbb{E}_{o\sim\mu}[\ell_f(o, g)]$$

Note that throughout the proof unless specified, the statements are true for any fixed $\delta \in (0, 1)$, integer $m > 0$ and integer $T > 0$. Also, we set $R = \sqrt{T}\varepsilon_{\text{gen}}$ throughout the proof. To simplify the proof, we will condition on the event that uniform convergence of ℓ holds throughout our algorithm, which we first show holds with high probability.

Lemma C.1 (Uniform Convergence). *For all $t \in [T]$ and $g \in \mathcal{H}$ and $h \in [H]$, with probability at least $1-\delta$, we have:*

$$\left|\mathcal{L}_{\mathcal{D}_{t;h},f_t}(g) - \mathcal{L}_{\mu_{t;h},f_t}(g)\right| \le \varepsilon_{gen}$$

Proof. This follows from the uniform convergence (Assumption 4.1) and then union bounding over all $t \in [T]$ and $h \in [H]$. □

We start by presenting our main lemma which shows if uniform convergence of ℓ holds throughout our algorithm, our algorithm finds a near-optimal policy. This lemma will be enough to prove our main results.

Lemma C.2 (Existence of high quality policy). *Suppose we run the algorithm for T iterations. Set $R = \sqrt{T}\varepsilon_{gen}$. Assume the event in Lemma C.1 holds and $\sup_{f\in\mathcal{H}}\|W_h(f)\|_2 \le B_W$ for all $h \in [H]$. Then, for all $\lambda \in \mathbb{R}^+$, there exists $t \in [T]$ such that the following is true for hypothesis f_t:*

$$V^\star - V^{\pi_{f_t}}(s_0) \le H\sqrt{(4\lambda B_W^2 + 4T\varepsilon_{gen}^2)\left(\exp\left(\frac{1}{T}\gamma_T(\lambda;\mathcal{X})\right) - 1\right)}$$

We now complete the proof of Theorem 4.1 and Theorem 4.2 using Lemma C.1, Lemma C.2 and setting the parameters using the definition of critical information gain.

Proof of Theorem 4.1 and Theorem 4.2. Fix $\lambda = \varepsilon_{\text{gen}}^2(m,\mathcal{H})/B_W^2$. From definition of critical information gain (Equation (5)), it follows that for $T = \widetilde{\gamma}(\lambda, \mathcal{X})$,

$$T \ge \gamma_T(\lambda, \mathcal{X})$$

Using Lemma C.2, we get that

$$V^\star - V^{\pi_{f_t}}(s_0) \le H\sqrt{\left(4\lambda B_W^2 + 4T\varepsilon_{\text{gen}}^2(m,\mathcal{H})\cdot\text{conf}^2(\delta/TH)\right)\left(\exp\left(\frac{1}{T}\gamma_T(\lambda;\mathcal{X})\right) - 1\right)}$$

Observing that for our choice of T, $\gamma_T(\lambda;\mathcal{X})/T \leq 1$ and $e-1<2$, we get

$$
\begin{aligned}
V^\star - V^{\pi_{f_t}}(s_0) &\leq \sqrt{8}H\sqrt{\left(\lambda B_W^2 + \widetilde{\gamma}(\lambda,\mathcal{X})\varepsilon_{\text{gen}}^2(m,\mathcal{H})\cdot \text{conf}^2(\delta/TH)\right)} \\
&\leq \sqrt{8}H\left(\sqrt{\lambda}B_W + \sqrt{\widetilde{\gamma}(\lambda,\mathcal{X})}\varepsilon_{\text{gen}}(m,\mathcal{H})\cdot \text{conf}(\frac{\delta}{\widetilde{\gamma}(\lambda,\mathcal{X})H})\right) \\
&= \sqrt{8}H\left(1 + \sqrt{\widetilde{\gamma}(\lambda,\mathcal{X})}\cdot \text{conf}(\frac{\delta}{\widetilde{\gamma}(\lambda,\mathcal{X})H})\right)\cdot \varepsilon_{\text{gen}}(m,\mathcal{H}) \\
&\leq 3H\left(1 + \sqrt{\widetilde{\gamma}(\lambda,\mathcal{X})}\cdot \text{conf}(\frac{\delta}{\widetilde{\gamma}(\lambda,\mathcal{X})H})\right)\cdot \varepsilon_{\text{gen}}(m,\mathcal{H})
\end{aligned}
$$

where the second last equality uses the definition of λ.

Moreover, each iteration of the algorithm, takes only mH trajectories, this gives the total trajectories as $mHT = mH\widetilde{\gamma}(\lambda,\mathcal{X})$. This proves Theorem 4.2. Theorem 4.1 follows from the upper bound on $\widetilde{\gamma}(\lambda,\mathcal{X})$ for finite dimensional $\mathcal{X}_h$ using Lemma G.3. □

In the rest of the section, we will prove our main lemma – Lemma C.2. The first step shows that under Assumption 4.1, our R is set properly so that $f^\star$ is always a feasible solution of the constrained optimization program in Algorithm 1.

Lemma C.3 (Feasibility of $f^\star$). *Assume the event in Lemma C.1 holds. Then for all $t \in [T]$, we have that $f^\star$ is always a feasible solution.*

Proof. Note that $\mathcal{L}_{\mu_{i;h},f_i}(f^*) = 0$ (Equation (2)). Thus using Lemma C.1, we have:

$$
\sum_{i=0}^{t-1}\left(\mathcal{L}_{\mathcal{D}_{i;h},f_i}(f^*)\right)^2 \leq t\varepsilon_{\text{gen}}^2 \qquad \forall h\in[H]\,.
$$

Noting that $t \leq T$ and in our parameter setup $R = \sqrt{T}\varepsilon_{\text{gen}}$ completes the proof. □

The feasibility result immediately leads to optimism.

Lemma C.4 (Optimism). *Assume the event in Lemma C.1 holds. Then for all $t \in [T]$, we have $V^\star \leq V_{f_t;0}(s_0)$.*

Proof. Lemma C.3 implies $f^\star$ is a feasible solution for the optimization program for all $t \in [T]$. This proves the claim. □

The following lemma relates the sub-optimality to a sum of bilinear forms. Using the performance difference lemma, we first show that sub-optimality is upper bounded by the Bellman errors of Q_{h,f_t}, which are further upper bounded by sum of bilinear forms via our assumption (Equation (1)).

Lemma C.5 (Bilinear Regret Lemma). *Assume the event in Lemma C.1 holds. Then, the following holds for all $t \in [T]$:*

$$
V^\star - V^{\pi_{f_t}}(s_0) \leq \sum_{h=0}^{H-1}\left|\langle W_h(f_t) - W_h(f^\star), X_h(f_t)\rangle\right|.
$$

Proof. We can upper bound the regret

$$
\begin{aligned}
& V^\star(s_0) - V^{\pi_{f_t}}(s_0) \\
& \leq V_{0,f_t}(s_0) - V^{\pi_{f_t}}(s_0) && \text{(since } V_{0,f_t}(s_0) \geq V^\star(s_0) \text{ (Lemma C.4))} \\
& = Q_{0,f_t}(s_0, a_0) - \mathbb{E}_{a_{0:h} \sim d^{\pi_{f_t}}} \left[\sum_{h=0}^{H-1} r(s_h, a_h) \right] && \text{(since } V_{f_t}(s_0) = Q_{f_t}(s_0, a_0),\ a_0 = \text{argmax}_a Q_{f_t}(s_0, a)) \\
& = \mathbb{E}_{a_{0:h} \sim d^{\pi_{f_t}}} \left[\sum_{h=0}^{H-1} \left(Q_{h,f_t}(s_h, a_h) - r(s_h, a_h) - Q_{h+1,f_t}(s_{h+1}, a_{h+1}) \right) \right] && \text{(by telescoping sum)} \\
& = \sum_{h=0}^{H-1} \mathbb{E}_{a_{0:h} \sim d^{\pi_{f_t}}} \left[Q_{h,f_t}(s_h, a_h) - r(s_h, a_h) - Q_{h+1,f_t}(s_{h+1}, a_{h+1}) \right] \\
& = \sum_{h=0}^{H-1} \mathbb{E}_{a_{0:h} \sim d^{\pi_{f_t}}} \left[Q_{h,f_t}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f_t}(s_{h+1}) \right] && \text{(since } V_{h+1,f_t}(s_{h+1}) = Q_{h+1,f_t}(s_{h+1}, a_{h+1})) \\
& = \sum_{h=0}^{H-1} |\langle W_h(f_t) - W_h(f^\star), X_h(f_t) \rangle|
\end{aligned}
$$

where the last step follows Equation (1) in the Bilinear Class definition. □

The following is a variant of the Elliptical Potential Lemma, central in the analysis of linear bandits (Dani et al., 2008; Srinivas et al., 2009; Abbasi-Yadkori et al., 2011).

Lemma C.6 (Elliptical potential). *Consider any sequence of vectors $\{x_0, \ldots, x_{T-1}\}$ where $x_i \in \mathcal{V}$ for some Hilbert space $\mathcal{V}$. Let $\lambda \in \mathbb{R}^+$. Denote $\Sigma_0 = \lambda I$ and $\Sigma_t = \Sigma_0 + \sum_{i=0}^{t-1} x_i x_i^\top$. We have that:*

$$
\min_{i \in [T]} \ln \left(1 + \|x_i\|^2_{\Sigma_i^{-1}} \right) \leq \frac{1}{T} \sum_{i=0}^{T-1} \ln \left(1 + \|x_i\|^2_{\Sigma_i^{-1}} \right) = \frac{1}{T} \ln \frac{\det(\Sigma_T)}{\det(\lambda I)}.
$$

Proof. By definition of Σ_t and matrix determinant lemma, we have:

$$
\begin{aligned}
\ln \det(\Sigma_{t+1}) &= \ln \det(\Sigma_t) + \ln \det \left(I + (\Sigma_t)^{-1/2} x_t x_t^\top (\Sigma_t)^{-1/2} \right) \\
&= \ln \det(\Sigma_t) + \ln \left(1 + \|x_t\|^2_{\Sigma_t^{-1}} \right).
\end{aligned}
$$

Using recursion completes the proof. □

Now, we will finish the proof of Lemma C.2 by showing that the sum of bilinear forms in Lemma C.5 is small for at least for one $t \in [T]$. More precisely, using Equation (2) together with elliptical potential argument (Lemma C.6), we can show that after $\widetilde{d}_m$ many iterations, we must have found a policy π_{f_t} such that $|\langle W_h(f_t) - W_h(f^\star), X_h(f_t) \rangle|$ is small for all h.

Proof of Lemma C.2. Our goal (as per Lemma C.5 and Equation (1)) is to find $t \in [T]$ such that

$$
|\langle W_h(f_t) - W_h(f^\star), X_h(f_t) \rangle| \quad \text{is small for all } h \in [H]
$$

To that end, we will show that

$$
\|W_h(f_t) - W_h(f^\star)\|_A \quad \|X_h(f_t)\|_{A^{-1}} \quad \text{is small for all } h \in [H]
$$

for appropriately chosen A. We will show existence of such $X_h(f_t)$ and A (Equation (7)) using the potential argument (Lemma C.6) and conditions on $W_h(f_t) - W_h(f^\star)$ follow from our optimization program. We now show this in more detail.

Let the hypothesis used by our algorithm at ith iteration be f_i. Consider the corresponding sequence of representations $\{X_h(f_i)\}_{i,h}$. Then, by Lemma C.6, we have that for all $h \in [H]$ and $\lambda \in \mathbb{R}^+$

$$\sum_{i=0}^{T-1} \ln\left(1 + \|X_h(f_i)\|^2_{\Sigma^{-1}_{i;h}}\right) \le \ln \frac{\det\left(\Sigma_{T;h}\right)}{\det(\lambda \mathrm{I})} \le \gamma_T(\lambda; \mathcal{X}_h)$$

where we have used definition of maximum information gain $\gamma_T(\lambda; \mathcal{X}_h)$ (Equation (3)) and

$$\Sigma_{i;h} = \lambda \mathrm{I} + \sum_{j=0}^{i-1} X_h(f_j) X_h(f_j)^\top$$

Summing these inequalities over all $h \in [H]$, we have that for all $\lambda \in \mathbb{R}^+$

$$\sum_{i=0}^{T-1} \sum_{h=0}^{H-1} \ln\left(1 + \|X_h(f_i)\|^2_{\Sigma^{-1}_{i;h}}\right) \le \sum_{h=0}^{H-1} \gamma_T(\lambda; \mathcal{X}_h) = \gamma_T(\lambda; \mathcal{X})$$

where the last equality follows from Equation (4). Since, each of these terms is ≥ 0, we get that there exists $t \in [T]$ such that

$$\sum_{h=0}^{H-1} \ln\left(1 + \|X_h(f_t)\|^2_{\Sigma^{-1}_{t;h}}\right) \le \frac{1}{T} \gamma_T(\lambda; \mathcal{X})$$

Again, since each of these terms is ≥ 0, we get that for all $h \in [H]$

$$\ln\left(1 + \|X_h(f_t)\|^2_{\Sigma^{-1}_{t;h}}\right) \le \frac{1}{T} \gamma_T(\lambda; \mathcal{X})$$

and simplifying, we get that for all $h \in [H]$,

$$\|X_h(f_t)\|^2_{\Sigma^{-1}_{t;h}} \le \exp\left(\frac{1}{T} \gamma_T(\lambda; \mathcal{X})\right) - 1 \tag{7}$$

Also, by construction of our program, for all iterations and in particular for t, it holds that for all $h \in [H]$

$$\sum_{j=0}^{t-1} \left(\mathcal{L}_{\mathcal{D}_{j;h}, f_j}(f_t)\right)^2 \le T \varepsilon^2_{\text{gen}}$$

and by Lemma C.1, for all $h \in [H]$

$$\begin{aligned} \sum_{j=0}^{t-1} \left(\mathcal{L}_{\mu_{j;h}, f_j}(f_t)\right)^2 &\le 2 \sum_{j=0}^{t-1} \left(\mathcal{L}_{\mathcal{D}_{j;h}, f_j}(f_t)\right)^2 + 2 \sum_{j=0}^{t-1} \varepsilon^2_{\text{gen}} \\ &\le 4T \varepsilon^2_{\text{gen}} \end{aligned}$$

where the first inequality follows from $(a+b)^2 \le 2a^2 + 2b^2$ and the last step follows from the frame above and $t \in [T]$. Using the definition of Bilinear Class (Equation (2)), for all $h \in [H]$

$$\sum_{j=0}^{t-1} |\langle W_h(f_t) - W_h(f^\star), X_h(f_j) \rangle|^2 \le 4T \varepsilon^2_{\text{gen}}$$

Using this, we get for all $h \in [H]$

$$\begin{aligned} &(W_h(f_t) - W_h(f^\star))^\top \Sigma_{t;h} (W_h(f_t) - W_h(f^\star)) \\ &\le \lambda \|(W_h(f_t) - W_h(f^\star))\|_2^2 + 4T \varepsilon^2_{\text{gen}} \\ &\le 4 \lambda B_W^2 + 4T \varepsilon^2_{\text{gen}} \end{aligned} \tag{8}$$

where the first inequality follows from the frame above and definition of $\Sigma_{t;h}$. Using Equation (7) and the frame above, this immediately shows that for all $h \in [H]$

$$\begin{aligned} |\langle W_h(f_t) - W_h(f^\star), X_h(f_t)\rangle|^2 &\leq \|W_h(f_t) - W_h(f^\star)\|^2_{\Sigma_{t;h}} \|X_h(f_t)\|^2_{\Sigma^{-1}_{t;h}} \\ &\leq (4\lambda B_W^2 + 4T\varepsilon^2_{\text{gen}}) \left(\exp\left(\frac{1}{T}\gamma_T(\lambda;\mathcal{X})\right) - 1\right) \end{aligned}$$

Summing over all $h \in [H]$, this gives

$$\sum_{h=0}^{H-1} |\langle W_h(f_t) - W_h(f^\star), X_h(f_t)\rangle| \leq H\sqrt{(4\lambda B_W^2 + 4T\varepsilon^2_{\text{gen}}) \left(\exp\left(\frac{1}{T}\gamma_T(\lambda;\mathcal{X})\right) - 1\right)}$$

Using Lemma C.5, this gives the desired result. □

D. Proofs for Section 4

Proof of Corollary 4.1. First, using Lemma G.1, we get that for any distribution μ over $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and for any $\delta \in (0,1)$, with probability of at least $1-\delta$ over choice of an i.i.d. sample $\mathcal{D} \sim \mu^m$ of size m, for all $g \in \mathcal{H}$

$$\begin{aligned} |\mathcal{L}_{\mathcal{D}}(g) - \mathcal{L}_\mu(g)| &\leq 2\sqrt{2}H\sqrt{\frac{\ln(|\mathcal{H}|/\delta)}{m}} \\ &\leq 2\sqrt{2}H\sqrt{\frac{\ln(|e\mathcal{H}|/\delta)}{m}} \\ &= 2\sqrt{2}H\sqrt{\frac{1+\ln(|\mathcal{H}|)+\ln(1/\delta)}{m}} \\ &\leq 2\sqrt{2}H\sqrt{\frac{1+\ln(|\mathcal{H}|)}{m}} \cdot \sqrt{\ln(1/\delta)} \end{aligned}$$

This satisfies our Assumption 4.1 with

$$\begin{aligned} \varepsilon_{\text{gen}}(m,\mathcal{H}) &= 2\sqrt{2}H\sqrt{\frac{1+\ln(|\mathcal{H}|)}{m}} \\ \text{conf}(\delta) &= \sqrt{\ln(1/\delta)} \end{aligned}$$

Using this in Theorem 4.1, we set

$$T = 4dH\ln\left(1 + 3B_X^2 B_W^2\sqrt{m}\right)$$

Therefore, we get ϵ-optimal policy by setting

$$3H \cdot 2\sqrt{2}H\sqrt{\frac{1+\ln(|\mathcal{H}|)}{m}} \cdot \left(1 + \sqrt{4dH\ln\left(1+3B_X^2B_W^2\sqrt{m}\right)}\right) \cdot \sqrt{\ln\frac{4dH^2\ln\left(1+3B_X^2B_W^2\sqrt{m}\right)}{\delta}} \leq \epsilon$$

or equivalently by setting m at least as large as

$$\begin{aligned} &\frac{720dH^5(1+\ln(|\mathcal{H}|))\ln(1+3B_X^2B_W^2\sqrt{m})}{\epsilon^2} \cdot \ln\frac{4dH^2\ln\left(1+3B_X^2B_W^2\sqrt{m}\right)}{\delta} \\ &\leq \frac{720dH^5\ln(4dH^2)(1+\ln(|\mathcal{H}|))\ln^2(1+3B_X^2B_W^2\sqrt{m})\ln(1/\delta)}{\epsilon^2} \end{aligned}$$

Using Lemma G.2, we get a solution for m

$$m = \frac{6480dH^5 \ln(4dH^2)\ln(1/\delta)(1+\ln(|\mathcal{H}|))}{\epsilon^2} \ln\Big(\frac{25920dH^5 B_X^2 B_W^2 (1+\ln(|\mathcal{H}|))\ln(4dH^2)\ln(1/\delta)}{\epsilon^2}\Big)$$

This gives the total trajectory complexity

$$mTH = \frac{cd^2H^7 \ln(dH^2)\ln(1/\delta)(1+\ln(|\mathcal{H}|))}{\epsilon^2} \ln^2\big(\frac{dHB_XB_W(1+\ln(|\mathcal{H}|))\ln(1/\delta)}{\epsilon^2}\big)$$

for some absolute constants c. □

D.1. Linear $Q^\star/V^\star$

Corollary D.1 (RKHS Linear $Q^\star/V^\star$). *Suppose MDP $\mathcal{M}$ is a linear $Q^\star/V^\star$ model. Assume $\sup_{(w,\theta)\in\mathcal{H}_h, h\in[H]} \|[w,\theta]\|_2 \le B_W$ and $\sup_{x\in\Phi\circ\Psi}\|x\|_2 \le B_X$. Fix $\delta \in (0,1/3)$, batch sample size m, and define:*

$$\widetilde{d}_m(\Phi\circ\Psi) = \widetilde{\gamma}\Big(\frac{1}{8B_W^2 m}; \Phi\circ\Psi\Big)\cdot\nu, \tag{9}$$

$$\widetilde{d}_m(\mathcal{X}) = \widetilde{\gamma}\left(\frac{144H^2\widetilde{d}_m(\Phi\circ\Psi)}{B_W^2 m}; \mathcal{X}\right), \tag{10}$$

where $\nu := \ln\left(1 + 3B_XB_W\sqrt{m\widetilde{\gamma}\Big(\frac{1}{8B_W^2 m}; \Phi\circ\Psi\Big)}\right)$.

Set the parameters as: $R = (12H/\sqrt{m})\sqrt{\widetilde{d}_m(\mathcal{X})\cdot\widetilde{d}_m(\Phi\circ\Psi)}\cdot\sqrt{\ln\big((\widetilde{d}_m(\mathcal{X})H)/\delta\big)}$ and $T = \widetilde{d}_m(\mathcal{X})$. With probability greater than $1-\delta$, $Algorithm$ 1 uses at most $mH\widetilde{d}_m(\mathcal{X})$ trajectories and returns a hypothesis f:

$$V^\star(s_0) - V^{\pi_f}(s_0) \le 72H^2\frac{\sqrt{\widetilde{d}_m(\mathcal{X})\cdot\widetilde{d}_m(\Phi\circ\Psi)}\cdot v}{\sqrt{m}}, \tag{11}$$

where $v := \sqrt{\ln\Big((\widetilde{d}_m(\mathcal{X})H)/\delta\Big)}$.

Proof. First, using Corollary F.3, we get that for any distribution μ over $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and for any $\delta\in(0,1)$, with probability of at least $1-\delta$ over choice of an i.i.d. sample $\mathcal{D}\sim\mu^m$ of size m, for all $g = ([w_0,\theta_0],\ldots,[w_{H-1},\theta_{H-1}])\in\mathcal{H}$ (note that $\mathcal{L}_\mu(g)$ only depends on $[w_h,\theta_h]$ for distribution μ over observed transitions $o_h = (r_h, s_h, a_h, s_{h+1})$ at timestep h.)

$$\begin{aligned}
|\mathcal{L}_\mathcal{D}(g) - \mathcal{L}_\mu(g)| &\le \frac{4}{\sqrt{m}} + 2H\sqrt{\frac{2\widetilde{\gamma}_m \ln\Big(1+3B_XB_W\sqrt{\widetilde{\gamma}_m m}\Big) + 2\ln(1/\delta)}{m}} \\
&= \frac{4 + 2H\sqrt{2\widetilde{\gamma}_m \ln\Big(1+3B_XB_W\sqrt{\widetilde{\gamma}_m m}\Big) + 2\ln(1/\delta)}}{\sqrt{m}} \\
&\le \frac{12H\sqrt{\widetilde{\gamma}_m \ln\Big(1+3B_XB_W\sqrt{\widetilde{\gamma}_m m}\Big)}\cdot\sqrt{\ln(1/\delta)}}{\sqrt{m}}
\end{aligned}$$

where we have used that $\ln(1/\delta) > 1$ and $\widetilde{\gamma}_m = \widetilde{\gamma}(1/(8B_W^2 m); \Phi\circ\Psi)$ (as defined in Equation (5)). Define

$$\widetilde{d}_m(\Phi\circ\Psi) := \widetilde{\gamma}_m \ln\Big(1+3B_XB_W\sqrt{\widetilde{\gamma}_m m}\Big)$$

This satisfies our Assumption 4.1 with

$$\varepsilon_{\text{gen}}(m, \mathcal{H}) = \frac{12H\sqrt{\widetilde{d}_m(\Phi \circ \Psi)}}{\sqrt{m}}$$
$$\text{conf}(\delta) = \sqrt{\ln(1/\delta)}$$

Substituting this in Theorem 4.2 gives the result

$$\begin{aligned}
\widetilde{d}_m(\mathcal{X}) &= \widetilde{\gamma}\Big(\varepsilon^2_{\text{gen}}(m, \mathcal{H})/B_W^2; \mathcal{X}\Big) \\
&= \widetilde{\gamma}\Big(144H^2\widetilde{d}_m(\Phi \circ \Psi)/mB_W^2; \mathcal{X}\Big) \\
V^\star(s_0) - V^{\pi_{f_t}}(s_0) &\le 6H\sqrt{\widetilde{d}_m(\mathcal{X})} \cdot \varepsilon_{\text{gen}}(m, \mathcal{H}) \cdot \text{conf}\big(\delta/(\widetilde{d}_m(\mathcal{X})H)\big) \\
&= 72H^2 \frac{\sqrt{\widetilde{d}_m(\mathcal{X}) \cdot \widetilde{d}_m(\Phi \circ \Psi)} \cdot \sqrt{\ln\big((\widetilde{d}_m(\mathcal{X})H)/\delta\big)}}{\sqrt{m}}
\end{aligned}$$

□

Next, we complete the proof of Corollary 4.2. Note that both $\widetilde{d}_m(\Phi \circ \Psi)$ and $\widetilde{d}_m(\mathcal{X})$ (related to critical information gain under Φ and $\mathcal{X}$ respectively) scale as $\widetilde{O}(d)$ if $\Phi \circ \Psi \subset \mathbb{R}^d$.

Proof of Corollary 4.2. First, from Lemma G.3, we have that

$$\begin{aligned}
\widetilde{\gamma}\Big(\frac{1}{8B_W^2 m}; \Phi \circ \Psi\Big) &\le 3d \ln\Big(1 + 24B_X^2 B_W^2 m\Big) + 1 \\
&\le 3d \ln\Big(25B_X^2 B_W^2 m\Big) + 1 \\
&\le 4d \ln\Big(25B_X^2 B_W^2 m\Big)
\end{aligned}$$

and substituting this in Equation (9)

$$\begin{aligned}
\widetilde{d}_m(\Phi \circ \Psi) &\le 4d \ln\Big(25B_X^2 B_W^2 m\Big) \cdot \ln\left(1 + 3B_X B_W \sqrt{m4d\ln\Big(25B_X^2 B_W^2 m\Big)}\right) \\
&\le 4d \ln\Big(25B_X^2 B_W^2 m\Big) \cdot \ln\left(4B_X B_W \sqrt{m4d\ln\Big(25B_X^2 B_W^2 m\Big)}\right) \\
&\le 4d \ln\Big(25B_X^2 B_W^2 m\Big) \cdot \Big(\ln(4B_X B_W) + \ln\Big(10m\sqrt{d}B_X B_W\Big)\Big) \\
&\le 8d \ln^2(25B_X^2 B_W^2 m\sqrt{d})
\end{aligned}$$

Similarly, as $\sup_{z\in\mathcal{X}}\|z\| \le \sup_{x\in\Phi\circ\Psi}\|x\|$, using Lemma G.3 and similar analysis as above (and $144H^2\widetilde{d}_m(\Phi \circ \Psi) \ge 1$), we get

$$\widetilde{\gamma}\left(\frac{144H^2\widetilde{d}_m(\Phi \circ \Psi)}{B_W^2 m}; \mathcal{X}_h\right) \le 4d \ln\Big(25B_X^2 B_W^2 m\Big)$$

and substituting this in Equation (10)

$$\widetilde{d}_m(\mathcal{X}) \le 4dH \ln\Big(4B_X^2 B_W^2 m\Big)$$

To get ϵ-optimal policy (from Equation (11)), we have to set

$$72H^2\frac{\sqrt{\widetilde{d}_m(\mathcal{X})\cdot\widetilde{d}_m(\Phi\circ\Psi)}\cdot\sqrt{\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}}{\sqrt{m}}\leq\epsilon$$
$$m\geq(72)^2H^4\frac{\widetilde{d}_m(\mathcal{X})\cdot\widetilde{d}_m(\Phi\circ\Psi)\cdot\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}{\epsilon^2}$$

Further upper bounding the right hand side of the above inequality by substituting in upper bounds for $\widetilde{d}_m(\mathcal{X})$ and $\widetilde{d}_m(\Phi\circ\Psi)$ from frames above, we can set m to be as large as:

$$(72)^2H^5\frac{32d^2\ln^3(25B_X^2B_W^2m\sqrt{d})\cdot\ln\left((16dH^2B_X^2B_W^2m)/\delta\right)}{\epsilon^2}$$
$$\leq 32\cdot(72)^2\frac{d^2H^5\ln^4(25B_X^2B_W^2mdH^2)\ln(1/\delta)}{\epsilon^2}$$

Using Lemma G.2 for $\alpha=4$, $a=32\cdot(72)^2d^2H^5\ln(1/\delta)/\epsilon^2$, $b=25B_X^2B_W^2dH^2$ and $c=5^4$, we get that

$$m=5^4\cdot 32\cdot(72)^2\frac{d^2H^5\ln(1/\delta)}{\epsilon^2}\ln^4\left(5^4\cdot 25\cdot 32\cdot(72)^2\frac{d^3H^7B_X^2B_W^2\ln(1/\delta)}{\epsilon^2}\right)$$
$$\ln\left(4B_X^2B_W^2m\right)\leq 5\ln\left(5^6\cdot 32\cdot(72)^2\frac{d^3H^7\ln(1/\delta)B_X^2B_W^2}{\epsilon^2}\right)$$

Substituting this in the expression above for $\widetilde{d}_m(\mathcal{X})$ and setting this upper bound to T, we get

$$T=20dH\ln\left(5^6\cdot 32\cdot(72)^2\frac{d^3H^7\ln(1/\delta)B_X^2B_W^2}{\epsilon^2}\right)$$

Since, we use on policy estimation, i.e., $\pi_{est}=\pi_{f_t}$ for all t, the trajectory complexity is mT which completes the proof. □

D.2. Bellman Complete

Corollary D.2 (RKHS Bellman Complete). *Suppose $\mathcal{H}$ is* Bellman Complete *with respect to MDP $\mathcal{M}$ for some Hilbert space $\mathcal{V}$. Assume $\sup_{h\in[H],\theta\in\mathcal{H}_h}\|\theta\|_2\leq B_W$ and $\sup_{x\in\Phi}\|x\|_2\leq B_X$. Fix $\delta\in(0,1/3)$, batch sample size m, and define:*

$$\widetilde{d}_m(\Phi)=\widetilde{\gamma}\Big(\frac{1}{8B_W^2m};\Phi\Big)\cdot\nu,$$
$$\widetilde{d}_m(\mathcal{X})=\widetilde{\gamma}\Big(\frac{400H^2d_m(\Phi)}{B_W^2m};\mathcal{X}\Big),$$

where $\nu=\ln\left(1+3B_XB_W\sqrt{m\widetilde{\gamma}\Big(\frac{1}{8B_W^2m};\Phi\Big)}\right)$.

Set the parameters as: $R=(12H/\sqrt{m})\sqrt{\widetilde{d}_m(\mathcal{X})\cdot\widetilde{d}_m(\Phi)}\cdot\sqrt{\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}$ and $T=\widetilde{d}_m(\mathcal{X})$. With probability at least $1-\delta$, $Algorithm$ 1 uses at most $mH\widetilde{d}_m(\mathcal{X})$ trajectories and returns a hypothesis f:

$$V^\star(s_0)-V^{\pi_f}(s_0)\leq 120H^2\frac{\sqrt{\widetilde{d}_m(\mathcal{X})\cdot\widetilde{d}_m(\Phi)}\cdot v}{\sqrt{m}},$$

where $v=\sqrt{\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}$.

Proof. First, using Corollary F.2, we get that for any distribution μ over $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and for any $\delta\in(0,1)$, with probability of at least $1-\delta$ over choice of an i.i.d. sample $\mathcal{D}\sim\mu^m$ of size m, for all $g=(\theta_0,\ldots,\theta_{H-1})\in\mathcal{H}$ (note that $\mathcal{L}_\mu(g)$ only depends on θ_h for distribution μ over observed transitions $o_h = (r_h, s_h, a_h, s_{h+1})$ at timestep h.)

$$
\begin{aligned}
|\mathcal{L}_{\mathcal{D}}(g) - \mathcal{L}_\mu(g)| &\leq \frac{8}{\sqrt{m}} + 2H\sqrt{\frac{2\widetilde{\gamma}_m \ln\left(1 + 3B_X B_W \sqrt{\widetilde{\gamma}_m m}\right) + 2\ln(1/\delta)}{m}} \\
&= \frac{8 + 2H\sqrt{2\widetilde{\gamma}_m \ln\left(1 + 3B_X B_W \sqrt{\widetilde{\gamma}_m m}\right) + 2\ln(1/\delta)}}{\sqrt{m}} \\
&\leq \frac{20H\sqrt{\widetilde{\gamma}_m \ln\left(1 + 3B_X B_W \sqrt{\widetilde{\gamma}_m m}\right)} \cdot \sqrt{\ln(1/\delta)}}{\sqrt{m}}
\end{aligned}
$$

where we have used that $\ln(1/\delta) > 1$ and $\widetilde{\gamma}_m = \widetilde{\gamma}(1/(8B_W^2 m); \Phi)$ (as defined in Equation (5)). Define

$$
\widetilde{d}_m(\Phi) := \widetilde{\gamma}_m \ln\left(1 + 3B_X B_W \sqrt{\widetilde{\gamma}_m m}\right)
$$

This satisfies our Assumption 4.1 with

$$
\begin{aligned}
\varepsilon_{\text{gen}}(m, \mathcal{H}) &= \frac{20H\sqrt{\widetilde{d}_m(\Phi)}}{\sqrt{m}} \\
\text{conf}(\delta) &= \sqrt{\ln(1/\delta)}
\end{aligned}
$$

Substituting this in Theorem 4.2 gives the result

$$
\begin{aligned}
\widetilde{d}_m(\mathcal{X}) &= \widetilde{\gamma}\Big(\varepsilon_{\text{gen}}^2(m, \mathcal{H})/B_W^2; \mathcal{X}\Big) \\
&= \widetilde{\gamma}\Big(400H^2 \widetilde{d}_m(\Phi \circ \Psi)/mB_W^2; \mathcal{X}\Big) \\
V^\star(s_0) - V^{\pi_{f_t}}(s_0) &\leq 6H\sqrt{\widetilde{d}_m} \cdot \varepsilon_{\text{gen}}(m, \mathcal{H}) \cdot \text{conf}\big(\delta/(\widetilde{d}_m H)\big) \\
&= 120H^2 \frac{\sqrt{\widetilde{d}_m(\mathcal{X}) \cdot \widetilde{d}_m(\Phi)} \cdot \sqrt{\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}}{\sqrt{m}}
\end{aligned}
$$

□

We now complete the proof of Corollary 4.3. Note that both $\widetilde{d}_m(\Phi)$ and $\widetilde{d}_m(\mathcal{X})$ (related to critical information gain under Φ and $\mathcal{X}$ respectively) scale as $\widetilde{O}(d)$ if $\Phi \subset \mathbb{R}^d$.

Proof of Corollary 4.3. Since the proof follows similar to proof of Corollary 4.2, we will only provide a proof sketch here. First, from Lemma G.3, we have that

$$
\widetilde{\gamma}\Big(\frac{1}{8B_W^2 m}; \Phi\Big) \leq 4d\ln\left(25B_X^2 B_W^2 m\right)
$$

and therefore

$$
\widetilde{d}_m(\Phi) \leq 8d\ln^2(25B_X^2 B_W^2 m\sqrt{d})
$$

Similarly, as $\sup_{z\in\mathcal{X}}\|z\| \leq \sup_{x\in\Phi}\|x\|$, using Lemma G.3 (and since $400H^2\widetilde{d}_m(\Phi) \geq 1$), we get

$$
\widetilde{\gamma}\left(\frac{400H^2\widetilde{d}_m(\Phi)}{B_W^2 m}; \mathcal{X}_h\right) \leq 4d\ln\left(25B_X^2 B_W^2 m\right)
$$

and therefore

$$\widetilde{d}_m(\mathcal{X}) \leq 4dH \ln\left(4B_X^2 B_W^2 m\right)$$

To get ϵ-optimal policy, we have to set

$$120H^2 \frac{\sqrt{\widetilde{d}_m(\mathcal{X}) \cdot \widetilde{d}_m(\Phi)} \cdot \sqrt{\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}}{\sqrt{m}} \leq \epsilon$$

The rest of the proof follows similarly to proof of Corollary 4.2. □

D.3. Linear Mixture Model

We omit proof of Corollary 4.4 since it follows same as proof of Corollary 4.2.

Corollary D.3 (RKHS linear mixture model). *Suppose MDP $\mathcal{M}$ is a* linear Mixture Model. *Assume $\sup_{\theta\in\mathcal{H}_h, h\in[H]}\|\theta\|_2 \leq B_W$ and $\sup_{x\in\Phi_h, h\in[H]}\|x\|_2 \leq B_X$. Fix $\delta \in (0, 1/3)$, batch sample size m, and define:*

$$\widetilde{d}_m(\Phi) = \max_{h\in[H]} \widetilde{\gamma}\Big(\frac{1}{8B_W^2 m}; \Phi_h\Big) \cdot \nu_h$$
$$\widetilde{d}_m(\mathcal{X}) = \widetilde{\gamma}\Big(\frac{256H^2\widetilde{d}_m(\Phi)}{B_W^2 m}; \mathcal{X}\Big),$$

where $\nu_h = \ln\left(1 + 3B_X B_W \sqrt{m\widetilde{\gamma}\Big(\frac{1}{8B_W^2 m}; \Phi_h\Big)}\right)$.

Set parameters as: $R = (12H/\sqrt{m})\sqrt{\widetilde{d}_m(\mathcal{X}) \cdot \widetilde{d}_m(\Phi)} \cdot \sqrt{\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}$ and $T = \widetilde{d}_m(\mathcal{X})$. With probability greater than $1-\delta$, Algorithm 1 uses at most $mH\widetilde{d}_m(\mathcal{X})$ trajectories and returns a hypothesis f

$$V^\star(s_0) - V^{\pi_f}(s_0) \leq 96H^2 \frac{\sqrt{\widetilde{d}_m(\mathcal{X}) \cdot \widetilde{d}_m(\Phi)} \cdot v}{\sqrt{m}}.$$

where $v = \sqrt{\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}$.

Proof. First, using Corollary F.3 and Lemma G.1, we get that for any distribution μ over $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and for any $\delta \in (0,1)$, with probability of at least $1-\delta$ over choice of an i.i.d. sample $\mathcal{D} \sim \mu^m$ of size m, for all $g = (\theta_0, \ldots, \theta_{H-1}) \in \mathcal{H}$ (note that $\mathcal{L}_\mu(g)$ only depends on θ_h for distribution μ over observed transitions $o_h = (r_h, s_h, a_h, s_{h+1})$ at timestep h.)

$$\begin{aligned}
|\mathcal{L}_\mathcal{D}(g) - \mathcal{L}_\mu(g)| &\leq \frac{4}{\sqrt{m}} + 2H\sqrt{\frac{2\widetilde{\gamma}_m \ln\left(1 + 3B_X B_W\sqrt{\widetilde{\gamma}_m m}\right) + 2\ln(1/\delta)}{m}} + \sqrt{2}H\sqrt{\frac{\ln(1/\delta)}{m}} \\
&= \frac{4 + 2H\sqrt{2\widetilde{\gamma}_m \ln\left(1 + 3B_X B_W\sqrt{\widetilde{\gamma}_m m}\right) + 2\ln(1/\delta)} + \sqrt{2}H\sqrt{\ln(1/\delta)}}{\sqrt{m}} \\
&\leq \frac{16H\sqrt{\widetilde{\gamma}_m \ln\left(1 + 3B_X B_W\sqrt{\widetilde{\gamma}_m m}\right)} \cdot \sqrt{\ln(1/\delta)}}{\sqrt{m}}
\end{aligned}$$

where we have used that $\ln(1/\delta) > 1$ and $\widetilde{\gamma}_m = \max_{h\in[H]} \widetilde{\gamma}(1/(8B_W^2 m); \Phi_h)$ (as defined in Equation (5)). Define

$$\widetilde{d}_m(\Phi) := \widetilde{\gamma}_m \ln\left(1 + 3B_X B_W\sqrt{\widetilde{\gamma}_m m}\right)$$

This satisfies our Assumption 4.1 with

$$\varepsilon_{\text{gen}}(m, \mathcal{H}) = \frac{16H\sqrt{\widetilde{d}_m(\Phi)}}{\sqrt{m}}$$
$$\text{conf}(\delta) = \sqrt{\ln(1/\delta)}$$

Substituting this in Theorem 4.2 gives the result

$$\begin{aligned}
\widetilde{d}_m(\mathcal{X}) &= \widetilde{\gamma}\Big(\varepsilon^2_{\text{gen}}(m, \mathcal{H})/B_W^2; \mathcal{X}\Big) \\
&= \widetilde{\gamma}\Big(256H^2\widetilde{d}_m(\Phi)/mB_W^2; \mathcal{X}\Big) \\
V^\star(s_0) - V^{\pi_{f_t}}(s_0) &\le 6H\sqrt{\widetilde{d}_m} \cdot \varepsilon_{\text{gen}}(m, \mathcal{H}) \cdot \text{conf}\big(\delta/(\widetilde{d}_m H)\big) \\
&= 96H^2 \frac{\sqrt{\widetilde{d}_m(\mathcal{X}) \cdot \widetilde{d}_m(\Phi)} \cdot \sqrt{\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}}{\sqrt{m}}
\end{aligned}$$

□

D.4. Low Occupancy Complexity

Proof of Corollary 4.5. First, using Lemma G.1, we get that for any distribution μ over $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over choice of an i.i.d. sample $\mathcal{D} \sim \mu^m$ of size m, for all $g \in \mathcal{H}$

$$\begin{aligned}
|\mathcal{L}_{\mathcal{D}}(g) - \mathcal{L}_\mu(g)| &\le 2\sqrt{2}H\sqrt{\frac{\ln(|\mathcal{H}|/\delta)}{m}} \\
&\le 2\sqrt{2}H\sqrt{\frac{\ln(|e\mathcal{H}|/\delta)}{m}} \\
&= 2\sqrt{2}H\sqrt{\frac{1 + \ln(|\mathcal{H}|) + \ln(1/\delta)}{m}} \\
&\le 2\sqrt{2}H\sqrt{\frac{1 + \ln(|\mathcal{H}|)}{m}} \cdot \sqrt{\ln(1/\delta)}
\end{aligned}$$

This satisfies our Assumption 4.1 with

$$\varepsilon_{\text{gen}}(m, \mathcal{H}) = 2\sqrt{2}H\sqrt{\frac{1 + \ln(|\mathcal{H}|)}{m}}$$
$$\text{conf}(\delta) = \sqrt{\ln(1/\delta)}$$

Substituting this in Theorem 4.2 gives the result

$$\begin{aligned}
\widetilde{d}_m(\mathcal{X}) &= \widetilde{\gamma}\Big(\varepsilon^2_{\text{gen}}(m, \mathcal{H})/B_W^2; \mathcal{X}\Big) \\
&= \widetilde{\gamma}\Big(\frac{8H^2(1 + \ln(|\mathcal{H}|))}{mB_W^2}; \mathcal{X}\Big) \\
V^\star(s_0) - V^{\pi_{f_t}}(s_0) &\le 6H\sqrt{\widetilde{d}_m(\mathcal{X})} \cdot \varepsilon_{\text{gen}}(m, \mathcal{H}) \cdot \text{conf}\big(\delta/(\widetilde{d}_m(\mathcal{X})H)\big) \\
&= 12\sqrt{2}H^2 \frac{\sqrt{\widetilde{d}_m(\mathcal{X})} \cdot \sqrt{1 + \ln\left(|\mathcal{H}|\right)} \cdot \sqrt{\ln\left((\widetilde{d}_m(\mathcal{X})H)/\delta\right)}}{\sqrt{m}}
\end{aligned}$$

□

D.5. Finite Bellman Rank

Proof of Corollary 4.6. First, as observed in (Jiang et al., 2016)[Lemma 14], we get that for any distribution μ over $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and for any $\delta \in (0,1)$, with probability of at least $1-\delta$ over choice of an i.i.d. sample $\mathcal{D} \sim \mu^m$ of size m, for all $g \in \mathcal{H}$

$$\begin{aligned}|\mathcal{L}_{\mathcal{D}}(g) - \mathcal{L}_\mu(g)| &\le \sqrt{\frac{8|\mathcal{A}|H^2\ln(|\mathcal{H}|/\delta)}{m}} + \frac{2H|\mathcal{A}|\ln(|\mathcal{H}|/\delta)}{m}\\ &\le 4\sqrt{2}H\sqrt{|\mathcal{A}|}\sqrt{\frac{\ln(|e\mathcal{H}|/\delta)}{m}}\\ &= 4\sqrt{2}H\sqrt{|\mathcal{A}|}\sqrt{\frac{1+\ln(|\mathcal{H}|)+\ln(1/\delta)}{m}}\\ &\le 4\sqrt{2}H\sqrt{|\mathcal{A}|}\sqrt{\frac{1+\ln(|\mathcal{H}|)}{m}}\cdot\sqrt{\ln(1/\delta)}\end{aligned}$$

where the second inequality holds as long as $m > 2H|\mathcal{A}|\ln(|\mathcal{H}|/\delta)$. This satisfies our Assumption 4.1 with

$$\begin{aligned}\varepsilon_{\text{gen}}(m,\mathcal{H}) &= 4\sqrt{2}H\sqrt{|\mathcal{A}|}\sqrt{\frac{1+\ln(|\mathcal{H}|)}{m}}\\ \text{conf}(\delta) &= \sqrt{\ln(1/\delta)}\end{aligned}$$

Substituting this in Theorem 4.2 gives the result

$$\begin{aligned}\widetilde{d}_m(\mathcal{X}) &= \widetilde{\gamma}\Big(\varepsilon^2_{\text{gen}}(m,\mathcal{H})/B_W^2;\mathcal{X}\Big)\\ &= \widetilde{\gamma}\Big(\frac{32H^2|\mathcal{A}|(1+\ln(|\mathcal{H}|))}{mB_W^2};\mathcal{X}\Big)\\ &\le H\Big(3d\ln\Big(1+3mB_W^2B_X^2\Big)+1\Big)\\ &\le 4dH\ln\Big(4mB_W^2B_X^2\Big)\end{aligned}$$

where the second last step follows from Lemma G.3. Substituting ε_{gen} and conf in Theorem 4.2 also gives

$$\begin{aligned}&V^\star(s_0) - V^{\pi_{f_t}}(s_0)\\ &\le 6H\sqrt{\widetilde{d}_m(\mathcal{X})}\cdot\varepsilon_{\text{gen}}(m,\mathcal{H})\cdot\text{conf}\big(\delta/(\widetilde{d}_m(\mathcal{X})H)\big)\\ &= 24\sqrt{2}H^2\sqrt{|\mathcal{A}|}\frac{\sqrt{4dH\ln\Big(4mB_W^2B_X^2\Big)}\cdot\sqrt{1+\ln\left(|\mathcal{H}|\right)}\cdot\sqrt{\ln\left((4dH^2\ln\Big(4mB_W^2B_X^2\Big)/\delta\right)}}{\sqrt{m}}\end{aligned}$$

To get ϵ-optimal policy, we have to set

$$m \ge \frac{4608dH^5|\mathcal{A}|\ln\Big(4mB_W^2B_X^2\Big)\cdot(1+\ln(|\mathcal{H}|)\cdot\ln\left((4dH^2\ln\Big(4mB_W^2B_X^2\Big)/\delta\right)}{\epsilon^2}$$

Further simplifying the RHS, we can write it as

$$\frac{4608dH^5|\mathcal{A}|(1+\ln(|\mathcal{H}|))\cdot\ln^2\Big(16dH^2mB_W^2B_X^2/\delta\Big)}{\epsilon^2}$$

Using Lemma G.2 for $\alpha = 2$, $a = 4608dH^5|\mathcal{A}|(1+\ln(|\mathcal{H}|))/\epsilon^2$, $b = 16dH^2B_W^2B_X^2/\delta$ and $c = 9$, we get that

$$\begin{aligned}m &= \frac{41472dH^5|\mathcal{A}|(1+\ln(|\mathcal{H}|))}{\epsilon^2}\ln^2\Big(\frac{663552d^2H^7|\mathcal{A}|B_W^2B_X^2(1+\ln(|\mathcal{H}|))}{\delta\epsilon^2}\Big)\\ \ln\Big(4mB_W^2B_X^2\Big) &= 3\ln\Big(\frac{663552d^2H^7|\mathcal{A}|B_W^2B_X^2(1+\ln(|\mathcal{H}|))}{\delta\epsilon^2}\Big)\end{aligned}$$

Substituting this in the expression above for $\widetilde{d}_m(\mathcal{X})$ and setting this upper bound to T, we get

$$T = 12dH \ln\Big(\frac{663552 d^2 H^7 |\mathcal{A}| B_W^2 B_X^2 (1 + \ln(|\mathcal{H}|))}{\delta \epsilon^2}\Big)$$

Since, we use on policy estimation, i.e., $\pi_{est} = U(\mathcal{A})$ for all t, the trajectory complexity is mTH which completes the proof. □

E. An Elliptical Cover for Hilbert Spaces

The following theorem is a key technical contribution which allows us to obtain a number of non-parametric convergence rates.

Theorem E.1. *Let $\mathcal{X} \subset \mathcal{V}$, where $\mathcal{V}$ is a Hilbert space. Suppose $T \in \mathbb{N}^+, \epsilon \in \mathbb{R}^+$; define $\mathcal{W} \subseteq \{w \in \mathcal{V} : \|w\| \le B_W\}$ for some real number B_W; and suppose for all $x \in \mathcal{X}$ that $\|x\|_2 \le B_X$. Set $\lambda = \epsilon^2/(8B_W^2)$.There exists a set $\mathcal{C} \subset \mathcal{W}$ (a cover of $\mathcal{W}$) such that: (i) $\log|\mathcal{C}| \le T \log(1 + 3B_W B_X \sqrt{T}/\epsilon)$ and (ii) for all $w \in \mathcal{W}$, there exists a $w' \in \mathcal{C}$, such that:*

$$\sup_{x\in\mathcal{X}} |(w - w')\cdot x| \le \epsilon \sqrt{\left(\exp\left(\frac{\gamma_T(\epsilon^2/(8B_W^2))}{T}\right) - 1\right)}.$$

Proof. Let us suppose that $\mathcal{X}$ is closed, in order for certain maximizers (and arg-maximizers) over $\mathcal{X}$ to exist. If $\mathcal{X}$ is not closed, then let us replace $\mathcal{X}$ with the closure of $\mathcal{X}$, which is possible since $\mathcal{X}$ is a bounded set. Consider the process: Set $\Sigma_0 = \lambda I$ with $\lambda \in \mathbb{R}^+$.

1. For $t = 0, \ldots T - 1$,
   (a) $x_t = \text{argmax}_{x\in\mathcal{X}} \|x\|^2_{\Sigma_t^{-1}}$
   (b) $\Sigma_{t+1} = \Sigma_t + x_t x_t^\top$

Via Lemma C.6, we have that:

$$\sum_{t=0}^{T-1} \ln\left(1 + \|x_t\|^2_{\Sigma_t^{-1}}\right) \le \ln \frac{\det(\Sigma_T)}{\det(\Sigma_0)}.$$

This implies that there must exist a $t \in 0, \ldots, T-1$, such that:

$$\ln\left(1 + \|x_t\|^2_{\Sigma_t^{-1}}\right) \le \frac{\gamma_T(\lambda)}{T},$$

which means that:

$$\|x_t\|^2_{\Sigma_t^{-1}} \le \exp\left(\frac{\gamma_T(\lambda)}{T}\right) - 1.$$

Note that $x_t = \text{argmax}_{x\in\mathcal{X}} \|x\|_{\Sigma_t^{-1}}$. Thus, we have that:

$$\max_{x\in\mathcal{X}} \|x\|^2_{\Sigma_t^{-1}} \le \exp\left(\frac{\gamma_T(\lambda)}{T}\right) - 1.$$

Note that the above derivation holds for any $\lambda \in \mathbb{R}^+$.

Define $M_T = \sum_{i=0}^T x_t x_t^\top$. Note that the range of M_T, Range(M_T) is a $T+1$-dimensional object. For an ϵ'-net, $\mathcal{C}$, in ℓ_2 distance over B_W-norm ball on Range(M_T), i.e., $\{v \in \mathcal{W} : v \in \text{Range}(M_T)\}$. With a standard covering number bound, we have that $\ln(|\mathcal{C}|) \le 2T \ln(1 + 2B_W/\epsilon')$ (e.g. see Lemma F.1).

Fix some $w \in \mathcal{W}$. Denote the projection of w on the the range of M_T by $\overline{w}$. Let $w' \in \mathcal{C}$ being the closest point to $\overline{w}$ in ℓ_2 distance. Note that $\|\overline{w} - w'\|_2 \leq \epsilon'$. For any $x \in \mathcal{X}$, we have:

$$\begin{aligned}
&\left((w-w')^\top x\right)^2 \leq \|w-w'\|^2_{\Sigma_T} \|x\|^2_{\Sigma_T^{-1}} \\
&\leq \|w-w'\|^2_{\Sigma_T} (\exp(\gamma_T(\lambda)/T) - 1) \\
&= \left(\lambda \|w-w'\|^2 + (w-w')^\top \left(\sum_{i=0}^{T} x_i x_i^\top\right)(w-w')\right)(\exp(\gamma_T(\lambda)/T) - 1) \\
&= \left(\lambda \|w-w'\|^2 + (\overline{w}-w')^\top \left(\sum_{i=0}^{T} x_i x_i^\top\right)(\overline{w}-w')\right)(\exp(\gamma_T(\lambda)/T) - 1) \\
&\leq \left(4\lambda B_W^2 + T\epsilon'^2 B_X^2\right)(\exp(\gamma_T(\lambda)/T) - 1),
\end{aligned}$$

where the equality in the third step uses that $(w-w')^\top x_i = (\overline{w}-w')^\top x_i$ for all $i \in 0, \ldots, T$. The proof is completed choosing $\lambda = \epsilon^2/(8B_W^2)$ and $(\epsilon')^2 = \epsilon^2/(2TB_X^2)$. □

F. Concentration Arguments for Special Cases

An application to RKHS Linear MDPs. Consider the RKHS linear MDP, where $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{H}$ with $\mathcal{H}$ being some Hilbert space. Define $\Phi = \{\phi(s,a) : s \in \mathcal{S}, a \in \mathcal{A}\}$.

Corollary F.1. *Suppose $T \in \mathbb{N}^+$ and $\epsilon \in \mathbb{R}^+$; define $\mathcal{W} \subseteq \{w \in \mathcal{H} : \|w\| \leq B_W\}$ for some real number B_W; and suppose for all $\phi(s,a) \in \Phi$ that $\|\phi(s,a)\|_2 \leq B_\phi$. There exists a set $\mathcal{C} \subset \mathcal{W}$ such that: (i) $\log|\mathcal{C}| \leq T \log(1 + 3B_\phi B_W \sqrt{T}/\epsilon)$ and (ii) for all $w \in \mathcal{W}$, there exists a $w' \in \mathcal{C}$ such that for all distributions d over $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we have:*

$$\begin{aligned}
&\Big| \mathbb{E}_{s,a,s'\sim d}\big[w \cdot \phi(s,a) - r(s,a) - \max_{a'} w \cdot \phi(s',a')\big] \\
&\quad - \mathbb{E}_{s,a,s'\sim d}\big[w' \cdot \phi(s,a) - r(s,a) - \max_{a'} w' \cdot \phi(s,a')\big] \Big| \\
&\quad \leq 2\epsilon \sqrt{\left(\exp\left(\frac{\gamma_T(\epsilon^2/(8B_W^2))}{T}\right) - 1\right)}
\end{aligned}$$

Proof. For any distribution d, we seek to bound:

$$\begin{aligned}
&\left| \mathbb{E}_{s,a,s'\sim d}\left[w \cdot \phi(s,a) - w' \cdot \phi(s,a) - \big(\max_{a'} w \cdot \phi(s',a') - \max_{a'} w' \cdot \phi(s,a')\big)\right] \right| \\
&\leq \sup_{s,a} \left| w \cdot \phi(s,a) - w' \cdot \phi(s,a) \right| + \left| \mathbb{E}_{s,a,s'\sim d}\left[\big(\max_{a'} w \cdot \phi(s',a') - \max_{a'} w' \cdot \phi(s,a')\big)\right] \right| \\
&\leq \sup_{s,a} \left| w \cdot \phi(s,a) - w' \cdot \phi(s,a) \right| + \sup_{s} \left| \sup_{a} w \cdot \phi(s,a) - \sup_{a} w' \cdot \phi(s,a) \right| \\
&\leq 2 \sup_{s,a} \left| w \cdot \phi(s,a) - w' \cdot \phi(s,a) \right|
\end{aligned}$$

where the last step follows using that $|\sup_x f(x) - \sup_x g(x)| \leq \sup_x |f(x) - g(x)|$ (which can be verified by considering both case of the sign inside the absolute value). The proof is completed by choose w' to be closest point $\mathcal{C}$ to w and applying Theorem E.1. □

Corollary F.2. *Define $\mathcal{W} =: \{w \in \mathcal{H} : \|w\| \leq B_W, w^\top \phi(s,a) \in [0,H]\ \forall s,a \in \mathcal{S} \times \mathcal{A}\}$ for some real number B_W; and suppose for all $\phi(s,a) \in \Phi$ that $\|\phi(s,a)\|_2 \leq B_\phi$. Let*

$$\ell(r,s,a,s',w) = w \cdot \phi(s,a) - r - \max_{a'} w \cdot \phi(s',a')$$

with $r \in [0,1]$. Then, for any distribution μ over $\mathbb{R} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and for any $\delta \in (0,1)$, with probability of at least $1-\delta$ over choice of an i.i.d. sample $\mathcal{D} \sim \mu^m$ of size m, for all $w \in \mathcal{H}$

$$|\mathcal{L}_\mathcal{D}(w) - \mathcal{L}_\mu(w)| \leq \frac{8}{\sqrt{m}} + 2H\sqrt{\frac{2\widetilde{\gamma}_m \ln\left(1 + 3B_\phi B_W \sqrt{\widetilde{\gamma}_m m}\right) + 2\ln(1/\delta)}{m}}$$

where $\widetilde{\gamma}_m = \widetilde{\gamma}(1/(8B_W^2 m); \Phi)$ *(as defined in Equation (5)).*

Proof. First note that for any $w \in \mathcal{W}$, we must have:

$$\ell(r, s, a, s', w) \in [-H-1, H],$$

since we eliminate all w such that $w^\top \phi(s,a) \notin [0, H]$ for some s, a.

Consider the cover $\mathcal{C}$ from Corollary F.1. From Lemma G.1 and a union bound over all $w' \in \mathcal{C}$, for all $w' \in \mathcal{C}$, we have that with probability at least $1-\delta$:

$$|\mathcal{L}_{\mathcal{D}}(w') - \mathcal{L}_\mu(w')| \le 2H\sqrt{\frac{2\ln(|\mathcal{C}|/\delta)}{m}}.$$

Now consider any $w \in \mathcal{W}$, via Corollary F.1, we know that there exists a $w' \in \mathcal{C}$ such that:

$$|\mathcal{L}_\mu(w) - \mathcal{L}_\mu(w')| \le 2\epsilon\sqrt{\left(\exp\left(\frac{\gamma_T(\lambda)}{T}\right) - 1\right)}.$$

Thus, together with the fact that Corollary F.1 holds for both μ and the uniform distribution over $\mathcal{D}$, we get:

$$\begin{aligned}
&|\mathcal{L}_\mu(w) - \mathcal{L}_{\mathcal{D}}(w)| \le |\mathcal{L}_\mu(w) - \mathcal{L}_\mu(w')| + |\mathcal{L}_\mu(w') - \mathcal{L}_{\mathcal{D}}(w')| + |\mathcal{L}_{\mathcal{D}}(w') - \mathcal{L}_{\mathcal{D}}(w)| \\
&\le 4\epsilon\sqrt{\left(\exp\left(\frac{\gamma_T(\lambda)}{T}\right) - 1\right)} + 2H\sqrt{\frac{2\ln(|\mathcal{C}|/\delta)}{m}} \\
&\le 4\epsilon\sqrt{\left(\exp\left(\frac{\gamma_T(\epsilon^2/(8B_W^2))}{T}\right) - 1\right)} + 2H\sqrt{\frac{2T\ln\left(1 + 3B_\phi B_W\sqrt{T}/\epsilon\right) + 2\ln(1/\delta)}{m}}
\end{aligned}$$

Let us set $\epsilon = 1/\sqrt{m}$ and rearrange terms, we get:

$$\begin{aligned}
&|\mathcal{L}_\mu(w) - \mathcal{L}_{\mathcal{D}}(w)| \\
&\le \frac{4}{\sqrt{m}}\sqrt{\left(\exp\left(\frac{\gamma_T(1/(8B_W^2 m))}{T}\right) - 1\right)} + 2H\sqrt{\frac{2T\ln\left(1 + 3B_\phi B_W\sqrt{Tm}\right) + 2\ln(1/\delta)}{m}}.
\end{aligned}$$

Denote $\widetilde{\gamma}_m = T$ where T is the smallest integer that satisfies $T \ge \gamma_T(1/(8B_W^2 m))$. Thus, we have:

$$\begin{aligned}
&|\mathcal{L}_\mu(w) - \mathcal{L}_{\mathcal{D}}(w)| \\
&\le \frac{8}{\sqrt{m}} + 2H\sqrt{\frac{2\widetilde{\gamma}_m\ln\left(1 + 3B_\phi B_W\sqrt{\widetilde{\gamma}_m m}\right) + 2\ln(1/\delta)}{m}},
\end{aligned}$$

where in the inequality we use $\exp\left(\frac{\gamma_T(1/(8B_W^2 m))}{T}\right) - 1 \le e - 1 \le 2$.

□

An application to RKHS linear functions Consider features $\zeta : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathcal{V}$ with $\mathcal{V}$ being some Hilbert space. Define $Z = \{\zeta(s,a,s') : (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}\}$.

Corollary F.3. *Define* $\mathcal{W} =: \{w \in \mathcal{V} : \|w\| \le B_W, w^\top \zeta(s,a,s') \in [0,H]\ \forall s,a,s' \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}\}$ *for some real number* B_W*; and suppose for all* $\zeta(s,a,s') \in Z$ *that* $\|\zeta(s,a,s')\|_2 \le B_\zeta$*. Let*

$$\ell(r, s, a, s', w) = w \cdot \zeta(s, a, s')$$

Then, for any distribution μ over $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and for any $\delta \in (0,1)$, with probability of at least $1-\delta$ over choice of an i.i.d. sample $\mathcal{D} \sim \mu^m$ of size m, for all $w \in \mathcal{H}$

$$|\mathcal{L}_{\mathcal{D}}(w) - \mathcal{L}_{\mu}(w)| \leq \frac{4}{\sqrt{m}} + 2H\sqrt{\frac{2\widetilde{\gamma}_m \ln\left(1 + 3B_\zeta B_W \sqrt{\widetilde{\gamma}_m m}\right) + 2\ln(1/\delta)}{m}}$$

where $\widetilde{\gamma}_m = \widetilde{\gamma}(1/(8B_W^2 m); Z)$ (as defined in Equation (5)).

Proof. The proof follows exactly as proof of Corollary F.2. □

Lemma F.1 (Covering number). *For any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R \in \mathbb{R}^+$, i.e., $\mathcal{B} = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$, is upper bounded by $(1 + 2R/\epsilon)^d$.*

G. Auxiliary Lemmas

Lemma G.1 (Azuma-Hoeffding). *Let $X_1, \ldots, X_m$ be independent random variables with mean μ such that $|X_i| \leq B$ for some $B > 0$ almost surely for all $i \in [m]$. Then, with probability $1-\delta$,*

$$\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \leq \sqrt{2}B\sqrt{\frac{\ln(1/\delta)}{m}}$$

Lemma G.2. *(Log Dominance Rule) Suppose $\alpha, a, b \geq 0$ and $c \geq (1+\alpha)^\alpha$. Then, $m = ca\ln^\alpha(abc)$ is a solution to*

$$m \geq a\ln^\alpha(bm)$$

Proof. First note that

$$\begin{aligned}
&a\ln^\alpha(bm)\\
&= a\ln^\alpha(abc\ln^\alpha(abc))\\
&= a\left(\ln(abc) + \alpha\ln\ln(abc)\right)^\alpha\\
&\leq a\left(\ln(abc) + \alpha\ln(abc)\right)^\alpha\\
&= a(1+\alpha)^\alpha \ln^\alpha(abc)\\
&\leq ca\ln^\alpha(abc)
\end{aligned}$$

□

Lemma G.3. *Let $\mathcal{X} \subset \mathbb{R}^d$ and $\sup_{x\in\mathcal{X}}\|x\|_2 \leq B_X$. Then, the maximum information gain*

$$\gamma_n(\lambda; \mathcal{X}) \leq d\ln\left(1 + \frac{nB_X^2}{d\lambda}\right)$$

Furthermore, the critical information gain

$$\widetilde{\gamma}(\lambda; \mathcal{X}) \leq \left\lceil 3d\ln\left(1 + \frac{3B_X^2}{\lambda}\right)\right\rceil$$

Proof.

$$\gamma_n(\lambda; \mathcal{D}) := \max_{x_0 \ldots x_{n-1} \in \mathcal{D}} \ln\det\left(\mathrm{I} + \frac{1}{\lambda}\sum_{t=0}^{n-1} x_t x_t^\top\right).$$

We have

$$\begin{aligned}
\mathrm{trace}\left(\mathrm{I} + \frac{1}{\lambda}\sum_{t=0}^{n-1} x_t x_t^\top\right) &= d + \frac{1}{\lambda}\sum_{t=0}^{n-1}\|x_t\|_2^2\\
&\leq d + nB_X^2/\lambda
\end{aligned}$$

Therefore, using the Determinant-Trace inequality, we get the first result

$$\begin{aligned} \ln\det\left(\mathrm{I} + \frac{1}{\lambda}\sum_{t=0}^{n-1} x_t x_t^\top\right) &\le d\ln\frac{\operatorname{trace}\left(\mathrm{I} + \frac{1}{\lambda}\sum_{t=0}^{n-1} x_t x_t^\top\right)}{d} \\ &\le d\ln\left(1 + \frac{nB_X^2}{d\lambda}\right) \end{aligned}$$

To get the second result, first note that for $n = cd\ln(1 + cB_X^2/\lambda)$ and $c = 3$,

$$\begin{aligned} d\ln\left(1 + \frac{nB_X^2}{d\lambda}\right) &= d\ln\left(1 + \frac{cB_X^2}{\lambda}\ln(1 + cB_X^2/\lambda)\right) \\ &\le d\ln\left(1 + \frac{cB_X^2}{\lambda}\max\{\ln(1 + cB_X^2/\lambda), 1\}\right) \\ &\le d\ln\left((1 + \frac{cB_X^2}{\lambda})\max\{\ln(1 + cB_X^2/\lambda), 1\}\right) \\ &\le d\left(\ln\big(1 + \frac{cB_X^2}{\lambda}\big) + \ln\big(\max\{\ln(1 + cB_X^2/\lambda), 1\}\big)\right) \\ &\le d\left(\ln\big(1 + \frac{cB_X^2}{\lambda}\big) + \ln(1 + cB_X^2/\lambda)\right) \\ &= 2d\ln\big(1 + \frac{cB_X^2}{\lambda}\big) \\ &\le n \end{aligned}$$

where the third last step follows from $\ln(1 + cB_X^2/\lambda) \ge 0$ and $\ln(1 + cB_X^2/\lambda) \ge \ln(\ln(1 + cB_X^2/\lambda))$ and last step follows from $c = 3 > 2$. $\square$