
Bilinear Classes: A Structural Framework for Provable Generalization in RL

Simon S. Du¹ Sham M. Kakade¹ Jason D. Lee² Shachar Lovett³ Gaurav Mahajan³ Wen Sun⁴
Ruosong Wang⁵

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).

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

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 work¹ 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

¹Full version appears as [arXiv:2103.10897].

Bilinear Classes: A Structural Framework for Provable Generalization in RL

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 contain 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^* “irrelevant” State Aggregation (Li, 2009)	✓	✗	✗	✓
Linear Q^*/V^* (this work)	✗	✗	✗	✓
RKHS Linear MDP (this work)	✗	✗	✗	✓
RKHS Linear Mixture MDP (this work)	✗	✗	✗	✓
Low Occupancy Complexity (this work)	✗	✗	✗	✓
Q^* State-action Aggregation (Dong et al., 2020)	✗	✗	✗	✗
Deterministic linear Q^* (Wen and Van Roy, 2013)	✗	✗	✗	✗
Linear Q^* (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 dis-

cussion 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 work shows that if we assume that both Q^* and V^* 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 be 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 $R(s, a)$, $P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ (where $\Delta(\mathcal{S})$ 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 state². 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

²Our results generalizes to any fixed initial state distribution

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^* and V_h^* 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 $\Delta(S)$ 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 $\mathcal{H} = \mathcal{H}_0 \times \dots \times \mathcal{H}_{H-1}$, 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

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., $\pi_{h,f}(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q_{h,f}(s, a)$, 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 $W_h : \mathcal{H} \rightarrow \mathbb{R}^d$ and $X_h : \mathcal{H} \rightarrow \mathbb{R}^d$, such that for all $f, g \in \mathcal{H}$:

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

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 $W_h : \mathcal{H} \rightarrow \mathbb{R}^d$ and $X_h : \mathcal{H} \rightarrow \mathbb{R}^d$, such that for all $f, g \in \mathcal{H}$

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

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 $f^* \in \mathcal{H}$ such that $Q_h^*(s, a) = Q_{h,f^*}(s, a)$, where Q_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 $\ell_f : (\mathbb{R} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}) \times \mathcal{H} \rightarrow \mathbb{R}$ (defined for each $f \in \mathcal{H}$), and a set of estimation policies $\Pi_{\text{est}} = \{\pi_{\text{est}}(f) : f \in \mathcal{H}\}$. 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 $W_h : \mathcal{H} \rightarrow \mathcal{V}$ and $X_h : \mathcal{H} \rightarrow \mathcal{V}$ 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:

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

2. The policy $\pi_{\text{est}}(f)$ 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’’)

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

Typically, $\pi_{\text{est}}(f)$ 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 $\mathcal{X}_h := \{X_h(f) : f \in \mathcal{H}\}$ 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 $|\langle W_h(f) - W_h(f^*), X_h(f) \rangle|$ (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 $|\langle W_h(g) - W_h(f^*), X_h(f) \rangle|$ 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, $(\mathcal{H}, \ell_f, \Pi_{est}, \mathcal{M})$ 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:

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

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, $(\mathcal{H}, \ell_f, \Pi_{est}, \mathcal{M})$ 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^*/V^* 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) $\theta^* \in \mathcal{V}$ for some Hilbert space \mathcal{V} such that for all $h \in [H]$ and $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$*

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

We denote hypothesis in our hypothesis class \mathcal{H} as tuples $(\theta_0, \dots, \theta_{H-1})$, 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 $g = \{\theta_0, \dots, \theta_{H-1}\} \in \mathcal{H}$, $V_{h,g}$ 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 $g = \{\theta_0, \dots, \theta_{H-1}\} \in \mathcal{H}$ 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 $\pi_{est}(f) = \pi_f$, $(\mathcal{H}, \ell_f, \Pi_{est}, \mathcal{M})$ is (implicitly) a Bilinear Class.*

3.3.2. LINEAR Q^*/V^* (NEW MODEL)

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

Definition 3.5 (Linear Q^*/V^*). *We say that a MDP \mathcal{M} is a linear Q^*/V^* 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) $(w^*, \theta^*) \in \mathcal{V}_1 \times \mathcal{V}_2$ for some Hilbert spaces $\mathcal{V}_1, \mathcal{V}_2$ such that for all $h \in [H]$ and for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

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

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

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

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 $Q^*/V^* \implies$ Bilinear Class). *Consider a MDP \mathcal{M} which is a linear Q^*/V^* model. Then, for the hypothesis class \mathcal{H} , the discrepancy function ℓ_f defined above and on-policy estimation policies $\pi_{est}(f) = \pi_f$, $(\mathcal{H}, \ell_f, \Pi_{est}, \mathcal{M})$ 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 $(\theta_0, \dots, \theta_{H-1})$, 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 $\mathcal{T}_h : \mathcal{V} \rightarrow \mathcal{V}$ such that for all $(\theta_0, \dots, \theta_{H-1}) \in \mathcal{H}$ 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 $g = (\theta_0, \dots, \theta_{H-1})$ 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, $(\mathcal{H}, \ell_f, \Pi_{est}, \mathcal{M})$ 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 $\phi_h : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{V}$ (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 $\pi_{est}(f) = \pi_f$, $(\mathcal{H}, \ell_f, \Pi_{est}, \mathcal{M})$ 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 $\mu_h : \mathcal{S} \rightarrow \mathcal{V}$. 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 $t = 0, 1, 2, \dots, T - 1$ **do**
- 3: Set f_t as the solution of the following program:

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

- 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 $\operatorname{conf}(\delta)$ 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 ,

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

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, $\varepsilon_{gen}(m, \mathcal{H}) \rightarrow 0$ as $m \rightarrow \infty$. 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 $(\mathcal{H}, \ell, \Pi_{est}, \mathcal{M})$ is a Bilinear Class with $\mathcal{X}_h \subset \mathbb{R}^d$ for all timesteps h and Assumption 4.1 holds. Assume $\sup_{f \in \mathcal{H}, h \in [H]} \|W_h(f)\|_2 \leq B_W$ and $\sup_{f \in \mathcal{H}, h \in [H]} \|X_h(f)\|_2 \leq B_X$. Fix $\delta \in (0, 1/3)$ and batch sample size m and define:

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

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

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

As discussed in the Remark 4.1, $\varepsilon_{gen}(m, \mathcal{H})$ and $\operatorname{conf}(\delta)$ measure the uniform convergence of discrepancy functions ℓ_f for the hypothesis class \mathcal{H} . Therefore, if $\varepsilon_{gen}(m, \mathcal{H})$ 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 $(\mathcal{H}, \ell, \Pi_{est}, \mathcal{M})$ 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 $\sup_{f \in \mathcal{H}, h \in [H]} \|W_h(f)\|_2 \leq B_W$ and $\sup_{f \in \mathcal{H}, h \in [H]} \|X_h(f)\|_2 \leq B_X$ for some $B_X, B_W \geq 1$. Assume the discrepancy function ℓ_f is bounded i.e. $\sup_{f \in \mathcal{H}} |\ell_f(\cdot)| \leq H + 1$. 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 $T = c_2 d H \ln(B_X B_W m)$ 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 $V^*(s_0) - V^{\pi_f}(s_0) \leq \epsilon$ 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

$$\nu = \ln \left(\frac{dHB_X B_W \ln(|\mathcal{H}|) \ln(1/\delta)}{\epsilon} \right).$$

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:

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

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 $\tilde{\gamma}(\lambda; \mathcal{X})$, as the smallest integer $k > 0$ s.t. $k \geq \gamma_k(\lambda; \mathcal{X})$, i.e.

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

(where k is an integer). Note that such a $\tilde{\gamma}(\lambda; \mathcal{X})$ 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 be 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 $\|x\| \leq B_X$ for any $x \in \mathcal{X}$, we have: $\gamma_n(\lambda; \mathcal{X}) \leq d \ln(1 + nB_X^2/d\lambda)$ and $\tilde{\gamma}(\lambda; \mathcal{X}) \leq 3d \ln(1 + 3B_X^2/\lambda)$ (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 $\mathcal{X}_h := \{X_h(f) : f \in \mathcal{H}\}$ 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 $\sup_{f \in \mathcal{H}, h \in [H]} \|W_h(f)\|_2 \leq B_W$. Fix $\delta \in (0, 1/3)$, batch sample size m , and define:

$$\tilde{d}_m = \tilde{\gamma} \left(\varepsilon_{\text{gen}}^2(m, \mathcal{H}) / B_W^2; \mathcal{X} \right).$$

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

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

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^*/V^* , 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^*/V^*

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

$$\Phi = \left\{ \phi(s, a) : (s, a) \in \mathcal{S} \times \mathcal{A} \right\}, \quad \Psi = \left\{ \psi(s') : s' \in \mathcal{S} \right\}.$$

and define the concatenation set³

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

Corollary 4.2 (Finite Dimensional Linear Q^*/V^*). Suppose MDP \mathcal{M} is a linear Q^*/V^* model with $\Phi \circ \Psi \subset \mathbb{R}^d$. Assume $\sup_{(w, \theta) \in \mathcal{H}_h, h \in [H]} \|[w, \theta]\|_2 \leq B_W$ and $\sup_{x \in \Phi \circ \Psi} \|x\|_2 \leq B_X$ for some $B_X, B_W \geq 1$. 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 $V^*(s_0) - V^{\pi_f}(s_0) \leq \epsilon$ using at most

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

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

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

³For 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 $\sup_{\theta \in \mathcal{H}_h, h \in [H]} \|\theta\|_2 \leq B_W$ and $\sup_{x \in \Phi} \|x\|_2 \leq B_X$ for some $B_X, B_W \geq 1$. 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 $V^*(s_0) - V^{\pi_f}(s_0) \leq \epsilon$ using at most*

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

trajectories for some absolute constant c_1 .

In comparison, (Jin et al., 2020) has sample complexity $\tilde{O}(d^3 H^3 / \epsilon^2 \log(1/\delta))$ and (Zanette et al., 2020) has $\tilde{O}(d^2 H^3 / \epsilon^2 \log(1/\delta))$. 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:

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

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 $\sup_{\theta \in \mathcal{H}_h, h \in [H]} \|\theta\|_2 \leq B_W$ and $\sup_{x \in \Phi_h, h \in [H]} \|x\|_2 \leq B_X$ for some $B_X, B_W \geq 1$. 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 $V^*(s_0) - V^{\pi_f}(s_0) \leq \epsilon$ using at most*

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

trajectories for some absolute constant c_1 .

In comparison, (Modi et al., 2020a) has sample complexity $\tilde{O}(d^2 H^2 / \epsilon^2 \log(1/\delta))$. 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 $\sup_{f \in \mathcal{H}_h, h \in [H]} \|W_h(f)\|_2 \leq B_W$. Fix $\delta \in (0, 1/3)$, batch sample size m , and define:*

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

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

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

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

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 $\sup_{f \in \mathcal{H}_h, h \in [H]} \|W_h(f)\|_2 \leq B_W$ and $\sup_{f \in \mathcal{H}, h \in [H]} \|X_h(f)\| \leq B_X$ for some $B_W, B_X \geq 1$. 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 $V^*(s_0) - V^{\pi_f}(s_0) \leq \epsilon$ using at most*

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

trajectories.

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

Niranjn 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.