Provably Efficient Reinforcement Learning in Partially Observable Dynamical Systems

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study Reinforcement Learning for partially observable dynamical systems using function approximation. We propose a new Partially Observable Bilinear Actor-Critic framework, that is general enough to include models such as tabular Partially Observable Markov Decision Processes (POMDPs), observable Linear-Quadratic-Gaussian (LQG), Predictive State Representations (PSRs), as well as a newly introduced model Hilbert Space Embeddings of POMDPs. Under this framework, we propose an actor-critic style algorithm that is capable of performing agnostic policy learning. Given a policy class that consists of memory based policies (i.e., policy that looks at a fixed-length window of recent observations), and a value function class that consists of functions taking both memory and future observations as inputs, our algorithm learns to compete against the best memory-based policy among the policy class. For certain examples such as undercomplete observable tabular POMDPs and LQGs, by implicitly leveraging their special properties, our algorithm is even capable of competing against the globally optimal policy without paying an exponential dependence on the horizon in its sample complexity.

1 Introduction

Large state space and partial observability are two key challenges of Reinforcement Learning (RL). While recent advances in RL for fully observable systems have focused on the challenge of scaling RL to large state space in both theory and in practice using rich function approximation, the understanding of large scale RL under partial observability is still limited. In POMDPs, for example, a core issue is that the optimal policy is not necessarily Markovian since the observations are not Markovian.

Empirically, the common heuristic to tackle large scale RL with partial observability is to simply maintain a time window of the history of observations, which is treated as a state to feed into the policy and the value function. Such a window of history can be often maintained explicitly via truncating away older history (e.g., DQN uses a window with length 4 for playing video games [44]; Open AI Five uses a window with length 16 [6] for LSTMs). Since even in planning with known dynamics, finding the globally optimal policy conditional on the entire history is generally NP-hard and extremely computationally expensive (i.e., the curse of the horizon) [39, 46, 20], searching for a short memory-based policy can be understood as a reasonable middelground that balances computation and optimality. The impressive empirical results of these prior works also demonstrate that empirically, there often exists a high-quality policy (not necessarily the globally optimal) that is only a function of a short window of recent observations. However, these prior works that search for the best memory-based policy unfortunately cannot ensure sample efficient PAC guarantees due to the difficulty of strategic exploration in POMDPs.

The key question that we aim to answer is: *Can we design RL algorithms that agnostically learn the best fixed-length memory policy using function approximation, in a provably efficient manner?*

We provide affirmative answers to the above question. More formally, we study RL for partially observable dynamical systems that include not only the classic Partially Observable MDPs

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Model	Observable tabular POMDPs	Observable LQG	Low-rank M-step decodable POMDPs	Observable HSE-POMDPs	PSRs
PO-Bilinear Rank	$O^M A^M S(\dagger)$	$O(M d_a^2 d_s^2)(\dagger)$	Rank $(\dagger)$	Feature dimension on (z, s)	$O^M A^M \times$ # of core tests
PAC Learning	Known	Known	Known	New	New

Table 1: Summary of settings that are from PO-Bilinear AC class. The 2nd row gives the parameters that bound the PO-Bilinear rank. Here M denotes the length of memory used to define memory-based policies $\pi(\cdot|\bar{z}_h)$ where $\bar{z}_h = (o_{h-M:h}, a_{h-M:h-1})$ denotes the M-step memory. In the 3rd row, "known" means that sample-efficient algorithms already exist. " New" means our result gives the first sample-efficient algorithm. The symbol † means we can compete with the globally optimal policy without paying an exponential dependence on horizon H. For LQG, d_a and d_s are the dimension of action and state spaces. For PSRs, O and A denote the size of observation and action spaces.

(POMDPs), but also a more general model called Predictive State Representations (PSRs) [38]. We design a model-free actor-critic framework, named *PO-Bilinear Actor-Critic Class*, where we have a policy class (i.e., actors) that consists of policies that take a fixed-length window of observations as input (memory-based policy), and a newly introduced value bridge function class (i.e., critics) that consists of functions that take the fixed-length window of history and *future observations* as inputs. A value bridge function class is an analog of the value function class tailored to partially observable systems that only relies on observable quantities (i.e., past and future observations and actions). In our algorithm, we *agnostically* search for the best memory-based policy from the given policy class.

Our framework is based on the idea of a newly introduced value bridge function class equipped with future observations. While the idea of using future observations has been used in the literature on POMDPs, our work is the first to use this idea to learn a high-quality policy in a model-free manner. Existing works discuss how to use future observations only in a model-based manner [9, 23]. By leveraging these model-based viewpoints, while recent works discuss strategic exploration to learn near-optimal policies, their results are limited to tabular models (non-scalable) [30, 21, 4, 59, 40] or tailored to specific non-tabular models and unclear how to incorporate general function approximation [49, 36, 10]. We break these barriers by devising the new actor-critic-based model-free view on POMDPs. We demonstrate the *scalability* and *generality* of PO-bilinear actor-critic framework by showing PAC-guarantee on many models as follows and Table 1.

Tabular POMDPs and Linear Quadratic Gaussian (LQG) In tabular POMDPs and LQG, our framework is also flexible enough to set the memory length according to the property of the problems in order to search for the globally optimal policy. For example, in (undercomplete) tabular POMDPs, using the latest planning result from [20] we only need to look at memory with length scaling $\ln(SH/\epsilon)$ with H being the horizon of the problem, S being the number of latent state, and ϵ being the optimality threshold. This allows us to compete against the globally optimal policy without paying an exponential dependence on H. For LQG, by implicitly leveraging the linear structure of the problem, we are able to look at the full length of history and learn a globally optimal policy in polynomial sample complexity. *To the best of our knowledge, this is the first unified algorithm that can solve both tabular POMDPs and LQG simultaneously without paying an exponential dependence on horizon H.*

Hilbert Space Embedding POMDPs (HSE-POMDPs) Our framework ensures the agnostic PAC guarantee on HSE-POMDPs [9, 8], which generalize tabular POMDPs and LQG via conditional mean embedding. This is the *first* PAC guarantee in HSE-POMDPs.

M-step decodable POMDPs [17] Our framework can capture M-step decodable POMDPs which is a generalization of block MDPs. We obtain a state-of-the-art PAC result as [17].

Predictive State Representations (PSRs) We give the *first* agnostic PAC-guarantee on PSRs. Our work significantly generalizes a prior PAC learning result for reactive PSRs (i.e., reactive PSRs require a strong condition that the optimal policy only depends on the latest observation) [27].

1.1 Related Works

We discuss related works. Additional works related to system identification, PSRs and value bridge functions are provided in Section A.

Generalization and function approximation of RL in MDPs. In Markovian environments, there is a growing literature that gives PAC bounds with function approximation under certain models. Some of the representative models are linear MDPs [30, 60], block MDPs [15, 62], and low-rank MDPs [2, 58]. Several general frameworks [27, 54, 29, 19, 14] characterize sufficient conditions for provably efficient RL. Each above model is captured in these frameworks as a special case. While our work builds on the bilinear/Bellman rank framework [14, 27], even in the tabular case, when

we naïvely reduce POMDPs to MDPs, the bilinear rank incurs $O(A^H)$, which we want to avoid, even if the policy class is restricted to M-memory policies. They also show the bilinear rank is S under reactive POMDPs where the optimal policy depends on the latest observation. However, this assumption is very strong.

Online RL for POMDPs. Prior works [32, 18] showed A^H-type sample complexity bounds for general POMDPs. Exponential dependence can be circumvented with more structures. First, in the tabular setting, under observability assumptions, [4, 21, 28] obtain favorable sample complexities by leveraging the spectral learning technique [25] (see section 1.1 in [28] for an excellent summary). Second, in LQG, which is a partial observable version of LQRs, [36, 49] propose sub-linear regret algorithms. These works use random policies for exploration, which is sufficient for LQG. Since random exploration strategy is not enough for tabular POMDPs, it is unclear if the existing techniques from LQG can be applied to solve general POMDPs. Third, the recent work [17] provides a new model called M-step decodable POMDP (when $M = 1$, it is Block MDP) and the efficient algorithm.

Our framework captures *all* above mentioned POMDP models. In addition, we propose a new model called HSE-POMDPs which extends prior works on HSE-HMM[8] to POMDPs and includes LQG and tabular POMDPs. Our algorithm delivers the first PAC bound for this model.

2 Preliminary

We introduce background for POMDPs here and defer the introduction of PSRs to Section 6. We consider an episodic POMDP specified by $\mathcal{M} = \langle \mathcal{S}, \mathcal{O}, \mathcal{A}, H, \mathbb{T}, \mathbb{O} \rangle$, where $\mathcal{S}$ is the *unobserved* state space, $\mathcal{O}$ is the observation space, $\mathcal{A}$ is the action space, H is the horizon, $\mathbb{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability, $\mathbb{O} : \mathcal{S} \to \Delta(\mathcal{O})$ is the emission probability, and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward. Here, $\mathbb{T}, \mathbb{O}$ are unknown distributions. For notation simplicity, we consider the time-homogeneous case; however, the extension to the time-inhomogeneous setting is straightforward.

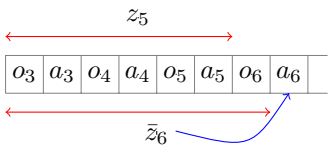

Figure 1: Case with $M = 3$. A 3-memory policy determines action a_6 based on $\bar{z}_6$.

In our work, we consider M-memory policies. Let $\mathcal{Z}_h = (\mathcal{O} \times \mathcal{A})^{\min\{h,M\}}$, i.e., for a $z_h \in \mathcal{Z}_h$, we have $z_h = [o_{\max(h-M+1,1):h}, a_{\max(h-M+1,1):h}]$, and $\bar{\mathcal{Z}}_h = \mathcal{Z}_{h-1} \times \mathcal{O}$. An element in $\bar{\mathcal{Z}}_h$ is represented as $\bar{z}_h = [o_{\max(h-M,1):h}, a_{\max(h-M,1):h-1}]$ (thus, $\bar{z}_h = [z_{h-1}, o_h]$). Figure 2 illustrates this situation. An M-memory policy is defined as $\pi = \{\pi_h\}_{h=1}^H$ where each π_h is a mapping from $\bar{\mathcal{Z}}_h$ to a distribution over actions $\Delta(\mathcal{A})$.

In a POMDP, an M-memory policy generates the data as follows. Each episode starts with initial state s_1 sampled from some unknown distribution. At each step $h \in [H]$, from $s_h \in \mathcal{S}$, the agent observes $o_h \sim \mathbb{O}(\cdot|s_h)$, executes action $a_h \sim \pi_h(\cdot|\bar{z}_h)$, receives reward $r(s_h, a_h)$, and transits to the next latent state $s_{h+1} \sim \mathbb{T}(\cdot|s_h, a_h)$. We denote $J(\pi)$ as the value of the policy π, i.e., $\mathbb{E}_\pi[\sum_{h=1}^H r_h; a_{1:H} \sim \pi]$ where the expectation is taken regarding the induced distribution by π.

We define a value function at step h to be the expected cumulative reward to go starting from $s \in \mathcal{S}, z \in \mathcal{Z}_{h-1}$. It is defined as $V_h^\pi : \mathcal{Z}_{h-1} \times \mathcal{S} \to \mathbb{R}$ where $V_h^\pi(z, s) = \mathbb{E}[\sum_{h'=h}^H r_{h'} \mid z_{h-1} = z, s_h = s; a_{h:H} \sim \pi]$. The notation $\mathbb{E}[\cdot\,; a_{h:H} \sim \pi]$ means the expectation is taken under a policy π from h to H. Compared to the standard MDP setting, the expectation is conditional on not only s_h but also z_{h-1} since we consider M-memory policies. The corresponding Bellman equation for V_h^π is $V_h^\pi(z_{h-1}, s_h) = \mathbb{E}\left[r_h + V_{h+1}^\pi(z_h, s_{h+1}) \mid z_{h-1}, s_h; a_h \sim \pi\right]$.

The Actor-critic function approximation setup Our goal is to find a near optimal policy that maximizes the policy value $J(\pi)$ in an online manner. Since any POMDPs can be converted into MDPs by setting the state at level h to the observable history up to h, any off-the-shelf online provably efficient algorithms for MDPs can be applied to POMDPs. By defining $\mathcal{H}_h$ as the whole history up to step $h \in [H]$, these naïve algorithms ensure that output policies can compete against the globally optimal policy $\pi_{gl}^\star = \operatorname{argmax}_{\pi \in \bar{\Pi}} J(\pi)$ where $\tilde{\Pi} = \{\bar{\Pi}_h\}, \tilde{\Pi}_h = [\mathcal{H}_h \to \Delta(\mathcal{A})]$. However, this conversion results in the error with exponential dependence on the horizon H, which is prohibitively large in the long horizon tabular setting.

Instead of directly competing against the globally optimal policy, we aim for *agnostic* policy learning, i.e., our goal is to compete against the best M-memory policy. Our function approximation setup

consists of two function classes. Given a policy class Π consisting of M-memory policies $\Pi := \{\Pi_h\}_{h=1}^H$ (i.e., actors), where $\Pi_h \subset [\bar{\mathcal{Z}}_h \to \Delta(\mathcal{A})]$. The second is a set of value bridge functions $\mathcal{G} = \{\mathcal{G}_h\}_{h=1}^H$ where $\mathcal{G}_h \subset [\overline{\mathcal{Z}}_h \to \mathbb{R}]$, of which the role is to approximate V_h^π (i.e., critics). We want to find an algorithm which outputs a policy $\hat{\pi} = \{\hat{\pi}_h\}$ so that it has a low excess risk defined by $R(\pi) := J(\hat{\pi}) - J(\pi^\star)$ where $\pi^\star = \operatorname{argmax}_{\pi\in\Pi} J(\pi)$ is the best in class Π. In practice, M-memory policies are also widely used, e.g., DQN [44] sets $M = 4$. Besides, there are natural examples where M-memory policies are close to the globally optimal policy with M being only polynomial with respect to parameters of problem instances rather than exponential, e.g., observable POMDPs [20] and LQG [36, 49, 41]. We will show the global optimatliy in these two examples later without exponential dependence on H.

Remark 1 (Limits of existing MDP actor-critic framework)**.** *While general actor-critic framework proposed in MDPs [27] is applicable to POMDPs via the naïve POMDP to MDP reduction, it will not be able to leverage the benefit of restricted policy class. Such reduction uses full history and will incur sample complexity that scales exponentially with respect to the horizon.*

Notation We streamline the notation as follows. Let $[H] = \{1, \cdots, H\}, [t] = \{1, \cdots, t\}$. Give a matrix A, we denote its pseudo inverse by $A^\dagger$ and the operator norm by $\|A\|$. The outer product is denoted by $\otimes$. Let $d_h^\pi(\cdot) \in \bar{Z}_h \times \mathcal{S}$ be the marginal distribution at h and $\delta(\cdot)$ be the Dirac delta function. We denote the policy $\delta(a = a')$ by $\mathrm{do}(a')$. We denote a uniform action by $U(\mathcal{A})$.

3 Value Bridge Functions and the PO-bilinear Framework

Unlike MDPs, in POMDPs, we cannot directly work with value functions V_h^π (or Q functions) since they depend on unobserved quantities s. To handle this issue, below we first introduce new value bridge functions by using future observations. Then, we introduce the PO-bilinear framework.

3.1 Value Bridge Functions

Definition 1 (K-step value bridge functions)**.** *Fix some set of M-memory policies $\pi^{out} = \{\pi_i^{out}\}_{i=1}^K$. Value bridge functions $g_h^\pi : \mathcal{Z}_{h-1} \times \mathcal{O}^K \times \mathcal{A}^{K-1} \to \mathbb{R}$ at step $h \in [H]$ are defined as the solution to the following integral equation : $\forall z_{h-1} \in \mathcal{Z}_{h-1}, s_h \in \mathcal{S}$,*

$$\mathbb{E}[g_h^\pi(z_{h-1}, o_{h:h+K-1}, a_{h:h+K-2}) \mid z_{h-1}, s_h; a_{h:h+K-2} \sim \pi^{out}] = V_h^\pi(z_{h-1}, s_h). \tag{1}$$

where the expectation is taken under a policy π^{out}.

Bridge functions do not necessarily exist nor need to be unique. K-step value bridge functions are embeddings of value functions onto the observation space. At a high level, its existence means that K-step futures have sufficient information to recover states. We will later discuss the choice of π^{out} in Section B.1. For example, we use uniform policy in the tabular case, and $\delta(a = 0)$ in LQG. For notational simplicity, we often focus on the case $K = 1$. The definition of the 1-step bridge functions is below. This definition is agnostic to π^{out}.

Definition 2 (1-step value bridge functions)**.** *One-step value bridge functions $g_h^\pi : \mathcal{Z}_{h-1} \times \mathcal{O} \to \mathbb{R}$ at step $h \in [H]$ are defined as the solution to the following integral equation:*

$$\forall z_{h-1} \in \mathcal{Z}_{h-1}, s_h \in \mathcal{S} : \quad \mathbb{E}[g_h^\pi(z_{h-1}, o_h) \mid z_{h-1}, s_h] = V_h^\pi(z_{h-1}, s_h). \tag{2}$$

In Section 4, we will demonstrate the form of the value bridge function through several examples.

The idea of encoding latent state information using the future has been widely used in learning models of HMMs [51, 25], PSRs [9]. Existing provably efficient (online) RL works for POMDPs elaborate on this viewpoint [30, 21, 4]. Compared to them, the novelty of bridge functions is that it is introduced to recover *value functions* but not *models*. This model-free view differs from the existing dominant model-based view in online RL for POMDPs. Our idea is that we can control systems if we can recover value functions on the underlying states even if we fail to identify underlying models.

3.2 The PO-Bilinear Actor-critic Framework for POMDPs

With the definition of value bridge functions, we are now ready to introduce the PO-bilinear class for POMDPs. We will focus on $K = 1$ here. We introduce a function class $\mathcal{G} = \{\mathcal{G}_h\}_{h=1}^H$ where $\mathcal{G}_h : \bar{\mathcal{Z}}_h \to \mathbb{R}$. We first need the following realizability assumption.

Assumption 1 (Realizability)**.** *We assume $\mathcal{G}$ is realizable, i.e., $\forall \pi \in \Pi, h \in [H]$,* at least one *value bridge function g_h^π is in $\mathcal{G}_h$. Note realizability implicitly requires the existence of bridge functions.*

Our key observation to introduce the PO-bilinear AC framework is that value bridge functions satisfy

$$0 = \mathrm{Br}_h(\pi, g^\pi; \pi^{in}),\ \mathrm{Br}_h(\pi, g; \pi^{in}) := \mathbb{E}[g_h(\bar{z}_h) - r_h - g_{h+1}(\bar{z}_{h+1}) : a_{1:h-1} \sim \pi^{in}, a_h \sim \pi].$$

for any M memory roll-in policy $\pi^{in} = \{\pi_h^{in}\}_{h=1}^H$, and any evaluation pair (π, g^π). This is an analog of Bellman equations on MDPs. The above equation tells us that $\mathrm{Br}_h(\pi, g; \pi^{in})$ is a right loss to quantify how much the estimator g is different from g_h^π. When this $\mathrm{Br}_h(\pi, g; \pi^{in})$ has a low-rank structure in a proper way, we can efficiently learn a near optimal M memory policy. The following definition precisely quantifies the low-rank structure we need.

Definition 3 (PO-bilinear AC Class)**.** *The model is a PO-bilinear AC class of rank d if $\mathcal{G}$ is realizable, and there exist $W_h : \Pi \times \mathcal{G} \to \mathbb{R}^d$ and $X_h : \Pi \to \mathbb{R}^d$ such that for all $\pi', \pi \in \Pi, g \in \mathcal{G}$ and $h \in [H]$,*

1. $|\mathbb{E}[g_h(\bar{z}_h) - r_h - g_{h+1}(\bar{z}_{h+1}); a_{1:h-1} \sim \pi', a_h \sim \pi]| \leq |\langle W_h(\pi, g), X_h(\pi')\rangle|$.
2. $W_h(\pi, g^\pi) = 0$ *for any* $\pi \in \Pi$ *and the corresponding value bridge function* g^π *in* $\mathcal{G}$ *.*

We call d as the PO-bilinear rank.

Remark 2. *The above definition is enough to capture most of the examples we discuss later in this work, including undercomplete tabular POMDPs, LQGs, HSE-POMDPs. In Appendix B, we provide two extensions. The first extension incorporates the minimax framework which can be used to capture the M-step decodable POMDPs. The second extension handles multiple-step futures (i.e., $K > 1$), which can be used to capture examples like overcomplete POMDPs and general PSRs.*

4 Examples of PO-Bilinear Actor-Critic Classes

We consider three examples that admit PO-bilinear rank decomposition. Our framework can also capture PSRs and M-step decodable POMDPs, of which the discussions are deferred to Section 6 and Section C, respectively. We mainly focus on one-step future, i.e., $K = 1$, and briefly discuss extension to $K > 1$.

4.1 Observable Tabular POMDPs

Example 1 (Observable undercomplete tabular POMDPs)**.** *Let $\mathbb{O} \in \mathbb{R}^{|\mathcal{O}| \times |\mathcal{S}|}$ where the entry indexed by a pair (o, s) is defined as $\mathbb{O}_{o,s} = \mathbb{O}(o|s)$. Assume $\mathrm{rank}(\mathbb{O}) = |\mathcal{S}|$, which we call observability. This requires undercompletenes $|\mathcal{O}| \geq |\mathcal{S}|$.*

The following lemma shows that $\mathbb{O}$ being full rank implies the existence of the value bridge functions.

Lemma 1. *In Example 1, for all $\pi \in \Pi, h \in [H]$, a one-step value bridge function g_h^π exists.*

Proof Consider any function $f : \mathcal{Z}_{h-1} \times \mathcal{S} \to \mathbb{R}$ (thus, this captures all possible V_h^π). Denote $\mathbf{1}(z)$ as the one-hot encoding of z over $\mathcal{Z}_{h-1}$ (similarly for $\mathbf{1}(s)$). We have $f(z, s) = \langle f, \mathbf{1}(z) \otimes \mathbf{1}(s)\rangle = \langle f, \mathbf{1}(z) \otimes (\mathbb{O}^\dagger \mathbb{O} \mathbf{1}(s))\rangle$, where we use the assumption that $\mathrm{rank}(\mathbb{O}) = |\mathcal{S}|$ and thus $\mathbb{O}^\dagger \mathbb{O} = I$. Then,

$$f(z, s) = \langle f, \mathbf{1}(z) \otimes (\mathbb{O}^\dagger \mathbb{E}_{o \sim O(s)} \mathbf{1}(o))\rangle = \mathbb{E}_{o \sim O(s)} \langle f, \mathbf{1}(z) \otimes \mathbb{O}^\dagger \mathbf{1}(o)\rangle,$$

which means that the value bridge function corresponding to f is $g(z, o) := \langle f, \mathbf{1}(z) \otimes \mathbb{O}^\dagger \mathbf{1}(o)\rangle$. □

Now we show that tabular POMDPs have PO-Bilinear rank at most $|\mathcal{S}|(|\mathcal{O}||\mathcal{A}|)^M$ using Definition 3.

Lemma 2. *Assume $\mathbb{O}$ is full column rank. Set value function class $\mathcal{G}_h = [\mathcal{Z}_{h-1} \times \mathcal{O} \to \mathbb{R}]$, policy class $\Pi_h = [\bar{\mathcal{Z}}_h \to \Delta(\mathcal{A})]$. Then, the model is a PO-biliner AC class (Definition 3) with PO-bilinear rank at most $|\mathcal{S}|(|\mathcal{O}||\mathcal{A}|)^M$.*

Lastly, we touch on overcomplete POMDPs ($|\mathcal{O}| \leq |\mathcal{S}|$) when we use multi-step futures. For details, refer to Section C.2. In this case, the existence of bridge function is ensured when $|\mathcal{O}|^K |\mathcal{A}|^{K-1} \times |\mathcal{S}|$ matrix with entries equal to $\mathbb{P}(o_{h:h+K-1}, a_{h:h+K-2} \mid s_h; a_{h:h+K-2} \sim U(\mathcal{A}))$ is full-column rank ($\pi_t^{out} = U(\mathcal{A})$). Then, the bilinear rank is still $|\mathcal{S}|(|\mathcal{O}||\mathcal{A}|)^M$. Crucially, it does not depend on the length of futures K.

4.2 Observable Linear Quadratic Gaussian

The next example is LQG with continuous state and action. The detail is deferred to Section K. Here, we set $M = H - 1$ so that the policy class Π can contain the globally optimal policy.

Example 2 (Linear Quadratic Gaussian (LQG)). *Consider LQG:*

$$s' = As + Ba + \epsilon, o = Cs + \tau, r = -(s^\top Qs + a^\top Ra)$$

where ϵ, τ are Gaussian distribution with mean 0 and variance $\Sigma_\epsilon, \Sigma_\tau$, respectively, and $s \in \mathbb{R}^{d_s}, o \in \mathbb{R}^{d_o}$, and $a \in \mathbb{R}^{d_a}$, and Q, R are positive definite matrices.

We restrict the policy class to be a linear policy class $\Pi_h = \{\delta(a_h = \bar{K}_h \bar{z}_h) : K_h \in \mathbb{R}^{|\mathcal{A}| \times d_{\bar{z}_h}}\}$ where $d_{\bar{z}_h}$ is a dimension of $\bar{z}_h$. This choice is natural since the globally optimal policy is known to be linear with respect to the entire history [7, Chapter 4]. We define two quadratic features, $\phi_h(z_{h-1}, s_h) = (1, [z_{h-1}^\top, s_h^\top] \otimes [z_{h-1}^\top, s_h^\top])^\top$ with $z_{h-1} \in \mathcal{Z}_{h-1}, s_h \in \mathcal{S}$, and $\psi_h(z_{h-1}, o_h) = (1, [z_{h-1}^\top, o_h^\top] \otimes [z_{h-1}^\top, o_h^\top])^\top$ with $z_{h-1} \in \mathcal{Z}_{h-1}, o_h \in \mathcal{O}$. We have the following lemma.

Lemma 3 (PO-bilinear rank in observable LQG). *Assume $\mathrm{rank}(C) = d_s$. Then, the following holds.*

- *For any policy π linear in $\bar{z}_h$, a one-step value bridge function $g_h^\pi(\cdot)$ exists, and is linear in $\psi_h(\cdot)$.*
- *Letting d_{ψ_h} be the dimension of ψ_h, we set $\mathcal{G}_h = \{\theta^\top \psi_h(\cdot) | \theta \in \mathbb{R}^{d_{\psi_h}}\}$ and Π being linear in $\bar{z}_h$. Then LQG satisfies Definition 3 with PO-bilinear rank at most $O(\{1 + (H-1)(d_o + d_a) + d_s\}^2)$*

We have two remarks. First, when $\pi_t^{out} = \delta(a = 0)$, K-step bridge functions exist when $[C^\top, (CA)^\top, \ldots, (CA^{K-1})^\top]^\top$ is full column rank. This assumption is referred to as observability in control theory [24]. Secondly, the PO-bilinear rank scales polynomially with respect to H, d_o, d_a, d_s even with $M = H - 1$. As we show in Section K, due to this fact, we can compete against the *globally* optimal policy with polynomial sample complexity.

4.3 Observable Hilbert Space Embedding POMDPs

We consider HSE-POMDPs which generalize tabular POMDPs and LQG. Proofs here are deferred to Section C.4. Consider any $h \in [H]$. Given a policy $\pi_h : \bar{\mathcal{Z}}_h \to \mathcal{A}$, we define the induced transition operator $\mathbb{T}_\pi = \{\mathbb{T}_{\pi;h}\}_{h=1}^H$ as $(z_h, s_{h+1}) \sim \mathbb{T}_{\pi;h}(z_{h-1}, s_h)$, where we have $o_h \sim \mathbb{O}(s_h), a_h \sim \pi_h(\bar{z}_h), s_{h+1} \sim \mathbb{T}(s_h, a_h)$. Namely, $\mathbb{T}_\pi$ is the transition kernel of the HMM induced by the policy π. The HSE-POMDP assumes two conditional distributions $\mathbb{O}(\cdot|s)$ and $\mathbb{T}_\pi(\cdot, \cdot|z, s)$ have conditional mean embeddings.

Example 3 (HSE-POMDPs). *We introduce features $\phi_h : \mathcal{Z}_{h-1} \times \mathcal{S} \to \mathbb{R}^{d_{\phi_h}}, \psi_h : \mathcal{Z}_{h-1} \times \mathcal{O} \to \mathbb{R}^{d_{\psi_h}}$. We assume the existence of the conditional mean embedding operators: (1) there exists a matrix K_h such that for all $z \in \mathcal{Z}_{h-1}, s \in \mathcal{S}$, $\mathbb{E}_{o \sim \mathbb{O}(\cdot|s)} \psi_h(z, o) = K_h \phi_h(z, s)$ and (2) for all $\pi \in \Pi$, there exists a matrix $T_{\pi;h}$, such that $\mathbb{E}_{z_h, s_{h+1} \sim \mathbb{T}_{\pi;h}(z_{h-1}, s_h)} \phi_{h+1}(z_h, s_{h+1}) = T_{\pi;h} \phi_h(z_{h-1}, s_h)$.*

The existence of conditional mean embedding is a common assumption in prior RL works on learning dynamics of HMMs, PSRs, [52, 8] and Bellman completeness linear MDPs [61, 16, 11, 22]. HSE-POMDPs naturally capture tabular POMDPs and LQG. For tabular POMDPs, ψ_h and ϕ_h are one-hot encoding features. In LQG, ϕ_h and ψ_h are quadratic features we define in Section 4.2. Here for simplicity, we focus on finite-dimensional features ϕ_h and ψ_h. Extension to infinite-dimensional Reproducing kernel Hilbert Space is deferred to Section C.4.

The following shows the existence of value bridge functions and the PO-bilinear rank decomposition.

Lemma 4 (PO-bilinear rank in observable HSE-POMDPs). *Assume K_h is full column rank (observability), and $V_h^\pi(\cdot)$ is linear in ϕ_h for any $\pi \in \Pi, h \in [H]$. Then the following holds.*

- *A one-step value bridge function $g_h^\pi(\cdot)$ exists for any $\pi \in \Pi, h \in [H]$, and is linear in ψ_h.*
- *We set a value function class $\mathcal{G}_h = \{w^\top \psi_h(\cdot) | w \in \mathbb{R}^{d_{\psi_h}}\}$, policy class $\Pi_h \subset [\bar{\mathcal{Z}}_h \to \Delta(\mathcal{A})]$. Then HSE-POMDP satisfies Definition 3 with PO-bilinear rank at most $\max_{h \in [H]} d_{\phi_h}$.*

The first statement can be verified by noting that when $V_h^\pi(\cdot) = \langle \theta_h, \phi_h(\cdot) \rangle$, value bridge functions take the following form $g_h^\pi(\cdot) = \langle (K_h^\dagger)^\top \theta_h), \psi_h(\cdot) \rangle$ where we leverage the existence of the conditional mean embedding operator K_h, and that K_h is full column rank (thus $K_h^\dagger K_h = \mathbb{I}_{d_{\phi_h}}$). Note that the PO-bilinear rank depends only on the dimension of the features ϕ_h without any explicit dependence on the length of memory.

Algorithm 1 PaRtially ObserVAble BiLinEar (**PROVABLE**) # multi-step version is in Section D

Input: Value class $\mathcal{G}$, Policy class Π, parameters m, R
Initialize $\pi^0 \in \Pi$
Form the first step dataset $\mathcal{D}^0 = \{o^i\}_{i=1}^m$, with $o^i \sim \mathbb{O}(\cdot|s_1)$
for $t = 0 \to T-1$ **do**
  For any $h \in [H]$, collect m i.i.d tuple as follows: $(\bar{z}_h, s_h) \sim d_h^{\pi^t}, a_h \sim U(\mathcal{A}), r_h = r_h(s_h, a_h), s_{h+1} \sim \mathbb{T}(s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(\cdot|s_{h+1})$.
  Define $\mathcal{D}_h^t = \{(\bar{z}_h^i, a_h^i, r_h^i, o_{h+1}^i)\}_{i=1}^m$ # note latent state s is not in the dataset
  Define the Bellman error $\forall(\pi, g) \in \Pi \times \mathcal{G}$,

$$\sigma_h^t(\pi, g) := \mathbb{E}_{\mathcal{D}_h^t}\left[\pi_h(a_h \mid \bar{z}_h)|\mathcal{A}|\{g_{h+1}(\bar{z}_{h+1}) + r_h - g_h(\bar{z}_h)\right].$$

  Select policy optimistically as follows

$$(\pi^{t+1}, g^{t+1}) := \text{argmax}_{\pi\in\Pi, g\in\mathcal{G}} \mathbb{E}_{\mathcal{D}^0}[g_1(o)] \quad \text{s.t.} \quad \forall h \in [H], \forall i \in [t], (\sigma_h^i(\pi, g))^2 \leq R.$$

end for
Output: Randomly choose $\hat{\pi}$ from $(\pi_1, \cdots, \pi_T)$.

5 Algorithm and Sample Complexity

In this section, we first give our algorithm, followed by a general sample complexity analysis. We then instantiate our analysis to specific models considered in Section 4.

5.1 Algorithm

Here, we focus on the cases where models satisfy the PO-Bilinear AC model (i.e., Definition 3) with finite action. We briefly discuss the extension to handle more general models in Remark 3.

We present our algorithm **PROVABLE** in Algorithm 1. Note **PROVABLE** is agnostic to the form of X_h and W_h. Inside iteration t, given the latest learned policy π^t, we define Bellman error for all pairs (π, g) where the Bellman error is averaged over the samples from π^t. A pair (π, g) that has a small total Bellman error intuitively means that given the data so far, g could still be a value bridge function for the policy π. Then in the constrained optimization formulation, we only focus on (π, g) pairs whose Bellman errors are small so far. Among these (π, g) pairs, we select the pair using the principle of optimism in the face of uncertainty. We remark the algorithm leverages some design choices from the Bilinear-UCB algorithm for MDPs [14]. The key difference between our algorithm and the Bilinear-UCB is that we leverage the actor-critic framework equipped with value bridge functions to handle partially observability and agnostic learning.

Remark 3 (Three key extensions: minimax version, continuous control, multi-step futures). *By extending Algorithm 1, we can consider more general algorithms to include three important cases. The first extension is the minimax version with discriminators for M-step decodable POMDPs. The second extension is for LQG, which has continuous action. By employing a G-optimal design over actions, our algorithm can handle the continuous action. The third extension is the multi-step future case, which can capture overcomplete POMDPs and general PSRs. The discussion is deferred to Section D.*

5.2 Sample Complexity

We show a sample complexity result by using reduction to supervised learning analysis. We begin by stating the following assumption which is ensured by standard uniform convergence results.

Assumption 2 (Uniform Convergence). *Fix $h \in [H]$. Let $\mathcal{D}'_h$ be a set of m i.i.d tuples following $(z_{h-1}, s_h, o_h) \sim d_h^{\pi^t}, a_h \sim U(\mathcal{A}), s_{h+1} \sim \mathbb{T}(s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(s_{h+1})$. With probability $1-\delta$,*

$$\sup_{\pi\in\Pi, g\in\mathcal{G}} |(\mathbb{E}_{\mathcal{D}'_h} - \mathbb{E})[\pi_h(a_h \mid \bar{z}_h)|\mathcal{A}|\{g_{h+1}(\bar{z}_{h+1}) + r_h - g_h(\bar{z}_h)\}]| \leq \epsilon_{gen,h}(m, \Pi, \mathcal{G}, \delta)$$

When $h = 1$, we also require $\sup_{g_1\in\mathcal{G}_1} |\mathbb{E}_{\mathcal{D}'_1}[g_1(o_1)] - \mathbb{E}[\mathbb{E}_{\mathcal{D}'_1}[g_1(o_1)]]| \leq \epsilon_{ini,1}(m, \mathcal{G}, \delta)$.

Remark 4. *As a simple example, we consider the case where Π and $\mathcal{G}$ are discrete. In this case, we have $\epsilon_{gen,h}(m, \Pi, \mathcal{G}, \delta) = O(\sqrt{\ln(|\Pi||\mathcal{G}|/\delta)/m})$, and $\epsilon_{ini,1}(m, \mathcal{G}, \delta) = O(\sqrt{\ln(|\mathcal{G}|/\delta)/m})$, which are standard statistical complexities for discrete function classes Π and $\mathcal{G}$.*

Under Assumption 2, when the model is PO-bilinear with rank d, we get the following.

Theorem 1 (PAC guarantee of **PROVABLE**). *Suppose we have a PO-bilinear AC class with rank d. Suppose Assumption 2, $\sup_{\pi\in\Pi}\|X_h(\pi)\| \le B_X$ and $\sup_{\pi\in\Pi,g\in\mathcal{G}}\|W_h(\pi,g)\| \le B_W$ for any $h\in[H]$. By setting $T = 2Hd\ln\left(4Hd\left(\frac{B_X^2B_W^2}{\tilde{\epsilon}_{gen}}+1\right)\right)$, $R=\epsilon_{gen}^2$ where*

$$\epsilon_{gen} := \max_h \epsilon_{gen,h}(m,\Pi,\mathcal{G},\delta/(TH+1)), \tilde{\epsilon}_{gen} := \max_h \epsilon_{gen,h}(m,\Pi,\mathcal{G},\delta/H).$$

With probability at least $1-\delta$, letting $\pi^\star = \mathrm{argmax}_{\pi\in\Pi} J(\pi^\star)$, we have

$$J(\pi^\star) - J(\hat{\pi}) \le 5\epsilon_{gen}\sqrt{dH^2\ln\left(4Hd\left(\frac{B_X^2B_W^2}{\tilde{\epsilon}_{gen}^2}+1\right)\right)} + 2\epsilon_{ini,1}(m,\mathcal{G},\delta/(TH+1)).$$

The total number of samples used in the algorithm is mTH.

5.3 Examples

Hereafter, we show the sample complexity result by using Theorem 1. For complete results, refer to Section G–M. The result of M-step decodable POMDPs is deferred to Section L.

Example 1 (continuing from p. 5). *In tabular models, recall the PO-bilinear rank is at most $d = |\mathcal{O}|^M|\mathcal{A}|^M|\mathcal{S}|$. We suppose $r_h\in[0,1]$ for any $h\in[H]$. Assuming $\mathbb{O}$ is full-column rank, to satisfy the realizability, we set $\mathcal{G}_h = \{\langle\theta,\psi_h(\cdot)\rangle \mid \|\theta\|_1 \le H|\mathcal{O}|^M|\mathcal{A}|^M|\mathcal{S}|/\sigma_1\}$ where $\sigma_1 = \min_h 1/\|K_h^\dagger\|_1$ and ψ_h is a one-hot encoding vector over $\bar{\mathcal{Z}}_h$. We set $\Pi = [\bar{Z}_h \to \Delta(\mathcal{A})]$. Then, the following holds.*

Theorem 2 (Sample complexity for tabular models (Informal)). *With probability $1-\delta$, we can achieve $J(\pi^\star) - J(\hat{\pi}) \le \epsilon$ when we use samples at most $\tilde{O}\left(|\mathcal{S}|^4|\mathcal{A}|^{6M}|\mathcal{O}|^{6M}H^7(1/\epsilon)^2(1/\sigma_1)^2\ln(1/\delta)\right)$. Here, $\mathrm{polylog}(|\mathcal{S}|,|\mathcal{O}|,|\mathcal{A}|,H,1/\sigma_1,\ln(1/\delta))$ are omitted.*

As shown in [20], it is enough to set $M = O((1/\sigma_1^4)\ln(SH/\epsilon))$ to complete with the globally optimal policy $\pi_{gl}^\star$.Thus we achieve a quasipolynomial sample complexity when competing against $\pi_{gl}^\star$.

Theorem 3 (Sample complexity for tabular models (Informal) — competing against $\pi_{gl}^\star$). *With probability $1-\delta$, we can achieve $J(\pi_{gl}^\star) - J(\hat{\pi}) \le \epsilon$ when we use samples at most*

$$\mathrm{poly}(|\mathcal{S}|,|\mathcal{A}|^{\ln(|\mathcal{S}|H/\epsilon)/\sigma_1^4},|\mathcal{O}|^{\ln(|\mathcal{S}|H/\epsilon)/\sigma_1^4},H,1/\sigma_1,1/\epsilon,\ln(1/\delta)).$$

Example 2 (continuing from p. 6). *In LQG, by setting $H = M-1$, we achieve a polynomial sample complexity when competing against the globally optimal policy $\pi_{gl}^\star$.*

Theorem 4 (Sample comlexity for LQG (informal) – competing against $\pi_{gl}^\star$). *Consider a linear policy class $\Pi_h = \{\delta(a_h = \bar{K}_{1h}\bar{z}_h) \mid \|\bar{K}_h\| \le \Theta\}$. and assume $\max(\|A\|,\|B\|,\|C\|,\|Q\|,\|R\|) \le \Theta$ and all policies induce a stable system (we formalize in Section K). With probability $1-\delta$, we can achieve $J(\pi_{gl}^\star) - J(\hat{\pi}) \le \epsilon$ when we use samples at most*

$$\mathrm{poly}(H,d_s,d_o,d_a,\Theta,\|C^\dagger\|,\ln(1/\delta)) \times (1/\epsilon)^2.$$

Example 3 (continuing from p. 6). *In HSE-POMDPs, PO-bilinear rank is at most $\max_h d_{\phi_h}$. Suppose $\|\psi_h\| \le 1$ and $V_h^\pi(\cdot) = \langle\theta_h^\pi,\phi_h(\cdot)\rangle$ such that $\|\theta_h^\pi(\cdot)\| \le \Theta_V$ for any $h\in[H]$. Then, to satisfies the realizability, we set $\mathcal{G}_h = \{\langle\theta,\psi_h(\cdot)\rangle \mid \|\theta\| \le \Theta_V/\sigma_{\min}(K)\}$ where $\sigma_{\min}(K) = \min_{h\in[H]} 1/\|K_h^\dagger\|$.*

Theorem 5 (Sample complexity for HSE-POMDPs (Informal)). *Let $d_\psi = \max_h\{d_{\psi_h}\}, d_\phi = \max_h\{d_{\psi_h}\}, |\Pi_{\max}| = \max_h(|\Pi_h|)$. Suppose r_h lies in $[0,1]$ for any $h\in[H]$. Then, with probability $1-\delta$, we can achieve $J(\pi^\star) - J(\hat{\pi}) \le \epsilon$ when we use samples*

$$\tilde{O}\left(d_\phi^2H^5|\mathcal{A}|^2\max(\Theta_V,1)^2\{d_\psi + \ln(|\Pi_{\max}|/\delta)\}(1/\sigma_{\min}(K))^2\cdot(1/\epsilon)^2\right).$$

Here, $\mathrm{polylog}(d_\phi,d_\psi,|\mathcal{A}|,\Theta_V,\ln(|\Pi_{\max}|),1/\sigma_{\min}(K),1/\epsilon,\ln(1/\delta),\sigma_{\max}(T),\sigma_{\max}(K))$ are omitted where $\sigma_{\max}(K) = \max_{h\in[H]}\|K_h\|, \sigma_{\max}(T) = \max_{\pi\in\Pi,h\in[H]}\|T_{\pi:h}\|$.

6 Predictive State Representations

In this section, we demonstrate that our definition and algorithm applies to PSRs — models that strictly generalize POMDPs [38, 50]. Below, we first briefly introduce PSRs, followed by showing that it is a PO-bilinear AC model. Throughout this section, we will focus on discrete linear PSRs.

We also suppose reward at h is deterministic function of (o_h, a_h) conditional on τ_{h+1} where $\tau_h = (o_1, a_1, \cdots, o_{h-1}, a_{h-1})$. Given τ_h, the dynamical system generates $o_h \sim \mathbb{P}(\cdot|\tau_h)$.

PSRs use the concept of *test*, which is a sequence of future observations and actions, i.e., for some test $t = (o_{h:h+W-1}, a_{h:h+W-2})$ with length $W \in \mathbb{N}^+$, we define the probability of test t being successful $\mathbb{P}(t|\tau_h)$ as $\mathbb{P}(t|\tau_h) := \mathbb{P}(o_{h:h+W-1}|\tau_h; \mathrm{do}(a_{h:h+W-2}))$ which is the probability of observing $o_{h:h+W-1}$ by actively executing actions $a_{h:h+W-2}$ conditioned on history τ_h.

We now explain one-step observable PSRs while deferring the general multi-step observable setting to Section E. A one-step observable PSR uses the observations in $\mathcal{O}$ as tests, i.e., tests with length 1.

Definition 4 (Core test set and linear PSRs)**.** *A core test set $\mathcal{T} \subset \mathcal{O}$ contains a finite number of tests (i.e., observations from $\mathcal{O}$). For any h, any history τ_h, any future test $t_h = (o_{h:h+W-1}, a_{h:h+W-2})$ for any $W \in \mathbb{N}^+$, there exists a vector $m_{t_h} \in \mathbb{R}^{|\mathcal{T}|}$, such that the probability of t_h succeeds conditioned on τ_h can be expressed as: $\mathbb{P}(t_h|\tau_h) = m_{t_h}^\top [\mathbb{P}(o|\tau_h)]_{o\in\mathcal{T}}$, where we denote $\mathbf{q}_{\tau_h} := [\mathbb{P}(o|\tau_h)]_{o\in\mathcal{T}}$ as a vector in $\mathbb{R}^{|\mathcal{T}|}$ with entries equal to $\mathbb{P}(o|\tau_h)$ for $o \in \mathcal{T}$. The vector $\mathbf{q}_{\tau_h}$ is called predictive state.*

PSRs are strictly more expressive than POMDPs in that all POMDPs can be embedded into PSRs whose size of core tests is at most $|\mathcal{S}|$; however, vice versa does not hold [38]. A core test set $\mathcal{T}$ that has the smallest number of tests is called a *minimum core test set* denoted as $\mathcal{Q}$. For example, in observable undercomplete POMDPs (i.e., $\mathbb{O}$ full column rank) , the observation set $\mathcal{O}$ can serve as a core test set, but the minimum core test set $\mathcal{Q}$ will have size $|\mathcal{S}|$. Hence, we assume we know a core test set $\mathcal{T}$ that contains $\mathcal{Q}$; however, we are agnostic to which set is the actual $\mathcal{Q}$. In the literature on PSRs, this setting is often referred to as transform PSRs [9, 47].

Now we define a value bridge function in PSRs. First, given a M-memory policy, define $\mathcal{V}_h^\pi(\tau_h) = \mathbb{E}[\sum_{t=h}^H r_t|\tau_h; a_{h:H} \sim \pi]$, i.e., the expected total reward under π, conditioned on the history τ_h.

Definition 5 (General value bridge functions)**.** *Consider an M-memory policy π. One-step general value bridge functions $g_h^\pi : \mathcal{Z}_{h-1} \times \mathcal{T} \to \mathbb{R}$ at step $h \in [H]$ are defined as solutions to*

$$\mathcal{V}_h^\pi(\tau_h) = \mathbb{E}[g_h^\pi(z_{h-1}, o_h) \mid \tau_h]. \tag{3}$$

This definition is more general than Definition 3 since (3) implies (2) in POMDPs by setting $\mathcal{O} = \mathcal{T}$. In PSRs, we can show the existence of this general value bridge function.

Lemma 5 (The existence of bridge functions for PSRs)**.** *Suppose $\mathcal{T}$ is a core test set. Then, a one-step value bridge function g_h^π always exists.*

The high-level derivation is as follows. Using the linear PSR property, one can first show that $\mathcal{V}_h^\pi(\tau_h)$ has a bilinear form $\mathcal{V}_h^\pi(\tau_h) = \mathbf{1}(z_{h-1})^\top M_h^\pi \mathbf{q}_{\tau_h}$, where recall $\mathbf{1}(z) \in \mathbb{R}^{|\mathcal{O}|^M|\mathcal{A}|^M}$ denotes the one-hot encoding vector over $\mathcal{Z}_{h-1}$, and M_h^π is a matrix in $\mathbb{R}^{|\mathcal{O}|^M|\mathcal{A}|^M\times|\mathcal{T}|}$. Then, given any τ_h and $o \sim \mathbb{P}(\cdot|\tau_h)$, for some M, we can show $g_h(z_{h-1}, o) := \mathbf{1}(z_{h-1})^\top M[\mathbf{1}(t = o)]_{t\in\mathcal{T}}$ satisfies the above, where $[\mathbf{1}(t = o)]_{t\in\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}|}$ serves as an unbiased estimate of $\mathbf{q}_{\tau_h}$.

Finally, we show that PSR admits PO-bilinear rank decomposition (Definition 3).

Lemma 6. *Suppose a core test set $\mathcal{T}$ includes a minimum core test set $\mathcal{Q}$. Set $\Pi_h = [\bar{\mathcal{Z}}_h \to \Delta(\mathcal{A})]$ and $\mathcal{G}_h = \{(z_{h-1}, o) \mapsto \boldsymbol{1}(z_{h-1})^\top M[\boldsymbol{1}(t = o)]_{t\in\mathcal{T}} \mid M \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{T}|}\}$, the PO-bilinear rank is at most $(|\mathcal{O}||\mathcal{A}|)^M|\mathcal{Q}|$.*

Then, Algorithm 1 is directly applicable to PSRs and we defer the detailed sample complexity result to Section M. Note that here the PO-bilinear fortunately rank scales with $|\mathcal{Q}|$ but not $|\mathcal{T}|$. The dependence $(|\mathcal{O}||\mathcal{A}|)^M$ comes from the dimension of the "feature" of memory $\mathbf{1}(z_{h-1})$. If one has a compact feature representation $\phi : \mathcal{Z}_{h-1} \to \mathbb{R}^d$, such that $\mathcal{V}_h^\pi(\tau_h) = \phi(z_{h-1})^\top M_h^\pi \mathbf{q}_{\tau_h}$ is linear with respect to feature $\phi(z_{h-1})$, then the PO-bilinear rank is $d|\mathcal{Q}|$. This implies that if one has a compact featurization of memory a priori, one can avoid exponential dependence on M.

7 Summary

We propose a PO-bilinear actor-critic framework which is the first unified framework for provably efficient RL on large-scale POMDPs. Our framework can capture not only many models where provably efficient learning has been known such as tabular POMDPs, LQGs, M-step decodable POMDPs but also models where provably efficient RL is not known such as HSE-POMDPs and general PSRs. Our unified actor-critic based algorithm **PROVABLE** performs agnostic learning by searching for the best memory-based policy.

References

[1] Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in Neural Information Processing Systems*, 33:13399–13412, 2020.

[2] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

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

[4] Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of pomdps using spectral methods. In *Conference on Learning Theory*, pages 193–256. PMLR, 2016.

[5] Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. 2021.

[6] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[7] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012.

[8] Byron Boots, Geoffrey Gordon, and Arthur Gretton. Hilbert space embeddings of predictive state representations. *arXiv preprint arXiv:1309.6819*, 2013.

[9] Byron Boots, Sajid M Siddiqi, and Geoffrey J Gordon. Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research*, 30(7):954–966, 2011.

[10] Qi Cai, Zhuoran Yang, and Zhaoran Wang. Sample-efficient reinforcement learning for pomdps with linear function approximations. *arXiv preprint arXiv:2204.09787*, 2022.

[11] Sayak Ray Chowdhury and Rafael Oliveira. No-regret reinforcement learning with value function approximation: a kernel embedding approach. *arXiv preprint arXiv:2011.07881*, 2020.

[12] Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411*, 2020.

[13] Ben Deaner. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2018.

[14] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

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

[16] Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

[17] Yonathan Efroni, Chi Jin, Akshay Krishnamurthy, and Sobhan Miryoosefi. Provable reinforcement learning with a short-term memory. *arXiv preprint arXiv:2202.03983*, 2022.

[18] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Reinforcement learning in pomdps without resets. 2005.

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

[20] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022.

[21] Zhaohan Daniel Guo, Shayan Doroudi, and Emma Brunskill. A pac rl algorithm for episodic pomdps. In *Artificial Intelligence and Statistics*, pages 510–518. PMLR, 2016.

[22] Botao Hao, Yaqi Duan, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning*, pages 4063–4073. PMLR, 2021.

[23] Ahmed Hefny, Carlton Downey, and Geoffrey J Gordon. Supervised learning for dynamical system learning. *Advances in neural information processing systems*, 28, 2015.

[24] Joao P Hespanha. *Linear systems theory*. Princeton university press, 2018.

[25] Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[26] Herbert Jaeger. *Discrete-time, discrete-valued observable operator models: a tutorial*. GMD-Forschungszentrum Informationstechnik Darmstadt, Germany, 1998.

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

[28] Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020.

[29] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[31] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

[32] Michael Kearns, Yishay Mansour, and Andrew Ng. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

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

[34] Alex Kulesza, Nan Jiang, and Satinder Singh. Spectral learning of predictive state representations with insufficient statistics. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[35] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret minimization in partially observable linear quadratic control. *arXiv preprint arXiv:2002.00082*, 2020.

[37] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[38] Michael Littman and Richard S Sutton. Predictive representations of state. *Advances in neural information processing systems*, 14, 2001.

[39] Michael Lederman Littman. *Algorithms for sequential decision-making*. Brown University, 1996.

[40] Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967*, 2022.

[41] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

[42] Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt J Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. *arXiv preprint arXiv:2105.04544*, 2021.

[43] Wang Miao, Xu Shi, and Eric Tchetgen Tchetgen. A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945*, 2018.

[44] 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.

[45] Yu Nishiyama, Abdeslam Boularias, Arthur Gretton, and Kenji Fukumizu. Hilbert space embeddings of pomdps. *arXiv preprint arXiv:1210.4887*, 2012.

[46] Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[47] Matthew Rosencrantz, Geoff Gordon, and Sebastian Thrun. Learning low dimensional predictive representations. In *Proceedings of the twenty-first international conference on Machine learning*, page 88, 2004.

[48] Chengchun Shi, Masatoshi Uehara, and Nan Jiang. A minimax learning approach to off-policy evaluation in partially observable markov decision processes. *arXiv preprint arXiv:2111.06784*, 2021.

[49] Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436. PMLR, 2020.

[50] Satinder Singh, Michael R James, and Matthew R Rudary. Predictive state representations: a new theory for modeling dynamical systems. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 512–519, 2004.

[51] Le Song, Byron Boots, Sajid Siddiqi, Geoffrey J Gordon, and Alex Smola. Hilbert space embeddings of hidden markov models. 2010.

[52] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

[53] 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.

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

[55] Wen Sun, Arun Venkatraman, Byron Boots, and J Andrew Bagnell. Learning to filter with predictive state inference machines. In *International conference on machine learning*, pages 1197–1205. PMLR, 2016.

[56] Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

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

[58] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

[59] Yi Xiong, Ningyuan Chen, Xuefeng Gao, and Xiang Zhou. Sublinear regret for learning pomdps. *arXiv preprint arXiv:2107.03635*, 2021.

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

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

[62] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Wen Sun, and Alekh Agarwal. Efficient reinforcement learning in block mdps: A model-free representation learning approach. *arXiv preprint arXiv:2202.00063*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We explicitly list up assumptions
   (c) Did you discuss any potential negative societal impacts of your work? [No] To the best of author's knowledge, our work does not have negative social impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Due to the page limit, most of the proofs are deferred in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?[N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]