Optimistic MLE: A Generic Model-Based Algorithm for Partially Observable Sequential Decision Making[*]

Qinghua Liu
qinghual@princeton.edu
Princeton University
USA

Praneeth Netrapalli
pnetrapalli@google.com
Google Research India
India

Csaba Szepesvari
szepesva@ualberta.ca
DeepMind and University of Alberta
Canada

Chi Jin
chij@princeton.edu
Princeton University
USA

ABSTRACT

This paper introduces a simple efficient learning algorithms for general sequential decision making. The algorithm combines Optimism for exploration with Maximum Likelihood Estimation for model estimation, which is thus named OMLE. We prove that OMLE learns the near-optimal policies of an enormously rich class of sequential decision making problems in a *polynomial* number of samples. This rich class includes not only a majority of known tractable model-based Reinforcement Learning (RL) problems (such as tabular MDPs, factored MDPs, low witness rank problems, tabular weakly-revealing/observable POMDPs and multi-step decodable POMDPs), but also many new challenging RL problems especially in the partially observable setting that were not previously known to be tractable.

Notably, the new problems addressed by this paper include (1) *observable* POMDPs with continuous observation and function approximation, where we achieve the first sample complexity that is completely independent of the size of observation space; (2) *well-conditioned* low-rank sequential decision making problems (also known as Predictive State Representations (PSRs)), which include and generalize all known tractable POMDP examples under a more intrinsic representation; (3) general sequential decision making problems under *SAIL* condition, which unifies our existing understandings of model-based RL in both fully observable and partially observable settings. SAIL condition is identified by this paper, which can be viewed as a natural generalization of Bellman/witness rank to address partial observability. This paper also presents a reward-free variant of OMLE algorithm, which learns approximate dynamic models that enable the computation of near-optimal policies for all reward functions simultaneously.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

Reinforcement Learning, Optimistic MLE, PSRs, POMDPs

ACM Reference Format:
Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. 2023. Optimistic MLE: A Generic Model-Based Algorithm for Partially Observable Sequential Decision Making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC '23), June 20–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3564246.3585161

1 INTRODUCTION

A wide range of modern artificial intelligence applications can be cast as sequential decision making problems, in which an agent interacts with an unknown environment through time, and learns to make a sequence of decisions using intermediate feedback. Sequential decision making covers not only problems like Atari games [27], Go [32], Chess [6] and basic control systems [35], where states are fully accessible to the learner (the *fully observable* setting), but also applications including StarCraft [37], Poker [4], robotics with local sensors [1], autonomous driving [24] and medical diagnostic systems [14], where observations only reveal partial information about the underlying states (the *partially observable* setting). While the fully observable sequential decision making problems have been under intense theoretical investigation over recent years, the partially observable problems remain comparatively less understood.

Distinguished from fully observable systems, a learner in partially observable systems is only able to see the observations that contain partial information about the underlying states. Observations in general are no longer Markovian. As a result, it is no longer sufficient for the learner to make decision based on the observation or information available at the current step. Instead, the learner is required to additionally infer the latent states using past histories (memories). Such histories of observations have exponentially many possibilities, leading to many well-known hardness results in the worst case in both computation [28–30, 38] and statistics [22]. To avoid these worst-case barriers, a recent line of results started to investigate rich subclasses of Partially Observable Markov Decision Process (POMDPs) under the basic settings of finite states and observations [see, e.g., 18, 26], which still only constitute a relatively small subset of all partially observable problems of practical interests.

[*] The full version of this paper is available at https://arxiv.org/abs/2209.14997.

STOC '23, June 20–23, 2023, Orlando, FL, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9913-5/23/06.
https://doi.org/10.1145/3564246.3585161

In this paper, we introduce a simple, generic, model-based algorithm—OMLE, which combines Optimism (O) for exploration with Maximum Likelihood Estimation (MLE) for model estimation. We prove that OMLE learns the near-optimal policies of an enormously rich class of sequential decision making problems in a *polynomial* number of samples. This rich class includes not only a majority of known tractable model-based Reinforcement Learning (RL) problems such as tabular MDPs, factored MDPs, low witness rank problems [34], tabular weakly-revealing/observable POMDPs [18, 26] and multistep decodable POMDPs [9], but also, more importantly, many new challenging RL problems especially in the partially observable setting *that were not previously known to be tractable* (see Section 1.1). To achieve these new results, this paper develops new frameworks and techniques which address a set of fundamental challenges that are uniquely presented in the partially observable systems:

Challenge 1: Continuous observation space and function approximation with partial observability. Modern applications of sequential decision making often involve an enormous (or even infinite) number of observations, where *function approximation* must be deployed to approximate dynamic models, value functions, or policies. While function approximation greatly expands the potential reach of existing frameworks, particularly via deep architectures, it raises a number of fundamental questions including generalization, model misspecification, and how to address those issues in presence of exploration. Function approximation becomes even more complicated in the partially observable setting when further coupled with the inference of latent states and the use of history dependent policies. As a result, existing results on function approximation in the partially observable setting remain very limited [5, 36]. They make rather restrictive assumptions, and do not provide efficient guarantees even to a relatively simple continuous-observation extension of the basic tabular weakly-revealing or observable POMDPs [13, 26]—GM-POMDPs (Section 5.1.2), which only add Gaussian noise to the observations in the original models.

Challenge 2: Learning under intrinsic representation of partially observable systems. Most existing works on efficient learning of partially observable problems focus on the model of POMDPs. POMDPs are based on latent states that are *unobservable* and subject to nontrivial ambiguity—there can exist multiple different POMDPs that represent the same sequential decision making problem. This ambiguity directly leads to the unidentifiability of latent states even in the benign settings where learning near-optimal policy is possible. This paper considers a more intrinsic modeling of partially observable dynamic system—Predictive State Representations (PSRs) [25, 33], which model a dynamic system using only *observable* experiments of futures. It is known that PSRs can represent any low-rank sequential decision making problems, which are more expressive than finite-state POMDPs [15]. However, it remains unclear how to learn large class of PSRs sample-efficiently.

Challenge 3: A unified understanding of fully observable and partially observable RL.. There has been a long line of important works on generic framework of reinforcement learning [8, 10, 16, 19, 34]. However, most of them focus on the fully observable problems and are only capable of dealing with very special partially observable problems such as reactive POMDPs. A majority of them critically rely on the complexity measures that are based on Bellman rank [16] or witness rank [34] (the model-based version), which assumes the Bellman error or the model estimation error (in the model-based setting) to have a bilinear structure. These bilinear-based complexity measures completely fail to explain the tractability of many basic partially observable problems [9, 13, 26]. It remains open to develop a unified theoretical framework which explain large classes of both fully observable and partially observable problems.

This work addresses all three challenges above. For Challenge 1, we prove that OMLE learns *observable* POMDPs with continuous observation and function approximation, where we achieve the first sample complexity that is completely independent of the size of observation space. For Challenge 2, we show that OMLE learns *well-conditioned* PSRs, which include and generalize all known tractable POMDP examples under a more intrinsic representation; For Challenge 3, we identify a new condition—Summation of Absolute values of Independent biLinear functions (SAIL)—which can be viewed as a natural generalization of Bellman/witness rank to address partial observability. We prove that OMLE learns general sequential decision making problems under *SAIL* condition, which include all problems considered in this paper, and unify our existing understanding for model-based RL in both fully observable and partially observable settings.

1.1 Overview of Our Results

This paper introduces a generic algorithm framework of OMLE, and prove it learns a very rich class of sequential decision making problems sample-efficiently. The OMLE algorithm (in its basic form) was first proposed in [26] for sample-efficient learning of tabular weakly-revealing POMDPs. Here we introduce some extra flexibility to the algorithm, address new challenges, and provide learning guarantees in a significantly more general setup. Specifically,

- We identify a sufficient condition for OMLE—generalized eluder-type condition (Condition 3.1), under which OMLE is guaranteed to find near-optimal policy in a polynomial number of samples. We will use this generalized eluder-type condition to analyze all problems considered in this paper.
- We consider sequential decision making with low-rank structure (also known as Predictive State Representations (PSRs)). We first show that learning generic PSRs is intractable. We then identify a rich subclass called *well-conditioned* PSRs, and prove that OMLE learn them sample-efficiently. Our sample complexity depends polynomially on the rank of PSRs and the size of core action sequences, and is independent of the size of core tests and the size of observation space.
- We show that a wide range of POMDP models fall in to the class of *well-conditioned PSRs*. They include not only previously known tractable problems such as tabular weakly-revealing/observable POMDPs [18, 26], multistep decodable POMDPs [9]; but also new problems including observable POMDPs with continuous observation (in particular, GM-POMDPs, see Section 5.1.2), and POMDPs with a few known core action sequence. Our PSR results immediately imply sample efficient guarantees of OMLE to learn these POMDP models.

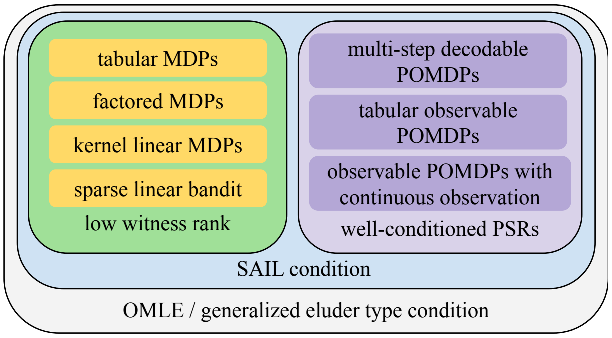
tabular MDPs
factored MDPs
kernel linear MDPs
sparse linear bandit
low witness rank
multi-step decodable POMDPs
tabular observable POMDPs
observable POMDPs with continuous observation
well-conditioned PSRs
SAIL condition
OMLE / generalized eluder type condition

Figure 1: A summary of sequential decision making problems that can be efficiently learned by OMLE.

- We identify a new SAIL condition which can be viewed as a natural generalization of Bellman/witness rank, and prove that OMLE sample-efficiently learns any sequential decision making problem with SAIL condition. We show that SAIL condition holds for well-conditioned PSRs and for all problems with low witness rank [34]. The latter covers a majority of known tractable model-based RL problems in the fully observable setting including factored MDPs, kernel linear MDPs, sparse linear bandits. Moreover, our sample complexity guarantees for learning low witness rank problems improve over the existing results [34] by a multiplicative factor of witness rank.
- We propose a variant of OMLE for reward-free learning. We show that Reward-free OMLE learns an approximate dynamic model sample-efficiently under a slightly stronger version of the SAIL condition. This approximate dynamic model allows us to compute the near-optimal policies for all reward functions simultaneously.

Besides above results, this paper also establishes the rigorous formulations for *overparameterized* PSRs, studies their properties, gives rigorous treatment for PSRs with continuous observation, and bounds the bracketing number of tabular PSRs, which might be of independent interests to the community.

1.2 Technical Contribution

Underlying our new results is a set of new techniques for handling PSRs with infinite observations.

- **New sharp elliptical potential style lemma for SAIL.** A crucial component for analyzing optimistic algorithms is pigeon-hole's principle [2, 17] or so-called elliptical potential lemma [23] which ensures that the size of confidence set is shrinking fast enough to guarantee near-optimality of the learned policy after a small number of rounds. Standard elliptical potential lemma applies to linear bandits whose reward is a linear function of form $\langle\theta, x\rangle$. To analyze POMDPs or PSRs, we establish a new generalized version of elliptical potential lemma which applies to Summation of Absolute values of Independent biLinear functions (SAIL) of form $\sum_{i=1}^{m}\sum_{j=1}^{n}|\langle\theta_i, x_j\rangle|$. A similar problem has been studied in [26] but the bounds derived therein *depend* on m, n, which scales with the size of observation space in PSRs/POMDPs. Such result becomes vacuous in the infinite-observation setting. We address this issue by developing a significantly sharper argument, which gives bounds completely *independent* of m, n (thus the size of observation space). Please see Appendix G.1 for details.
- **Projection that approximately preserve the ℓ_1-norm.** To apply the new sharp elliptical potential lemma discussed above, we need a projection operator which maps a function (or high-dimensional vector) defined on the observation space into a low-dimensional Euclidean space whose dimension is equal to the intrinsic complexity of POMDPs or PSRs. Our analysis further requires the resulting vector after projection to have a small ℓ_1-norm. In POMDPs, we can directly construct such a projection by taking the pseudo-inverse of emission matrices (as in [26]). However, such choice does not apply to PSRs as it has less structure than POMDPs. To address this issue, we consider the general problem of projecting high-dimensional vectors (that lie in a low-dimensional subspace) to a low-dimensional Euclidean space without significantly increasing their ℓ_1-norm. We achieve so by constructing a projection using the Barycentric spanner technique. Please see Lemma G.3 and Step 3 in Appendix C.4 for details.
- **Matrix pseudo-inverse with small ℓ_1-norm.** To establish efficient guarantees for learning observable POMDPs, we need to construct operator $\mathbf{M}$ as in the framework of PSR, and bound the ℓ_1-norm of the operator. All previous works [e.g., 3, 18, 26, 40, etc] construct such operators using the pseudo-inverse of emission matrices $\mathbb{O}^\dagger$, whose ℓ_1-norm scales with the size of observation space even under the *observable* condition (Condition 5.1). Such dependency prevents their analysis from generalizing to the infinite observation setting. We address this issue by adding a matrix $\mathbf{Y}$ that lies in the subspace complementary to $\mathbb{O}^\dagger$. We show that with an optimal choice of $\mathbf{Y}$, $\mathbb{O}^\dagger + \mathbf{Y}$ has a small ℓ_1-norm which is independent of the size of observation space. To our best knowledge, this operator design is completely new and has not been considered in the previous POMDP literature.

2 PRELIMINARIES

Notation. For a positive integer n, we let $[n] = \{1, \dots, n\}$. We use the notation $x_{1:n}$ to denote the sequence $(x_1, \dots, x_n)$. We use bold upper-case letters $\mathbf{B}$ to denote matrices and bold lower-case letters $\mathbf{b}$ to denote vectors. Given a matrix $\mathbf{B} \in \mathbb{R}^{m\times n}$, we use $\mathbf{B}_{ij}$ to denote its $(i, j)^{\text{th}}$ entry, $\|\mathbf{B}\|_p = \max_{\|\mathbf{z}\|\neq 0}\|\mathbf{B}\mathbf{z}\|_p/\|\mathbf{z}\|_p$ to denote its matrix p-norm, and $\mathbf{B}^\dagger$ to denote its Moore-Penrose inverse. For a vector $\mathbf{b} \in \mathbb{R}^m$, we use $\mathbf{b}_i$ to denote its i^{th} entry, $\|\mathbf{b}\|_p$ to denote its vector p-norm, and $\mathrm{diag}(\mathbf{b})$ to denote a diagonal matrix with $[\mathrm{diag}(\mathbf{b})]_{ii} = \mathbf{b}_i$. Given a set $\mathcal{X}$, we use $2^{\mathcal{X}}$ to denote the collections of all subsets of $\mathcal{X}$.

2.1 Sequential Decision Making

We consider the general episodic sequential decision making problems, which can be specified by a tuple $(\mathscr{O}, \mathscr{A}, H, \mathbb{P}, R)$. Here $\mathscr{O}$ and $\mathscr{A}$ denote the space of observation and action respectively. H denotes the length of each episode. $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^{H}$ specifies the joint distribution over observations $o_{1:H}$ conditioned on action sequence $a_{1:H}$, which can be factorized as:

$$\mathbb{P}(o_{1:H}|a_{1:H}) = \prod_{h=1}^{H} \mathbb{P}_h(o_h|o_{1:h-1}, a_{1:h-1})$$

$\mathbb{P}$ is also known as the *system dynamics*. $R = \{R_h\}_{h\in[H]}$ are the known reward functions from $\mathscr{O}$ to $[0,1]$ such that the agent will receive reward $R_h(o)$ when she observes $o \in \mathscr{O}$ at step h.[1] To simplify the presentation, we also use the notation $\overline{\mathbb{P}}(o_{1:h}, a_{1:h}) := \mathbb{P}(o_{1:h}|a_{1:h})$ for any trajectory $(o_{1:h}, a_{1:h})$ to represent the conditional probability over observations conditioned on actions. Throughout this paper we assume the finite action space $\mathscr{A}$ with $|\mathscr{A}| = A$, but allow infinitely large observation space $\mathscr{O}$.

At each step $h \in [H]$ of each episode, the environment first samples an observation o_h according to $\mathbb{P}_h(\cdot|o_{1:h-1}, a_{1:h-1})$ based on the observation-action sequence in the past, and then the agent takes an action a_h. The current episode terminates immediately after a_H is taken.

Policy and value. A policy $\pi = \{\pi_h\}_{h=1}^{H}$ is a collection of H functions where $\pi_h : (\mathscr{O} \times \mathscr{A})^{h-1} \times \mathscr{O} \to \Delta_A$ maps a length-h observation-action sequence to a distribution over actions. Given a policy π, we use V^π to denote its value, which is defined as the expected total reward received under policy π:

$$V^\pi := \mathbb{E}_\pi\left[\sum_{h=1}^{H} R_h(o_h)\right],$$

where the expectation is with respect to the randomness within the system dynamics $\mathbb{P}$ and the policy π.

Since the action space and the episode length are both finite, the maximal value over all policies $\max_\pi V^\pi$ always exists. We call $\max_\pi V^\pi$ the optimal value denoted by $V^\star$, and call the policy that achieves this optimal value the optimal policy denoted by $\pi^\star$.

Learning objective. Our goal is to learn an ε-optimal policy π in the sense that $V^\pi \geq V^\star - \varepsilon$, using a number of samples polynomial in all relevant parameters. We also consider the problem of learning with low regret. Suppose the agent interacts with the sequential decision making problem for K episodes, and plays policy π_k in the k^{th} episode for all $k \in [K]$. The total (expected) regret is then defined as:

$$\text{Regret}(K) = \sum_{k=1}^{K} [V^\star - V^{\pi_k}].$$

The question then is whether a learner can keep the regret small.

Below we describe several widely studied reinforcement learning models that can be cast into the framework of sequential interactive decision making.

Example 1 (Contextual bandit). In a contextual bandit, the observation is the context of the problem. The episode length H is equal to 1 and there exists a distribution $\mu \in \Delta_{\mathscr{O}}$ so that the first-step observation o_1 of each episode is independently sampled from μ, i.e., $\mathbb{P}(o_1 = \cdot) = \mu$.

Example 2 (MDP). In Markov decision process (MDP), the observation is the state of MDP. The observation-action pair satisfies the Markovian property. That is, there exist a collection of transition kernels $\mathbb{T} = \{\mathbb{T}_h\}_{h=1}^{H}$ so that $\mathbb{P}_h(o_h|o_{1:h-1}, a_{1:h-1}) = \mathbb{T}_{h,a_{h-1}}(o_h \mid o_{h-1})$ for all $h \in [H]$.

Example 3 (POMDP). In partially observable Markov decision process (POMDP), there is an additional latent state space $\mathscr{S}$, a collection of transition kernels $\mathbb{T} = \{\mathbb{T}_h\}_{h=1}^{H}$, an initial distribution over the latent state space μ_1, and a collection of emission kernels $\mathbb{O} = \{\mathbb{O}_h\}_{h=1}^{H}$. In a POMDP, the latent states are *hidden* from the agent. At the beginning of each episode, the environment samples an initial state s_1 from μ_1. At each step $h \in [H]$, the agent first observes o_h that is sampled from $\mathbb{O}_h(\cdot \mid s_h)$, the emission distribution of hidden state s_h at step h. Then the agent takes action a_h and receives reward $r_h(o_h)$. After this, the environment transitions to s_{h+1}, whose distribution follows $\mathbb{T}_{h,a_h}(\cdot \mid s_h)$.

We note that MDPs are fully observable models while POMDPs are partially observable models. Distinguished from MDPs where the optimal policies only depend on the current observation, the near-optimal policies of POMDPs in general depend on the entire history. This makes both learning and planning in POMDPs significantly more challenging than in MDPs.

2.2 Model-Based Function Approximation

We consider the interactive decision making problems where the observation space $\mathscr{O}$, the action space $\mathscr{A}$, the horizon H, and the reward function R are known, while the system dynamics $\mathbb{P}$ is unknown. To address infinitely large observation space, we consider the setting where we are given a model class Θ, which specifies a class of system dynamics $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$. We denote the system dynamics of the real model as $\mathbb{P}_{\theta^\star}$. Throughout this paper, we make the following realizability assumption.

Assumption 2.1 (Realizability). $\theta^\star \in \Theta$.

Realizability states that the true model resides in the given model class, so there is no misspecification error. Realizability is a standard assumption which appears in a majority of theoretical works in RL.

Following the convention in analyzing MLE [e.g., 11], we use the bracketing number to control the complexity of the model class Θ.

Definition 2.2 (Bracketing number). Given two functions l and u, the *bracket* $[l, u]$ is the set of all functions f satisfying $l \leq f \leq u$. An ε-bracket is a bracket $[l, u]$ with $\|u - l\| < \varepsilon$. The bracketing number $\mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$.

The bracketing number is required in the existing MLE analysis [11], which is in general equal or greater than the standard covering number. Across this paper, we use $\mathcal{N}_\Theta(\varepsilon)$ to denote the ε-bracketing number of function class $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ with respect to the policy-weighted ℓ_1-distance, where the policy-weighted ℓ_1-distance between two functions l and u defined on $(\mathscr{O} \times \mathscr{A})^H$ is equal to $\max_\pi \sum_{\tau_H} |l(\tau_H) - u(\tau_H)| \times \pi(\tau_H)$ [2], where the maximum is taken

[1] This is equivalent to assuming that reward information is contained in the observation. We consider this setup to avoid the leakage of information about the dynamic system through rewards beyond observations. We remark that all results in this paper immediately extend to the more general setting where reward $R(\tau_H)$ can be a function of the entire observation-action trajectory $\tau_H = (o_{1:H}, a_{1:H})$, and is only received at the end of each episode.

[2] In the settings with infinite observations, we replace the summation with integral, i.e., $\max_\pi \int_{\tau_H} |l(\tau_H) - u(\tau_H)| \times \pi(\tau_H) d\tau_H$.

Algorithm 1 Optimistic Maximum Likelihood Estimation (Θ, β)

```
1: initialize: $\mathscr{B}^1 = \Theta$, $\mathfrak{D} = \{\}$
2: for $k = 1, \dots, K$ do
3:     compute $(\theta^k, \pi^k) \leftarrow \mathrm{argmax}_{\theta \in \mathscr{B}^k, \pi} V^\pi(\theta)$
4:     compute exploration policies $\Pi^k_{\mathrm{exp}} \leftarrow \Pi_{\mathrm{exp}}(\pi^k)$
5:     for each $\pi \in \Pi^k_{\mathrm{exp}}$ do
6:         execute policy $\pi$ and collect a trajectory
               $\tau = (o_1, a_1, \dots, o_H, a_H)$
7:         add $(\pi, \tau)$ into dataset $\mathfrak{D}$
8:     update confidence set
```

$$\mathscr{B}^{k+1} = \left\{ \hat{\theta} \in \Theta : \sum_{(\pi,\tau) \in \mathfrak{D}} \log \mathbb{P}^\pi_{\hat{\theta}}(\tau) \geq \max_{\theta' \in \Theta} \sum_{(\pi,\tau) \in \mathfrak{D}} \log \mathbb{P}^\pi_{\theta'}(\tau) - \beta \right\} \bigcap \mathscr{B}^k$$

```
9: output $\pi^{\mathrm{out}}$ that is a uniform mixture of $\{\pi^k\}_{k=1}^K$
```

over all policy π. Intuitively, we need this maximization, because $\mathbb{P}_\theta$ is a conditional probability of observations given actions.

3 OPTIMISTIC MLE

In this section, we present the generic *Optimistic Maximum Likelihood Estimation* (OMLE) algorithm. Moreover, we provide a general sufficient condition—a generalized eluder-type condition (Condition 3.1), and prove that for any RL problems satisfying this condition, OMLE learns them within a polynomial number of samples.

3.1 Algorithm

The pseudocode of OMLE is provided in Algorithm 1. We remark that the OMLE algorithm was first proposed in [26] for sample-efficient learning of weakly-revealing POMDPs and here we introduce some extra flexibility in the data collection steps to handle more general learning problems.

Formally, OMLE is a model-based algorithm which takes as input a model class Θ, and executes the following three key steps in each iteration $k \in [K]$:

- **Optimistic planning** (Line 3): OMLE computes the most optimistic model θ^k in the model confidence set $\mathscr{B}^k$ and its corresponding optimal policy π^k.
- **Data collection** (Line 4-7): Based on the optimistic policy π^k, OMLE constructs a set of exploration policies $\Pi_{\mathrm{exp}}(\pi^k)$ and then the learner executes each of them to collect a trajectory. As will be explained in later sections, these exploration policies could simply be π^k or some composite policies that combine π^k with random or certain action sequences, depending on the structure of the problems to solve. Intuitively, by actively trying exploratory action sequences after π^k, the learner could gather more information about the system dynamics under π^k. As an example, when applying OMLE to learning PSRs, the exploration policies will execute the core action sequences after π^k, which we will explain in details in Section 4.
- **Confidence set update** (Line 8): Finally, OMLE updates the model confidence set using the newly collected data. Specifically, it constructs $\mathscr{B}^{k+1}$ to include all the models $\theta \in \Theta$ whose log likelihood on all the historical data collected so far is close to the maximal log likelihood up to an additive factor β. This can be viewed as a relaxation of the classic maximal likelihood estimation (MLE) approach which chooses the model estimate to be the one exactly maximizing the log likelihood. In particular, when $\beta = 0$, $\mathscr{B}^{k+1}$ reduces to the solution set of MLE. One important reason behind this construction is that by choosing the relaxation parameter β properly, we can guarantee the true model $\theta^\star$ lies in the confidence set for all $k \in [K]$ with high probability, under the realizability assumption.

3.2 Theoretical Guarantees

In this section, we present the theoretical guarantees for OMLE. To present our results in the most general form, we first introduce a sufficient condition, called *generalized eluder-type condition*. We then provide the sample-efficiency guarantees for OMLE in learning any RL problems that satisfy this condition. Let $\mathbb{P}^\pi_\theta$ denote the distribution over $(o, a, r)_{1:H}$ induced by executing policy π in model θ.

Condition 3.1 (Generalized eluder-type condition)**.** There exists a real number $d_\Theta \in \mathbb{R}_+$ and a function ξ such that: for any $(K, \Delta) \in \mathbb{N} \times \mathbb{R}^+$, and for the models $\{\theta^k\}_{k \in [K]}$ and the policies $\{\pi^k\}_{k \in [K]}$, $\{\Pi^k_{\mathrm{exp}}\}_{k \in [K]}$ in Algorithm 1, we have

$$\begin{aligned} &\forall k \in [K], \sum_{t=1}^{k-1} \sum_{\pi \in \Pi^t_{\mathrm{exp}}} d^2_{\mathrm{TV}}(\mathbb{P}^\pi_{\theta^k}, \mathbb{P}^\pi_{\theta^\star}) \leq \Delta \\ &\Rightarrow \sum_{k=1}^{K} d_{\mathrm{TV}}(\mathbb{P}^{\pi^k}_{\theta^k}, \mathbb{P}^{\pi^k}_{\theta^\star}) \leq \xi(d_\Theta, K, \Delta, |\Pi_{\mathrm{exp}}|) \end{aligned} \tag{1}$$

where $|\Pi_{\mathrm{exp}}| := \max_\pi |\Pi_{\mathrm{exp}}(\pi)|$ is the largest possible number of exploration policies in each iteration.

At a high level, Condition 3.1 resembles the pigeonhole principle and the elliptical potential lemma widely used in tabular MDPs [e.g., 2, 17] and linear bandits/MDPs [e.g., 20, 23] respectively. Such type of condition is widely used as a sufficient condition for algorithms using optimistic exploration [31]. Importantly, *we will prove that Condition 3.1 holds for all the problems studied in this paper*, with moderate d_Θ and function ξ whose leading term scales as $\tilde{O}(\sqrt{d_\Theta \Delta |\Pi_{\mathrm{exp}}| K})$.

For an intuitive understanding of this generalized eluder-type condition, imagine that in each k^{th} iteration, the learner chooses a model θ^k such that θ^k can accurately predict the behavior of the historical exploration policies in $\Pi^1_{\mathrm{exp}}, \dots, \Pi^{k-1}_{\mathrm{exp}}$ up to cumulative error Δ (i.e., the left inequality of (1)). Since θ^k could be different from $\theta^\star$, the learner will still suffer an instantaneous error in predicting the behavior of policy π_k using model θ_k. And $\xi(d_\Theta, K, \Delta, |\Pi_{\mathrm{exp}}|)$ essentially measures the worst-case growth rate of the cumulative instantaneous error with respect to K.

The key motivation behind Condition 3.1 is that because of the way OMLE constructs the confidence set $\mathcal{B}^k$, we can use the classical analysis of MLE [11] to guarantee that any model inside $\mathcal{B}^k$ is close to the true model $\theta^\star$ in TV-distance under the historical policies in $\Pi^1_{\rm exp},\ldots,\Pi^{k-1}_{\rm exp}$ with high probability. As a result, if the problem further satisfies the generalized eluder-type condition, then OMLE immediately enjoys low-suboptimality guarantee by the optimism of $\{\pi^k\}_{k=1}^K$ and Condition 3.1. Formally, we have the following theoretical guarantee for OMLE.

Theorem 3.2. *There exists absolute constant* $c_1, c_2 > 0$ *such that for any* $\delta\in(0,1]$ *and* $K\in\mathbb{N}$, *if we choose* $\beta = c_1\log(T\mathcal{N}_\Theta(T^{-1})/\delta)$ *with* $T = K|\Pi_{\rm exp}|$ *in OMLE (Algorithm 1) and assume Condition 3.1 holds, then with probability at least* $1-\delta$, *we have* $\sum_{k=1}^K[V^\star - V^{\pi_k}]\le H\xi(d_\Theta, K, c_2\beta, |\Pi_{\rm exp}|)$.

As mentioned before, for all problems studied in this paper, the leading term (in terms of K dependency) of function ξ scales as $\tilde{O}(\sqrt{d_\Theta\beta|\Pi_{\rm exp}|K})$. Then, Theorem 3.2 immediately leads to a guarantee $\sum_{k=1}^K[V^\star - V^{\pi_k}]\le \tilde{O}(H\sqrt{d_\Theta\beta|\Pi_{\rm exp}|K}) + o(\sqrt{K})$, which gives the optimal $\sqrt{K}$ dependency up to a polylogarithmic factor. We remark that Theorem 3.2 is not a regret guarantee unless $\Pi^k_{\rm exp} = \{\pi^k\}$, because it is the policies in $\{\Pi^k_{\rm exp}\}_{k=1}^K$ that are executed by OMLE, not $\{\pi^k\}_{k=1}^K$.

Sample complexity. Since the output policy $\pi^{\rm out}$ is a uniform mixture of $\{\pi^k\}_{k=1}^K$, we have $V^{\pi^{\rm out}} = (\sum_{k=1}^K V^{\pi_k})/K$. As a result, Theorem 3.2 immediately implies that with probability at least $1-\delta$, $\pi^{\rm out}$ of OMLE is ε-optimal as long as $H\xi(d_\Theta, K, \beta)/K\le\varepsilon/2$. In particular, when $\xi(d_\Theta, K,\beta)$ scales as $\tilde{O}(\sqrt{K})$ with respect to K, it suffices to run OMLE for $K\ge\tilde{O}(\varepsilon^{-2})$ episodes, where the dependency on ε is again optimal up to a polylogarithmic factor.

4 LOW-RANK SEQUENTIAL DECISION MAKING

In this section, we consider an important large class of sequential decision making problems which has a low-rank structure. Note that the entire dynamics of the sequential decision making problem is fully specified by the joint probability $\mathbb{P}(o_{1:H}|a_{1:H})$. We can equivalently view this joint probability as system-dynamic matrices $\{\mathbb{D}_h\}_{h\in[H]}$: for each fixed step h, we call an observation-action sequence in previous steps up to h, i.e., $\tau_h = (o_{1:h}, a_{1:h})$ a **history**, and call an observation-action sequence in future steps, i.e., $\omega_h = (o_{h+1:m}, a_{h+1:m})$ for any $m\in[h+1,H]$ a **future** (or test). Denote the set of all possible histories at step h as $\mathcal{T}_h$ and the set of all possible futures as Ω_h. Then we can define the system-dynamic matrix $\mathbb{D}_h\in\mathbb{R}^{|\mathcal{T}_h|\times|\Omega_h|}$ as a matrix with histories as rows and futures as columns[3] whose entry is specified as

$$[\mathbb{D}_h]_{\tau_h,\omega_h} = \overline{\mathbb{P}}(\tau_h,\omega_h) := \mathbb{P}(o_{1:H}|a_{1:H}) \tag{2}$$

The **rank** of the sequential decision making problem is simply defined as $\max_{h\in[H]}\operatorname{rank}(\mathbb{D}_h)$, which is the maximal rank of the system-dynamic matrices $\{\mathbb{D}_h\}_{h\in[H]}$.

4.1 Predicative State Representations

Predicative State Representations (PSRs) are proposed by [25, 33] as a generic approach to model low-rank sequential decision making problems. Consider a fixed step $h\in[H-1]$, and denote $r = \operatorname{rank}(\mathbb{D}_h)$. For any integer $d\ge r$, there always exist d columns (denoted as Q_h) of matrix $\mathbb{D}_h$, such that the submatrix restricted to these columns $\mathbb{D}_h[Q_h]$ satisfies $\operatorname{rank}(\mathbb{D}_h[Q_h]) = r$. These d columns correspond to d futures $Q_h = \{q_1,\ldots,q_d\}$, which are called *core tests*. Throughout this section, we assume all models in our model class Θ share the same sets of core tests, which are known to the learner. While most literature in PSRs often choose $d = r$, in many applications (as shown in the next section), learner only knows a set of core tests with a larger size. Therefore, we also consider the setting when $d > r$, to which we refer as *overparameterized* PSR.

Core tests allow the system-dynamic matrix $\mathbb{D}_h$ to be factorized as follows for certain matrix $\mathbf{W}_h$:

$$\mathbb{D}_h = \mathbb{D}_h[Q_h]\cdot\mathbf{W}_h^\top,\quad \mathbb{D}_h[Q_h]\in\mathbb{R}^{|\mathcal{T}_h|\times d}, \mathbf{W}_h\in\mathbb{R}^{|\Omega_h|\times d} \tag{3}$$

This implies an important property: for any history τ_h, the $\tau_h^{\rm th}$ row of $\mathbb{D}_h[Q_h]$, which we denote as $\boldsymbol{\psi}(\tau_h) := (\overline{\mathbb{P}}(\tau_h, q_1),\ldots,\overline{\mathbb{P}}(\tau_h, q_d))$, serves as a sufficient statistics for the history τ_h in predicting the the probabilities of all futures conditioned on τ_h. In sum, PSR captures the state of a dynamic system using $\boldsymbol{\psi}(\tau_h)$—a vector of predictions for future tests.

Formally, PSR models the dynamic system using a tuple $(\boldsymbol{\phi}, \mathbf{M}, \boldsymbol{\psi}_0)$, where $\boldsymbol{\phi} = \{\boldsymbol{\phi}_H(o,a)\}_{(o,a)\in\mathcal{O}\times\mathcal{A}}$ is a set of vectors where $\boldsymbol{\phi}_H(o,a)\in\mathbb{R}^{|Q_{H-1}|}$; $\mathbf{M} = \{\mathbf{M}_h(o,a)\}_{(h,o,a)\in[H-1]\times\mathcal{O}\times\mathcal{A}}$ is a set of matrices where $\mathbf{M}_h(o,a)\in\mathbb{R}^{|Q_h|\times|Q_{h-1}|}$, and $\boldsymbol{\psi}_0$ is a vector in $\mathbb{R}^{|Q_0|}$. The tuple satisfies following two equations:

$$\mathbb{P}(o_{1:H}|a_{1:H}) = \boldsymbol{\phi}_H(o_H, a_H)^\top\mathbf{M}_{H-1}(o_{H-1}, a_{H-1})\cdots\mathbf{M}_1(o_1,a_1)\boldsymbol{\psi}_0, \tag{4}$$

$$\boldsymbol{\psi}(o_{1:h}, a_{1:h}) = \mathbf{M}_h(o_h,a_h)\cdots\mathbf{M}_1(o_1,a_1)\boldsymbol{\psi}_0, \tag{5}$$

for any $h\in[0,H-1]$ and any observation-action sequence $(o_{1:h}, a_{1:h})$. That is, in PSR, the joint probability $\mathbb{P}(o_{1:H}|a_{1:H})$ can be factorized as a product of matrices and vectors where each matrix only depends on the observation and action at the corresponding step. The second condition (5) further requires the product of the first h matrices to have a probabilistic interpretation—the sufficient statistics $\boldsymbol{\psi}(o_{1:h}, a_{1:h})$ for the history $(o_{1:h}, a_{1:h})$. In condition (5), we include the special case $h = 0$, where the history τ_h is empty $\emptyset$, and the condition becomes $\boldsymbol{\psi}(\emptyset) := (\overline{\mathbb{P}}(q_1),\ldots,\overline{\mathbb{P}}(q_d)) = \boldsymbol{\psi}_0$ for core tests $\{q_1,\ldots,q_d\}$ in Q_0. We call the sets of core tests $\{Q_h\}_{h\in[H-1]}$ along with the tuple $(\boldsymbol{\phi},\mathbf{M},\boldsymbol{\psi}_0)$ the PSR representation of the dynamic system. Finally, we define the rank of a PSR to be the rank of the underlying sequential decision making problem that the PSR describes (according to (4)).

Representation power of PSRs. The following theorem [see e.g., 25, 33] guarantees the existence of such PSR representation $(\boldsymbol{\phi},\mathbf{M},\boldsymbol{\psi}_0)$ for any low-rank sequential decision making problem.

Theorem 4.1. *Any rank-r sequential decision making problem can be represented by a PSR with sets of core tests whose sizes are no larger than r. That is, there always exist sets of core tests* $\{Q_h\}_{h\in[H-1]}$

[3] For clean presentation, here we write $\mathbb{D}_h$ as a matrix, which requires $|\Omega_h|$ or $|\mathcal{O}|$ to be finite. We remark that our framework immediately extends to the infinite observation setting. See Appendix A for more details.

with size $\max_{h\in[H-1]}|Q_h| \le r$*, and a corresponding tuple* $(\boldsymbol{\phi}, \mathbf{M}, \boldsymbol{\psi}_0)$ *which jointly satisfy Equation* (4) (5).

Theorem 4.1 demonstrates the superior expressive power of PSR, in the sense that any low-rank sequential decision making problem admits an equivalent and compact PSR representation. This is in sharp contrast to other models of dynamical systems such as POMDPs which not only implicitly require the system dynamics being low-rank but also explicitly assume the existence of latent nominal states so that the current state of the system can be represented as a probability distribution over these unobservable nominal states. As a result, PSRs can model strictly more complex dynamical systems than POMDPs with finite states, e.g., the probability clock introduced in [15].

Linear weight vectors. According to low rank factorization (3), we know there exist linear weight vectors $\{\mathbf{m}(\omega_h)\}_{\omega_h\in\Omega_h}$ only depending on the futures (where $\mathbf{m}(\omega_h)$ can be the ω_h^{th} row of $\mathbf{W}_h$ matrix) such that for any future ω_h and history τ_h, the joint probability can be written in the bilinear form

$$\overline{\mathbb{P}}(\tau_h, \omega_h) = \mathbf{m}(\omega_h)^\top \boldsymbol{\psi}(\tau_h). \tag{6}$$

Equation (4) and (5) give two natural constructions for weight vectors. First, consider futures of full length $\Omega_h^{(1)} := (\mathscr{O}\times\mathscr{A})^{H-h}$. Equation (4) gives the weight vector of any future $\omega_h = (o_{h+1:H}, a_{h+1:H}) \in \Omega_h^{(1)}$ as:

$$\mathbf{m}_1(\omega_h)^\top = \boldsymbol{\phi}_H(o_H, a_H)^\top \mathbf{M}_{H-1}(o_{H-1}, a_{H-1}) \cdots \mathbf{M}_{h+1}(o_{h+1}, a_{h+1}) \tag{7}$$

Next, consider the future set of $\Omega_h^{(2)} := \mathscr{O}\times\mathscr{A}\times Q_{h+1}$. Equation (5) gives the weight vector of any future $\omega_h = (o_{h+1}, a_{h+1}, q_i) \in \Omega_h^{(2)}$ (where $q_i \in Q_{h+1}$ is the i^{th} core test of Q_{h+1})as:

$$\mathbf{m}_2(\omega_h)^\top = \boldsymbol{e}_i^\top \mathbf{M}_{h+1}(o_{h+1}, a_{h+1}) \tag{8}$$

We note that in the overparameterized setting ($|Q_h| > \text{rank}(\mathbb{D}_h)$), the choice of linear weights $\mathbf{m}(\cdot)$ in (6) may not be unique. As a result, the constructions in (7) and (8) are not necessarily related in general, unless a further *self-consistent* condition is satisfied (see discussion in Appendix C.3 for more details).

Core action sequences. We note that multiple core tests might use the same action sequence $a_{h+1:m}$ for $m \in [h+1, H]$. Therefore, in many occasions, it is convenient to consider the set of core action sequences Q_h^{A}, which is the set of unique actions sequences within the set of core tests Q_h. We know immediately that $|Q_h^{\text{A}}| \le |Q_h|$ and any rank-r system-dynamic matrix $\mathbb{D}_h$ admits at least one set of core action sequences with size $|Q_h^{\text{A}}| \le r$. The size of core action sequences $|Q_h^{\text{A}}|$ determines the number of experiments we need to conduct in the dynamic system in order to estimate $\boldsymbol{\psi}(\tau_h)$. As we will see later, all our sample complexity results only depend on $|Q_h^{\text{A}}|$ instead of $|Q_h|$. WLOG, we assume that no core action sequence is a prefix of another core action sequence.

Continuous observation. For clean presentation, we write the results in this section using the formulation with finite observations. As we will see, our sample complexity results are completely independent of the number of observations, which allows our results to readily extend to the setting of continuous observation. For rigorous treatment, we note that in our current definition each core test is a single observation-action sequence, which has probability 0 to be observed if the observation is continuous. In Appendix A, we provide two approaches to modify the PSR formulation to resolve this issue. One approach is to consider a dense set of core tests with infinitely many futures, and generalize $\boldsymbol{\psi}(\tau)$ and $\mathbf{M}(o, a)$ from vectors and matrices to functions and linear operators in Hilbert space. Our results remain meaningful even with infinitely many core tests as long as the number of core action sequences is small. The second approach is to generalize the definition of core test to be an event of whether the future lands in a measurable subset of future space. We defer the details of rigorous treatment of continuous observation to Appendix A.

4.2 Well-Conditioned PSRs

Since PSR includes POMDP as a special case, it naturally inherits all the hardness results of learning POMDPs. In particular, even when the observation space, the action space and the sets of core tests are all small, finding a near-optimal policy still requires an exponential number of samples in the worst case.

Proposition 4.2. *There exists a family of PSRs with* $|\mathscr{O}|, |\mathscr{A}|, \max_h |Q_h| = O(1)$ *so that any algorithm requires at least* $\Omega(2^H)$ *samples to learn a* $(1/4)$*-optimal policy with probability* $1/6$ *or higher.*

The proof of Proposition 4.2 essentially follows from Theorem 6 in [26] which shows the hardness for learning POMDPs when the weakly-revealing coefficient is bad. See Appendix C.2 for details.

Intuitively, the hard instances in Proposition 4.2 is due to the following reason: in the definition of PSR, we require that for each step h, the core tests Q_h satisfies $\text{rank}(\mathbb{D}_h[Q_h]) = \text{rank}(\mathbb{D}_h) := r$. However, this requirement alone does not prohibit the submatrix $\mathbb{D}_h[Q_h]$ to be extremely close to some matrix whose rank is strictly less than r. That is, matrix $\mathbb{D}_h[Q_h]$ can be highly ill-conditioned. This will lead to high non-robustness in predicting the probability of $\overline{\mathbb{P}}(\tau_h, \omega_h) = \mathbf{m}(\omega_h)^\top\boldsymbol{\psi}(\tau_h)$ when the vector $\boldsymbol{\psi}(\tau_h)$ needs to be estimated—the corresponding linear weight $\mathbf{m}(\omega_h)$ can be extremely large such that we need to estimate $\boldsymbol{\psi}(\tau_h)$ up to an extremely high accuracy. Indeed, in the hard instances of Proposition 4.2, there exists some future ω_h such that $\|\mathbf{m}(\omega_h)\|_1 \ge \Omega(2^H)$.

To rule out such hard instances, core tests are required to not only guarantee $\text{rank}(\mathbb{D}_h[Q_h]) := r$, but also ensure $\mathbb{D}_h[Q_h]$ to be "well-conditioned" in certain sense. In this paper, we enforce such condition by assuming an upper bound on the magnitude of linear weight vectors.

Condition 4.3 (γ-well-conditioned PSR)**.** We say a PSR is γ*-well-conditioned* if for any $h \in [H-1]$ and any policy π independent of the history before step $h+1$, the weight vectors $\mathbf{m}_1(\cdot), \mathbf{m}_2(\cdot)$ and the corresponding future sets $\Omega_h^{(1)}, \Omega_h^{(2)}$ in (7) (8) satisfy:

$$\max_{i\in\{1,2\}} \max_{\substack{\mathbf{x}\in\mathbb{R}^{|Q_h|}\\ \|\mathbf{x}\|_1\le 1}} \sum_{\omega_h\in\Omega_h^{(i)}} \pi(\omega_h)\cdot|\mathbf{m}_i(\omega_h)^\top\mathbf{x}| \le \frac{1}{\gamma}. \tag{9}$$

Remark 4.4. *Condition 4.3 requires* $\max_{i\in\{1,2\}}$ *because for overparameterized PSR, linear weight vectors are not unique. Thus,* $\mathbf{m}_1, \mathbf{m}_2$ *in general are not related. However, if they are related by self-consistency* *(see Appendix C.3 for details), then it is sufficient to assume the inequality* (9) *only for* $\mathbf{m}_1, \Omega_h^{(1)}$.

Intuitively, the parameter γ^{-1} above measures how much the future weight vectors $\{\mathbf{m}(\omega_h)\}_{\omega_h \in \Omega_h}$ can amplify the error $\mathbf{x}$ arising from estimating the probability of core tests, in an averaged sense that the future ω_h is sampled from policy π. Being γ-well-conditioned naturally requires this error amplification to be not extremely large since otherwise the hard instances mentioned before will come into play. In Section 5, we will prove many common partially observable RL problems are naturally γ-well-conditioned PSRs with moderate γ, e.g., observable POMDPs and multistep decodable POMDPs.

4.3 Theoretical Results

In this subsection, we present the theoretical guarantees for learning well-conditioned PSRs with OMLE. To analyze OMLE, we first need to specify the exploration policy function $\Pi_{\exp}$. Denote by $\nu(\pi, h, \mathbf{a})$ a composite policy that first executes policy π for step 1 to step $h-1$, then takes random action at step h, and after that executes action sequence $\mathbf{a} = (a_{h+1}, \ldots, a_m)$ till certain step m, and finally finishes the remaining steps of the current episode by taking random actions. We construct the following exploration policy function:

$$\Pi_{\exp}(\pi) := \bigcup_{h \in [H-1]} \{\nu(\pi, h, \mathbf{a}) : \mathbf{a} \in \mathcal{Q}_h^{\mathrm{A}}\}. \tag{10}$$

By using the above exploration policy function in OMLE, we have the following polynomial sample-efficiency guarantee for learning well-conditioned PSRs.

Theorem 4.5. *Let $c > 0$ be an absolute constant large enough and Θ be a rank-r γ-well-conditioned PSR class. For any $\delta \in (0, 1]$ and $K \in \mathbb{N}$, if we choose $\beta = c \log(T\mathcal{N}_\Theta(T^{-1})\delta^{-1})$ with $T = KH \max_h |\mathcal{Q}_h^A|$ and $\Pi_{\exp}$ specified by Equation* (10) *in OMLE (Algorithm 1), then with probability at least $1-\delta$, we have*

$$V^\star - V^{\pi^{\mathrm{out}}} \le \mathrm{poly}(r, \gamma^{-1}, \max_h |\mathcal{Q}_h^A|, A, H, \log K) \times \sqrt{\frac{\beta}{K}}\,.$$

The result in Theorem 4.5 scales polynomially with respect to the rank of the PSR r, the inverse well-conditioned parameter γ^{-1}, the number of core action sequences $\max_h |\mathcal{Q}_h^{\mathrm{A}}|$, the log-bracketing number of the model class $\log \mathcal{N}_\Theta$, the number of actions A, and the episode length H. In particular, (1) it does *not* depend on the size of core tests, but instead only depend on the size of core action sequence; (2) it is completely independent of the size of the observation space. Both empower our results to handle problems with continuous observations. Moreover, when the bracketing number satisfies $\log \mathcal{N}_\Theta(T^{-1}) \le O(\mathrm{polylog}(T))$ (e.g., in tabular PSRs and POMDPs with mixture of Gaussian observations), Theorem 4.5 guarantees that $K = \tilde{O}(\varepsilon^{-2})$ episodes suffices for finding an ε-optimal policy, which is optimal up to a polylogarithmic factor.

The proof of Theorem 4.5 relies on the following key lemma, which states that any class of well-conditioned PSRs satisfy the generalized eluder-type condition (Condition 3.1) with favorable d_Θ and ζ.

Lemma 4.6. *Let Θ be a family of rank-r γ-well-conditioned PSRs. Then Condition 3.1 holds with $\Pi_{\exp}$ defined in Equation* (10)*,*

$$d_\Theta = (r\gamma^{-2}A^2 \max_h |\mathcal{Q}_h^A|)^2 \mathrm{poly}(H),$$

and

$$\xi(d_\Theta, \Delta, |\Pi_{\exp}|, K) = \tilde{O}(\sqrt{d_\Theta \Delta |\Pi_{\exp}| K}).$$

Once Lemma 4.6 is established, Theorem 4.5 follows immediately from combining it with the guarantee of OMLE (Theorem 3.2).

Technical challenge. One of the key steps in proving Lemma 4.6 is to establish a generalized version of elliptical potential lemma for Summation of Absolute values of Independent biLinear (SAIL) functions of form $\sum_{i=1}^m \sum_{j=1}^n |\langle \theta_i, x_j \rangle|$. Despite similar problems have been investigated in the previous analysis of OMLE [26], the bound derived therein scales with m, n, which depend on the number of observations. As a result, that bound is incapable to handle the settings with infinite observations. To address this issue, we develop a much tighter elliptical potential lemma which completely get rids of the m, n dependence. With the help of this strengthened elliptical potential lemma and other newly developed techniques, we are able to prove Lemma 4.6 without suffering any dependence on the size of the observation space. We refer an interesting reader to Appendix G.1 for more technical details.

4.3.1 *Special cases: tabular PSRs.*

To apply Theorem 4.5, we still need to upper bound the bracketing number of model class Θ. The following proposition states that in tabular PSRs (i.e., PSRs with finite observations and actions) the log-bracketing number of Θ is always upper bounded.

Theorem 4.7 (bracketing number of tabular PSRs). *Let Θ be the collections of all rank-r PSRs with O observations, A actions and episode length H. Then $\log \mathcal{N}_\Theta(\varepsilon) \le O(r^2 OAH^2 \log(rOAH/\varepsilon))$.*

We remark that the bracketing number in Theorem 4.7 is independent of the size of core tests or core action sequences. This is because the representation power of rank-r PSRs is limited to rank-r sequential decision making problems regardless the choices of core tests.

The key intermediate step in proving Theorem 4.7 is to show every low rank sequential decision making problem admits an observable operator model (OOM) representation wherein the norm of the operators are well controlled. Once this argument is established, we can upper bound the bracketing number by discretizing those operators. In comparison, recent works on PSRs [40] simply *assume* every PSR representation has bounded operator norm without proving it. To our knowledge, Theorem 4.7 provides the first polynomial upper bound for the bracketing number of tabular PSRs *without any additional assumptions.*

Finally, by plugging the above upper bound back into Theorem 4.5, we immediately obtain the following sample complexity bound for learning tabular PSRs:

$$\mathrm{poly}(r, \gamma^{-1}, \max_h |\mathcal{Q}_h^{\mathrm{A}}|, O, A, H, \log(\varepsilon^{-1}\delta^{-1})) \cdot \varepsilon^{-2}.$$

5 IMPORTANT PSR SUBCLASSES

In this section, we introduce several partially observable RL problems of interests and prove that they are all special subclasses of

γ-well-conditioned PSRs with moderate γ. All the proofs for this section are deferred to Appendix D.

5.1 Observable POMDPs

We first consider observable POMDPs [13] [4] —an important, natural and rich subclass of POMDPs wherein there exists an integer $m \in [H]$ so that any two different distributions over latent states induce different m-step observation-action distributions. We will prove a new result that OMLE can sample-efficiently learn any observable POMDP even with *infinite* or *continuous* observation. We remark that while such a result have been proved in the setting of finite observations [26], the sample complexity in [26] has a polynomial dependency on the number of observations, thus does not extend to the setting of continuous observation. Our new result is highly non-trivial: in addition to the sample-complexity guarantees of well-conditioned PSRs with continuous observation (Theorem 4.5), our result further requires new techniques on matrix pseudo-inverse with small ℓ_1-norm (Appendix G.3) and a new core tests design technique (Appendix D.2),

To formally state the observability condition, we first define the m-step observation-action probability kernels as follows:

$$\{\mathbb{G}_h \in \{\mathscr{O}^m \times \mathscr{A}^{m-1} \to \mathbb{R}\}^S\}_{h\in[H-m+1]}$$

For an observation sequence $\mathbf{o}$ of length m, a latent state s and an action sequence $\mathbf{a}$ of length $m-1$, the value of the s^{th} probability function in $\mathbb{G}_h$ at point $(\mathbf{o},\mathbf{a}) \in \mathscr{O}^m \times \mathscr{A}^{m-1}$, denoted as $\mathbb{G}_{h,s}(\mathbf{o},\mathbf{a})$, is equal to the probability density of observing $\mathbf{o}$ provided that the action sequence $\mathbf{a}$ is used from state s and step h:

$$\mathbb{G}_{h,s}(\mathbf{o},\mathbf{a}) := \mathbb{P}(o_{h:h+m-1} = \mathbf{o} \mid s_h = s, a_{h:h+m-2} = \mathbf{a}). \tag{11}$$

And we say a POMDP is m-step α-observable ($m \in [H]$ and $\alpha > 0$), if its m-step observation-action probability kernels satisfy the following condition.

Condition 5.1 (m-step α-observable condition)**.** For any $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \Delta_{\mathscr{S}}$ and $h \in [H-m+1]$,

$$\|\mathbb{E}_{s\sim\boldsymbol{\nu}_1}[\mathbb{G}_{h,s}] - \mathbb{E}_{s\sim\boldsymbol{\nu}_2}[\mathbb{G}_{h,s}]\|_1 \geq \alpha\|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2\|_1. \tag{12}$$

In the above condition, we use $\|f-g\|_1 = \int_{x\in\mathcal{X}} |f(x)-g(x)|dx$ to denote the ℓ_1-distance between two functions from $\mathcal{X}$ to $\mathbb{R}$. Intuitively, Condition 5.1 can be viewed as a robust version of assuming that the S probability functions in each $\mathbb{G}_h$ are linearly independent, which guarantees that for any two different latent state mixtures $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \Delta_S$, there exists an action sequence $\mathbf{a}$ of length $m-1$ so that these two mixtures can be distinguished from the distributions over the next m-step observations provided that action sequence $\mathbf{a}$ is executed.

The following theorem states that any m-step α-observable POMDP admits an $\alpha/(S+A^{m-1})$-well-conditioned PSR representation with core action sets equal to $\mathscr{A}^{m-1}$.

Theorem 5.2. *Let Θ be a model class of m-step α-observable POMDPs. Then Θ satisfies Condition 4.3 with $\gamma = O(\alpha/S)$ and $Q_h^A = \mathscr{A}^{\min\{m-1,H-h\}}$.*

New PSR operators for observable POMDPs. The key challenge in proving Theorem 5.2 is to construct a set of PSR operators that satisfy Condition 4.3 with parameter γ independent of the number of observations O. For simplicity of illustration, let us consider 1-step α-observable tabular POMDPs as examples in this paragraph. Previous work [26] and concurrent works [7, 40] all adopt the following operator construction:

$$\mathbf{M}_h(o,a) = \mathbb{O}_{h+1}\mathbb{T}_{h,a}\text{diag}(\mathbb{O}_h(o \mid \cdot))\mathbb{O}_h^{\dagger} \in \mathbb{R}^{O\times O},$$

where $\mathbb{T}_{h,a} \in \mathbb{R}^{S\times S}$ is the transition matrix of action a, $\mathbb{O}_h \in \mathbb{R}^{O\times S}$ is the observation matrix and $\mathbb{O}_h(o \mid \cdot)$ is the o^{th} row of $\mathbb{O}_h$, all for step h. However, the above operators have γ scaling as $O(\alpha/\sqrt{O})$ in the worst case, which hinders generalization to the infinite-observation settings. To address this issue, we propose a different operator construction based on a novel ℓ_1-norm matrix inverse technique (Lemma G.4):

$$\mathbf{M}_h(o,a) = \mathbb{O}_{h+1}\mathbb{T}_{h,a}\text{diag}(\mathbb{O}_h(o \mid \cdot))(\mathbf{Y}_h + \mathbb{O}_h^{\dagger}) \in \mathbb{R}^{O\times O}$$
$$\text{where } \mathbf{Y}_h \in \operatorname*{argmin}_{\tilde{\mathbf{Y}}\in\mathbb{R}^{S\times O}} \|\tilde{\mathbf{Y}} + \mathbb{O}_h^{\dagger}\|_1,$$

which, importantly, satisfies Condition 4.3 with $\gamma = O(\alpha/S)$ completely independent of O. When moving from the single-step observable tabular setting to the more challenging multi-step observable infinite-observation setting, the same idea still plays an important role in constructing well-conditioned PSR operators, where we first use a novel partition technique to group different observations to obtain an $(\alpha/2)$-observable meta-POMDP with finite but exponentially many meta-observations and then apply the above operator construction on top of the meta-POMDP. For more technical details, please refer to Appendix D.1 and D.2.

Sample complexity. By combining Theorem 5.2 with Theorem 4.5, we immediately obtain the following sample-efficiency guarantee for learning observable POMDPs with OMLE.

Corollary 5.3. *Let Θ be a model class of m-step α-observable POMDPs. There exists an absolute constant $c > 0$ such that for any $\delta \in (0,1]$ and $K \in \mathbb{N}$, if we choose $\beta = c\log(T\mathcal{N}_\Theta(T^{-1})\delta^{-1})$ with $T = KHA^m$ in OMLE (Algorithm 1), then with probability at least $1-\delta$,*

$$V^\star - V^{\pi^{\text{out}}} \leq \text{poly}(\alpha^{-1}, S, A^m, H, \log K) \times \sqrt{\frac{\beta}{K}}.$$

Different from previous works on tabular POMDPs [12, 20, 26] where the sample complexity scales with the number of observations, The result in Corollary 5.3 completely gets rid of the dependence on O thanks to our novel PSR operator design as is discussed above. As a result, it also applies to learning observable POMDPs with continuous observations as long as the log-bracketing number of model class Θ is well controlled, whereas previous works cannot.

5.1.1 Observable tabular POMDPs.

We first consider tabular observable POMDPs where the number of observations is finite. In this case, the m-step observation-action probability kernel $\mathbb{G}_h$ is equivalent to an O^mA^{m-1} by S matrix wherein the entry at the

[4] [26] considers a similar subclass called *weakly-revealing* POMDPs, which assumes the S^{th} singular value of matrix $\mathbb{G}_h$ to be lower bounded. Here $\mathbb{G}_h$ is a matrix of size $O^mA^{m-1} \times S$ whose entry is defined in (11). [26] proved that observable POMDPs and weakly-revealing POMDPs are equivalent up to a polynomial factor that depends on the number of states and observations. Therefore, there is essentially no difference in proving polynomial sample complexity for two classes in the tabular setting. However, observable POMDPs extend more naturally to the setting of continuous observation, and natural examples in continuous observation such as GM-POMDPs in Section 5.1.2 do not satisfy weakly-revealing condition.

intersection of the $(\mathbf{o},\mathbf{a})^{\text{th}}$ row and the s^{th} column is equal to $\mathbb{P}(o_{h:h+m-1}=\mathbf{o} \mid s_h = s, a_{h:h+m-2}=\mathbf{a})$. And the observable condition (Condition 5.1) can be equivalently written as:

$$\max_{h\in[H-m+1]} \|\mathbb{G}_h(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)\|_1 \ge \alpha\|\boldsymbol{\nu}_1-\boldsymbol{\nu}_2\|_1. \tag{13}$$

To apply OMLE to tabular POMDPs with S states, O observations and A actions, we choose the model class Θ to consist of all the legitimate POMDP parameterizations $\theta = (\mathbb{T},\mathbb{O},\boldsymbol{\mu}_1)$ whose corresponding m-step observation-action probability matrices satisfy Equation (13). By simple discretization argument [e.g., see Appendix B in 26], we can bound the ε-bracketing number of Θ by

$$\log \mathcal{N}_\Theta(\varepsilon) \le O(H(S^2A+SO)\log(SAOH\varepsilon^{-1})). \tag{14}$$

Plugging the above upper bound back into Theorem 5.2, we immediately recover the sample efficiency guarantee for learning tabular observable POMDPs in [26].

5.1.2 Observable POMDPs with Gaussian emission. To showcase the power of Theorem 5.2 in handling POMDPs with continuous observations, we consider the model of POMDPs with Gaussian mixture emissions (abbreviated as GM-POMDP hereafter), which can be intuitively viewed as tabular observable or weakly revealing POMDPs with observations corrupted by Gaussian noise. The Gaussian emissions further allow us to directly control the bracketing number. We start with the formal definition of GM-POMDPs.

Definition 5.4 (GM-POMDPs). A d-dimensional n-components GM-POMDP is a POMDP where the observation distributions are d-dimensional Gaussian mixtures of size n, i.e.,

$$\mathbb{O}_h(\cdot \mid s) = \sum_{i=1}^{n} \mathbb{W}_h(i \mid s) \times \mathrm{Gauss}(\mathbf{x}_{h,i}, \sigma_h \cdot \mathbf{I}_{d\times d})$$

where $\mathbb{W}_h(\cdot \mid s) \in \Delta_n$, $\mathbf{x}_{h,i}\in\mathbb{R}^d$ and $\sigma_h > 0$.

Without further assumptions on GM-POMDPs, the observable condition can be arbitrarily violated and sample-efficient learning is in general impossible. Therefore, we introduce the following natural separation condition on the Gaussian mixtures in GM-POMDPs, which, once being satisfied, immediately implies the observable condition holds. To condense notations, denote $\mathbb{W} := [\mathbb{W}_h(\cdot \mid s)]_{s\in\mathscr{S}} \in \mathbb{R}^{n\times S}$.

Condition 5.5 (η-separable condition). For all $h\in[H]$, $i\neq j\in[n]$ and $\boldsymbol{\nu}_1,\boldsymbol{\nu}_2\in\Delta_{\mathscr{S}}$, we have

$$\begin{cases} \|\mathbf{x}_{hi}-\mathbf{x}_{hj}\|_2 \ge 4\sqrt{\log(d+1)}\times\sigma_h, \\ \|\mathbb{W}_h(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)\|_1 \ge \eta\|\boldsymbol{\nu}_1-\boldsymbol{\nu}_2\|_1. \end{cases}$$

Condition 5.5 requires that (a) different base Gaussian components are well separated, which is standard in learning Gaussian mixtures in classic theory of statistics, and (b) different latent state distributions induce different weights over the base Gaussian components, which resembles the one-step observable condition for tabular POMDPs. Importantly, in Lemma D.2 in Appendix D.3, we show that any GM-POMDPs satisfying the η-separable condition are $\Omega(\eta)$-observable POMDPs. We remark that GM-POMDPs belongs to the infinite observation extension of tabular observable POMDP but not tabular weakly-revealing POMDPs, this is also the major reason we choose to present observable POMDPs in section 5.1.

To apply OMLE to learning η-separable GM-POMDPs with S states, A actions and n base Gaussian components in $\mathbb{R}^d$, we construct the model class Θ to include all the valid POMDP models wherein (a) the observation distributions are η-separable (Condition 5.5) and (b) the norm of the mean and variance of the base Gaussian components are well behaved. Formally, define

$$\Theta := \Big\{ \big(\mathbb{T},\mathbb{W},\{(\mathbf{x}_{h,i},\sigma_h\cdot\mathbf{I}_{d\times d})\}_{h,i},\boldsymbol{\mu}_1\big) : \\ \eta\text{-separable, } \|\mathbf{x}_{h,i}\|_2 \le C_x \text{ and } \underline{C_\sigma}\le\sigma_h\le\overline{C_\sigma}\Big\}.$$

By carefully discretizing the parameter space and constructing the envelope functions, we can derive the following upper bound for the bracketing number of model class Θ (Lemma D.3 in Appendix D.3):

$$\begin{aligned} \mathcal{N}_\Theta(\varepsilon) \le \exp\Bigg(\Theta\Bigg(H(S^2A+Sn+nd) \\ \times\log(HSAnd(\overline{C_\sigma}/\underline{C_\sigma})(C_x/\overline{C_\sigma})\cdot\varepsilon^{-1})\Bigg)\Bigg) := \overline{\mathcal{N}}_\Theta(\varepsilon). \end{aligned} \tag{15}$$

Now that we know η-separable GM-POMDPs are $\Omega(\eta)$-observable and have bounded bracketing number, we can invoke Theorem 5.2, which gives the following sample complexity guarantee for learning η-separable GM-POMDPs with OMLE.

Proposition 5.6. *Suppose Condition 5.5 holds. There exists an absolute constant $c>0$ such that for any $\delta\in(0,1]$ and $K\in\mathbb{N}$, if we choose $\beta = c\log(K\overline{\mathcal{N}}_\Theta(K^{-1})\delta^{-1})$ in OMLE (Algorithm 1) with $\overline{\mathcal{N}}_\Theta$ specified by Equation (15), then with probability at least $1-\delta$,*

$$\begin{aligned} V^\star - V^{\pi^{\text{out}}} \le \mathrm{poly}\big(\eta^{-1}, S, A, H, n, d, \\ \log K, \log(\overline{C_\sigma}/\underline{C_\sigma}), \log(C_x/\overline{C_\sigma})\big) \times K^{-1/2}. \end{aligned}$$

Despite the observation space being infinitely large and unbounded, the above sample complexity only scales polynomially with respect to the dimension of the observation space and other relevant finite parameters. Finally, we emphasize that although we only focus on POMDPs with Gaussian mixture observations in this subsection, our main result (Theorem 5.2) also applies to learning other types of continuous observation distributions as long as the observable condition (Condition 5.1) holds and the model class has bounded bracketing number.

5.2 Multi-Step Decodable POMDPs

Multi-step decodable POMDPs [9] is subclass of POMDPs in which a suffix of length m of the most recent history contains sufficient information to decode the latent state. To simplify notations, denote $m(h) = \min\{h-m+1, 1\}$. Formally,

Condition 5.7. [m-step decodable POMDPs, 9] There exists an unknown decoder $\zeta = \{\zeta_h\}_{h=1}^H$ such that for every $(o,a)_{1:H}$ we have $s_h = \zeta(z_h)$ for all $h\in[H]$, where $z_h = [(o,a)_{m(h):h-1}, o_h]$.

We remark that neither of multi-step decodable POMDPs or multi-step observable POMDPs is more general than the other. That is, each of them contains statistically tractable POMDP instances that are not included by the other class (see Lemma D.4 in Appendix

D.5 for the concrete constructions). Nonetheless, the following theorem states that multi-step decodable POMDPs also falls into the family of γ-well-conditioned PSRs with $\gamma = 1$ and the sets of core test actions equal to $\mathscr{A}^m$. As a result, OMLE also enjoys polynomial sample efficiency guarantee for learning multi-step decodable POMDPs.

Theorem 5.8. *Let Θ be a model class of m-step decodable POMDPs. Then Θ admits rank-r PSR representations with $Q_h^A = \mathscr{A}^{\min\{m,H-h\}}$ and satisfies Condition 4.3 with $\gamma = 1$. Moreover, there exists an absolute constant $c > 0$ such that for any $\delta \in (0,1]$ and $K \in \mathbb{N}$, if we choose $\beta = c\log(T\mathcal{N}_\Theta(T^{-1})\delta^{-1})$ with $T = KHA^m$ in OMLE (Algorithm 1), then with probability at least $1-\delta$,*

$$V^\star - V^{\pi^{\text{out}}} \leq \text{poly}(r, A^m, H, \log K) \times \sqrt{\frac{\beta}{K}},$$

where we always have $r \leq S$ in any POMDP and $r \leq d_{\text{lin}}$ when the underlying MDP can be represented as a d_{lin}-dimensional kernel linear MDP.

Similar to the results in previous sections, the sample complexity in Theorem 5.8 is independent of the number of observations, which means it also applies to the cases with infinite observations as long as the log-bracketing number of Θ is finite. Moreover, the above result scales with the rank of the PSR representations r instead of the number of latent states S. Although it is well-known $r \leq S$ in any POMDPs, the rank can be much smaller than the number of latent states in certain settings of interest. For example, when the underlying MDP can be represented as a d_{lin}-dimensional linear kernel MDP [39], we have $r \leq d_{\text{lin}}$ while S can be arbitrarily large.

Finite observations. When the number of observations is finite, we can easily upper bound the bracketing number of Θ by the standard discretization arguments as in Equation (14). And by plugging the bound back into Theorem 5.8, we immediately obtain a $\text{poly}(S, A^m, O, H, \log\varepsilon^{-1}) \times \varepsilon^{-2}$ sample complexity upper bound for finding an ε-optimal policy with OMLE in tabular m-step decodable POMDPs.

5.3 POMDPs with A Few Known Core Action Sequences

In Section 5.1, we prove that if a POMDP satisfies that any two state mixtures can be distinguished from the observation distributions induced by taking m-step random actions, then it can be represented as an well-conditioned PSR and OMLE can learn it sample efficiently. However, the sample complexity there scales exponentially with respect to m due to m-step random exploration, which could be prohibitively large even for moderate m. In this subsection, we show that it is possible to get rid of this exponential dependence when there exist a small set of *known* exploratory action sequences so that any two state mixtures can be distinguished from the observation distributions induced by at least one exploratory action sequence.

To simplify notations, we first define the observation-action probability kernel $\mathbb{K}_h$ at step $h \in [H]$: For a latent state s and an action sequence $\mathbf{a}$ of length $l \leq H - h$, $\mathbb{K}_h(s, \mathbf{a})$ is equal to the probability density function over $o_{h:h+l}$ provided that action sequence $\mathbf{a}$ is used from state s and step h. Formally, we consider the following observable-style condition.

Condition 5.9. For any $h \in [H]$, there exists *known* $\mathcal{A}_h$ so that for any $\theta \in \Theta$ and $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \Delta_S$:

$$\max_{\mathbf{a}\in\mathcal{A}_h} \left\| \mathbb{E}_{s\sim\boldsymbol{\nu}_1}\left[\mathbb{K}_h(s,\mathbf{a})\right] - \mathbb{E}_{s\sim\boldsymbol{\nu}_2}\left[\mathbb{K}_h(s,\mathbf{a})\right] \right\|_1 \geq \alpha \|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2\|_1. \quad (16)$$

Notice that in Condition 5.9, the exploratory action sequences in $\mathcal{A}_h$ can be length-$\Omega(H)$, which means a POMDP class Θ that satisfies Condition 5.9 could satisfy the m-step observable condition only for $m = \Omega(H)$. Nonetheless, the following theorem states that as long as Θ satisfies Condition 5.9 with $\mathcal{A}_h$ of small cardinality, then OMLE is guaranteed to learn a near-optimal policy for any $\theta \in \Theta$ within a number of samples that scales only polynomially with respect to $\max_h |\mathcal{A}_h|$.

Theorem 5.10. *Let Θ be a model class of POMDPs that satisfy Condition 5.9 with α and $\{\mathcal{A}_h\}_{h=1}^H$. Then Θ satisfies Condition 4.3 with $\gamma = \alpha/(S + |\mathcal{A}_h|)$ and $Q_h^A = \mathcal{A}_h$. Moreover, there exists an absolute constant $c > 0$ such that for any $\delta \in (0,1]$ and $K \in \mathbb{N}$, if we choose $\beta = c\log(T\mathcal{N}_\Theta(T^{-1})\delta^{-1})$ with $T = KH\max_h|\mathcal{A}_h|$ in OMLE (Algorithm 1), then with probability at least $1-\delta$,*

$$V^\star - V^{\pi^{\text{out}}} \leq \text{poly}(\alpha^{-1}, S, \max_h|\mathcal{A}_h|, H, \log K) \times \sqrt{\frac{\beta}{K}}.$$

When the number of exploratory action sequences ($\max_h|\mathcal{A}_h|$) is small but their length ($\max_h \max_{\mathbf{a}\in\mathcal{A}_h} |\mathbf{a}|$) is large, Theorem 5.10 offers exponentially sharper sample complexity guarantee than Theorem 5.2. As an extreme case, when each $\mathcal{A}_h$ contains a single action sequence of length $H - h$, Theorem 5.10 improves over Theorem 5.2 by a factor of $A^{\Omega(H)}$.

6 BEYOND LOW-RANK SEQUENTIAL DECISION MAKING

In this section, we extend the sample efficiency guarantees of OMLE to any sequential decision making problems under a new structural condition—SAIL condition. We will show that SAIL condition holds not only in all well-conditioned low-rank sequential decision making problems studied in Section 4, but also in problems beyond low-rank sequential decision making, such as factored MDPs, low witness rank problems.

6.1 SAIL Condition

In the fully observable setting, RL with general function approximation has been intensively studied in the theory community, and various complexity measures have been proposed, including Bellman rank [16], witness rank [34], and more [8, 19]. Most of them critical relies on the Bellman error (model-free setting) or the error in model estimation (model-based setting) to have a bilinear structure. Unfortunately, partially observability significantly complicates the learning problem, and neither structure mentioned above hold for even the basic tabular weakly-revealing POMDPs.

Here, we introduce a new general structural condition that is also capable of addressing partially observable setting. Our new condition can be viewed as a generalizations of the bilinear structures mentioned above. Since our focus is OMLE which is a model-based algorithm, our new condition requires the model estimation error to be upper and lower bounded by **S**ummation of **A**bsolute values of **I**ndependent bi**L**inear functions (SAIL). Formally, let Π denote the universal policy space.

Condition 6.1 (SAIL condition)**.** We say model class Θ satisfies (d,κ,B)-SAIL condition with exploration policy function $\Pi_{\exp}:\Pi\rightarrow 2^{\Pi}$, if there exist two sets of mappings $\{f_{h,i}\}_{(h,i)\in[H]\times[m]}$, $\{g_{h,i}\}_{(h,i)\in[H]\times[n]}$ from Θ to $\mathbb{R}^d$ such that for any $\theta,\theta'\in\Theta$, and the optimal policy π_θ of model θ:

$$\sum_{\tilde{\pi}\in\Pi_{\exp}(\pi_\theta)} d_{\mathrm{TV}}(\mathbb{P}^{\tilde{\pi}}_{\theta^\star},\mathbb{P}^{\tilde{\pi}}_{\theta'}) \geq \kappa^{-1}\sum_{h=1}^{H}\sum_{i=1}^{m}\sum_{j=1}^{n}|\langle f_{h,i}(\theta),g_{h,j}(\theta')\rangle|,$$

$$d_{\mathrm{TV}}(\mathbb{P}^{\pi_\theta}_{\theta^\star},\mathbb{P}^{\pi_\theta}_{\theta}) \leq \sum_{h=1}^{H}\sum_{i=1}^{m}\sum_{j=1}^{n}|\langle f_{h,i}(\theta),g_{h,j}(\theta)\rangle|,$$

$$\left(\sum_{i=1}^{m}\|f_{h,i}(\theta)\|_1\right)\cdot\left(\sum_{j=1}^{n}\|g_{h,j}(\theta')\|_\infty\right)\leq B.$$

The first inequality requires the model estimation error of θ' (measured by TV distance) on the exploration policies computed using θ to be lower bounded by a coefficient κ^{-1} times SAIL. In particular, the summand $\langle f_{h,i}(\theta),g_{h,j}(\theta')\rangle$ is a bilinear function, because it is a linear function of $f_{h,i}(\theta)$ (features of θ) when θ' is fixed, and it is also a linear function of $g_{h,j}(\theta')$ (features of θ') when θ is fixed. The second inequality requires the model estimation error of θ on its optimal policy π_θ to be upper bounded by SAIL. The third inequality is a normalization condition.

At a high-level, standard Bellman rank or witness rank can be viewed as conditions similar to SAIL, with the LHS of the first two inequalities replaced by appropriate error measure and the RHS of the first two inequalities replaced by a bilinear function $\langle f(\theta),g(\theta')\rangle$. SAIL condition generalize them by allowing multiple feature functions $\{f_i\}_{i\in[m]},\{g_j\}_{j\in[n]}$ which are indexed by i,j, and taking summation of them. One key structure here is that the indexes are decoupled between two features f,g, and summation is taken over two indexes *independently*. This is crucial in many partially observable applications where m,n are extremely large and we do not want to suffer any dependency on m,n in the sample complexity.

We will prove in Section 6.3 that SAIL condition is very general, which holds not only in all well-conditioned low-rank sequential decision making problems studied in Section 4, but also in problems beyond low-rank sequential decision making, such as factored MDPs, low witness rank problems.

6.2 Theoretical Guarantees for SAIL

Now we present the theoretical guarantees for OMLE in learning sequential decision problems that satisfy the SAIL condition.

Theorem 6.2. *There exists an absolute constant $c>0$ such that for any $\delta\in(0,1]$ and $K\in\mathbb{N}$, if we choose $\beta=c\log(T\mathcal{N}_\Theta(T^{-1})\delta^{-1})$ with $T=K|\Pi_{\exp}|$ in OMLE (Algorithm 1) and assume (d,κ,B)-SAIL condition holds, then with probability at least $1-\delta$, we have*

$$\sum_{k=1}^{K}\left(V^\star-V^{\pi^k}\right)\leq \mathrm{poly}(H)d\left(B+\kappa\sqrt{\beta|\Pi_{\exp}|K}\right)\log^2(K).$$

The result in Theorem 6.2 scales polynomially with respect to the parameters (d,κ,B) and the number of exploration policies $|\Pi_{\exp}|$ in the SAIL condition. Moreover, the result is completely independent of the number of the feature mappings m and n, which is key in addressing the case of well-conditioned PSRs where the SAIL condition requires exponentially many feature mappings. When the log bracketing number has a reasonable growth rate $\log\mathcal{N}_\Theta(T^{-1})\leq\mathrm{polylog}(T)$, Theorem 6.2 guarantees that $K=\tilde{O}(\kappa^2d^2\log\mathcal{N}_\Theta(\varepsilon^{-1})\cdot\varepsilon^{-2})$ episodes suffices for finding an ε-optimal policy. The ε-dependency is optimal up to polylogarithmic factors.

The critical step in proving Theorem 6.2 is our new elliptical potential style lemma for SAIL, which significantly generalizes the standard elliptical potential lemma that only applies to bilinear functions. Our new lemma immediately implies the following result.

Lemma 6.3. *(d,κ,B)-SAIL condition implies the generalized eluder-type condition (Condition 3.1) with $d_\Theta=\kappa^2d^2|\Pi_{\exp}|\mathrm{poly}(H)$ and $\xi(d_\Theta,\Delta,|\Pi_{\exp}|,K)=\tilde{O}\left(\sqrt{d_\Theta\Delta|\Pi_{\exp}|K}+dB\mathrm{poly}(H)\right)$.*

With this lemma, we can directly invoke the guarantee for OMLE (Theorem 3.2), which gives the bound in Theorem 6.2.

Sharper guarantee for single feature mapping. For sequential decision making problems that satisfy the SAIL condition with a single pair of feature mappings (f_h,g_h) for each $h\in[H]$, e.g., sparse linear bandits, factored MDPs, and linear MDPs, we can further derive the following sharper sample complexity guarantee.

Theorem 6.4. *Suppose (d,κ,B)-SAILcondition holds with $m=n=1$. Then under the same choice of parameters as in Theorem 6.2, OMLE satisfies that with probability at least $1-\delta$,*

$$\sum_{k=1}^{K}\left(V^\star-V^{\pi^k}\right)=\tilde{O}\left(\mathrm{poly}(H)\left(dB+\kappa\sqrt{d\beta|\Pi_{\exp}|K}\right)\right).$$

Theorem 6.4 directly implies a regret bound with leading-order term $\tilde{O}(\kappa\sqrt{d\log\mathcal{N}_\Theta(K^{-1})K})$ when the exploration policy function $\Pi_{\exp}$ is equal to identity. This improves a $\sqrt{d}$ factor over Theorem 6.2. Theorem 6.4 also implies a $\tilde{O}(\kappa^2d\log\mathcal{N}_\Theta(\varepsilon^{-1})\cdot\varepsilon^{-2})$ sample complexity upper bound for finding an ε-optimal policy when the log-bracketing number of Θ grows polylogarithmically with respect to the covering precision, which improves a d factor over the sample complexity implied by Theorem 6.2.

6.3 Important Examples of SAIL

In this section, we present several widely studied sequential decision making problems that satisfy the SAIL condition. We remark that all problems considered in this section are MDPs so we will use $\{s_h\}_{h=1}^H$ to denote states.

6.3.1 Low-rank sequential decision making. To demonstrate the generality of the SAIL condition, we prove the following proposition which states that (a) any well-conditioned PSR satisfies the SAIL condition with moderate (d,κ,B), and (b) there exist sequential decision making problems, whose system dynamics matrices have exponentially large rank though, which still satisfy the SAIL condition with mild (d,κ,B).

Proposition 6.5 (well-conditioned PSR $\subseteq$ SAIL$\not\subseteq$ low-rank sequential decision making)**.**

(a) Any rank-r γ-well-conditioned PSR class Θ satisfies the SAIL condition with $d=r$ and $\kappa,B=\mathrm{poly}(r,\gamma^{-1},\max_h|Q_h^A|,A,H)$ and the same choice of $\Pi_{\exp}$ as in Theorem 4.5.

(b) For any $n\in\mathbb{N}$, there exists Θ satisfying the SAIL condition with $d,\kappa,B=O(n)$ and $\Pi_{\exp}(\pi)=\pi$, but for some $\theta\in\Theta$ the system dynamics matrices have rank $\Omega(2^n)$.

6.3.2 *Fully observable problems with low witness rank.*

Witness rank. Witness rank [34] was introduced as a structural parameter for measuring the difficulty of model-based RL. [34] proved that the witness rank of a model class being small suffices to guarantee sample-efficient learning, and several RL settings of interest (e.g., factored MDPs) possess rather moderate witness rank. To simplify notations, let $\mathbb{D}_\theta(s_h, a_h) := \mathbb{P}_\theta((r_h, s_{h+1}) = \cdot \mid s_h, a_h)$. And the witness rank is defined as following:

Definition 6.6 (Q/V-type witness conditions (slightly modified version[5] of [34])). We say model class Θ satisfies (d, κ, B)-witness condition, if there exist two sets of mappings $\{f_h\}_{h\in[H]}, \{g_h\}_{h\in[H]}$ from Θ to $\mathbb{R}^d$, so that for any $\theta, \theta' \in \Theta$ and $h \in [H]$:

$$\begin{cases} \mathbb{E}_{s_h\sim\mathbb{P}^{\pi_\theta}_{\theta^\star},\, a_h\sim\nu(s_h)}\left[\|\mathbb{D}_{\theta'}(s_h,a_h) - \mathbb{D}_{\theta^\star}(s_h,a_h)\|_1\right] \\ \qquad \geq \kappa^{-1}\left|\langle f_h(\theta), g_h(\theta')\rangle\right|, \\ \mathbb{E}_{s_h\sim\mathbb{P}^{\pi_\theta}_{\theta^\star},\, a_h\sim\nu(s_h)}\left[\|\mathbb{D}_{\theta'}(s_h,a_h) - \mathbb{D}_{\theta^\star}(s_h,a_h)\|_1\right] \\ \qquad \leq \left|\langle f_h(\theta), g_h(\theta')\rangle\right|, \\ \|f_h(\theta)\|_1 \times \|g_h(\theta')\|_\infty \leq B, \end{cases}$$

where ν is typically chosen as π_θ (Q-type) or $\pi_{\theta'}$ (V-type).

The Q-type witness condition requires that at each single step the expected model discrepancy between the true model $\theta^\star$ and model candidate θ' under the state-action distribution induced by the optimal policy of θ is roughly proportional to the inner product of the features of θ and θ'. And the V-type version is defined similarly except that the last action a_h is sampled from $\pi_{\theta'}$ instead of π_θ. By basic algebra, we can easily relate the above per-step model discrepancy in witness condition to the whole-trajectory model discrepancy in SAIL condition, which leads to the following conclusion that the SAIL condition is satisfied with almost the same (d, κ, B) whenever either Q-type or V-type witness condition holds.

Proposition 6.7. *For any model class Θ, we always have*

- *Q-type (d, κ, B)-witness condition implies $(d, 2\kappa, B)$-SAIL condition with $\Pi_{\exp}(\pi) = \{\pi\}$, and $m = n = 1$.*
- *V-type (d, κ, B)-witness condition implies $(d, 2A\kappa, B)$-SAIL condition with $\Pi_{\exp}(\pi) = \{\pi_{1:h} \circ \mathrm{Uniform}(\mathscr{A}) : h \in [0, H-1]\}$, and $m = n = 1$.*

In the case when $\log \mathcal{N}_\Theta$ grows polylogarithmically with respect to the covering precision, by plugging Proposition 6.7 back into Theorem 6.4, we immediately obtain a $\tilde{O}(A^2\kappa^2 d \log \mathcal{N}_\Theta(\varepsilon^{-1})\varepsilon^{-2})$ sample complexity upper bound for OMLE in the V-type witness rank setting, which improves over the quadratic dependence on d in [34]. Moreover, OMLE further enjoys a $\tilde{O}(\kappa\sqrt{d \log \mathcal{N}_\Theta(K^{-1})K})$ regret guarantee in the Q-type witness rank setting, which is new to our knowledge.

Factored MDPs. In factored MDPs, the state admits a factored structure. Concretely, each state s consists of m factors denoted as $(s[1], \ldots, s[m]) \in \mathcal{X}^m$. Moreover, each factor $i \in [m]$ has a parent set denoted by $\mathrm{pa}_i \subseteq [m]$, with respect to which the transition admits the following factorized form:

$$\mathbb{P}(s_{h+1} \mid s_h, a_h) = \prod_{i=1}^m \mathbb{P}^i(s_{h+1}[i] \mid s_h[\mathrm{pa}_i], a_h). \tag{17}$$

In other words, the transition of the i^{th} factor of states are only determined by a subset of all factors, that is pa_i, instead of the whole state. In factored MDPs, it is standard to assume the factorization structure and the reward function are *known* [21, 34]. Therefore, our model class Θ only needs to parameterize the transitions under the given factorization structure.

The following proposition states that when the factorization structure is known, factored MDPs admit low witness rank structure.

Proposition 6.8. *Let Θ consist of all the factored MDPs with the same factorization structure $\{\mathrm{pa}_i\}_{i=1}^m$. Then Θ satisfies Q-type witness condition with $d = A\sum_{i=1}^m |\mathcal{X}|^{|\mathrm{pa}_i|}$, $\kappa = m$, and $B = \sum_{i=1}^m |\mathcal{X}|^{|\mathrm{pa}_i|}$.*

Kernel linear MDPs. In kernel linear MDPs [39], the transition functions can be represented as a linear functions of the tensor product of two *known* feature mappings. Formally, the learner is provided with features $\phi : \mathscr{S} \times \mathscr{A} \to \mathbb{R}^{d_{\text{lin}}}$ and $\psi : \mathscr{S} \to \mathbb{R}^{d_{\text{lin}}}$ so that for any $h \in [H]$, there exists $\mathbf{W}_h \in \mathbb{R}^{d_{\text{lin}}\times d_{\text{lin}}}$ satisfying $\mathbb{P}_h(s_{h+1} \mid s_h, a_h) = \phi(s_h, a_h)^\top \mathbf{W}_h \psi(s_{h+1})$ for all $(s_h, a_h, s_{h+1}) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}$. Besides, kernel linear MDPs satisfy the normalization condition: (a) $\|\phi(s_h, a_h)\|_2 \leq C_\phi$ for all (s_h, a_h), (b) $\|\sum_{s_{h+1}} \psi(s_{h+1}) f(s_{h+1})\|_1 \leq C_\psi$ for all $\|f\|_\infty \leq 1$, and (c) $\|\mathbf{W}_h\|_2 \leq C_W$.

For simplicity, we assume the reward function is known. Previous works [e.g., 39] have shown that kernel linear MDPs are capable of representing various examples with moderate dimension d_{lin}, e.g., tabular MDPs with $d_{\text{lin}} = SA$. The following proposition states that kernel linear MDPs also fall into the low witness rank framework with the same ambient dimension.

Proposition 6.9. *Let Θ be the family of d_{lin}-dimensional kernel linear MDPs. Then Θ satisfies V-type witness condition with $d = d_{\text{lin}}$, $\kappa = 1$, and $B = 2(\sqrt{d_{\text{lin}}}C_\phi C_W C_\psi + 1)$.*

Sparse linear bandits. In sparse linear bandits, the mean reward function can be represented as a sparse linear function of the arm feature. Formally, we have $R_\theta(a) = \langle a, \theta\rangle$ where (i) $a \in \mathscr{A} \subseteq B^{d_{\text{lin}}}_{C_\mathscr{A}}(0)$, (ii) $\Theta := \{\theta \in B^{d_{\text{lin}}}_{C_\Theta}(0) : \|\theta\|_0 \leq m$ and $\langle \theta, a\rangle \in [0,1]$ for any $a \in \mathscr{A}\}$. Without loss of generality, assume the stochastic reward feedback is binary[6]. The following proposition states that the witness rank of sparse linear bandits is no larger than the ambient dimension d_{lin}.

Proposition 6.10. *Let Θ be the family of d_{lin}-dimensional m-sparse linear bandit. Then Θ satisfies Q-type witness condition with $d = d_{\text{lin}}$, $\kappa = 1$, and $B = 4\sqrt{d_{\text{lin}}}C_\Theta C_\mathscr{A}$.*

By combining Proposition 6.10 with Proposition 6.7 and 6.4, we recover the optimal regret for sparse linear bandits $\tilde{O}(\sqrt{md_{\text{lin}}K})$ up to a polylogarithmic factor.

ACKNOWLEDGMENTS

Chi Jin gratefully acknowledges the funding support from Google LLC, Princeton Project X innovation fund, and NSF CAREER IIS-2239297.

[6] If the reward feedback $\hat{r}$ is a real number in $[0, 1]$, we can binarize it by sampling x from Bernoulli$(\hat{r})$ and then using x as the reward feedback instead. Such modification will not change the mean reward.

REFERENCES

[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. 2019. Solving Rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113* (2019).

[2] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. 2017. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*. PMLR, 263–272.

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

[4] Noam Brown and Tuomas Sandholm. 2019. Superhuman AI for multiplayer poker. *Science* 365, 6456 (2019), 885–890.

[5] Qi Cai, Zhuoran Yang, and Zhaoran Wang. 2022. Reinforcement learning from partial observation: Linear function approximation with provable sample efficiency. In *International Conference on Machine Learning*. PMLR, 2485–2522.

[6] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. Deep blue. *Artificial intelligence* 134, 1-2 (2002), 57–83.

[7] Fan Chen, Yu Bai, and Song Mei. 2022. Partially Observable RL with B-Stability: Unified Structural Condition and Sharp Sample-Efficient Algorithms. *arXiv preprint arXiv:2209.14990* (2022).

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

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

[10] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. 2021. The Statistical Complexity of Interactive Decision Making. *arXiv preprint arXiv:2112.13487* (2021).

[11] Sara A Geer, Sara van de Geer, and D Williams. 2000. *Empirical Processes in M-estimation*. Vol. 6. Cambridge University Press.

[12] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. 2022. Learning in Observable POMDPs, without Computationally Intractable Oracles. *arXiv preprint arXiv:2206.03446* (2022).

[13] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. 2022. Planning in Observable POMDPs in Quasipolynomial Time. *arXiv preprint arXiv:2201.04735* (2022).

[14] Milos Hauskrecht and Hamish Fraser. 2000. Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artificial Intelligence in Medicine* 18, 3 (2000), 221–244.

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

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

[17] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. 2018. Is Q-learning provably efficient? *Advances in neural information processing systems* 31 (2018).

[18] Chi Jin, Sham M Kakade, Akshay Krishnamurthy, and Qinghua Liu. 2020. Sample-Efficient Reinforcement Learning of Undercomplete POMDPs. *NeurIPS* (2020).

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

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

[21] Michael Kearns and Daphne Koller. 1999. Efficient reinforcement learning in factored MDPs. In *IJCAI*, Vol. 16. 740–747.

[22] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. 2016. PAC reinforcement learning with rich observations. *Advances in Neural Information Processing Systems* 29 (2016).

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

[24] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. 2011. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*. IEEE, 163–168.

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

[26] Qinghua Liu, Alan Chung, Csaba Szepesvari, and Chi Jin. 2022. When Is Partially Observable Reinforcement Learning Not Scary?. In *Proceedings of Thirty Fifth Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 178)*. PMLR, 5175–5220. https://proceedings.mlr.press/v178/liu22f.html

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

[28] Elchanan Mossel and Sébastien Roch. 2005. Learning nonsingular phylogenies and hidden Markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*. 366–375.

[29] Martin Mundhenk, Judy Goldsmith, Christopher Lusena, and Eric Allender. 2000. Complexity of finite-horizon Markov decision process problems. *Journal of the ACM (JACM)* 47, 4 (2000), 681–720.

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

[31] Daniel Russo and Benjamin Van Roy. 2013. Eluder Dimension and the Sample Complexity of Optimistic Exploration.. In *NIPS*. Citeseer, 2256–2264.

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

[33] Satinder Singh, Michael James, and Matthew Rudary. 2012. Predictive state representations: A new theory for modeling dynamical systems. *arXiv preprint arXiv:1207.4167* (2012).

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

[35] Emanuel Todorov and Weiwei Li. 2005. A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the 2005, American Control Conference, 2005.* IEEE, 300–306.

[36] Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. 2022. Provably efficient reinforcement learning in partially observable dynamical systems. *arXiv preprint arXiv:2206.12020* (2022).

[37] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michael Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[38] Nikos Vlassis, Michael L Littman, and David Barber. 2012. On the computational complexity of stochastic controller optimization in POMDPs. *ACM Transactions on Computation Theory (TOCT)* 4, 4 (2012), 1–8.

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

[40] Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. 2022. PAC Reinforcement Learning for Predictive State Representations. *arXiv preprint arXiv:2207.05738* (2022).

Received 2022-11-07; accepted 2023-02-06