Submitted to Statistical Science

Partially Observable RL: Benign Structures and Simple Generic Algorithms

Qinghua Liu , Chi Jin **(draft version; under review)**

Abstract. Partially observable Reinforcement Learning (RL) applications, where agents must make a series of decisions without complete knowledge of the underlying states of the system, are widespread. Partially observable RL can be notoriously difficult – well-known information-theoretic results show that learning partially observable Markov decision processes (POMDPs) requires an exponential number of samples in the worst case. However, this does not preclude the possibility of identifying rich subclasses of structured POMDPs for which learning remains feasible.

This survey aims to offer a high-level overview of recent advancements in learning structured POMDPs. We will identify clean and practical problem structures that facilitate sample-efficient learning. Additionally, we will introduce simple and generic algorithms for learning POMDPs under different structural conditions. Finally, we will provide a unified view of these results under the framework of well-conditioned predictive state representation, which further reveals new tractable classes of partially observable problems along the way.

Key words and phrases: partially observable reinforcement learning.

1. INTRODUCTION

A wide range of modern artificial intelligence applications can be framed as sequential decision-making problems, where an agent interacts with an unknown environment over time and learns to make a sequence of decisions based on intermediate feedback. Sequential decision-making encompasses problems like Atari games [59], Go [67], Chess [14], and basic control systems [72], where states are fully accessible to the learner (the *fully observable* setting). It also includes applications such as StarCraft [77], Poker [12], robotics with local sensors [1], autonomous driving [49], and medical diagnostic systems [29], where observations only reveal partial information about the underlying states (the *partially observable* setting). While fully observable sequential decision-making problems have been extensively studied theoretically in recent years, partially observable problems remain comparatively less understood.

Unlike fully observable systems, a learner in partially observable systems can only access observations that provide partial information about the underlying states. Observations, in general, are no longer Markovian. Consequently, it is insufficient for the learner to make decisions solely based on the current observation or information. Instead, the learner must also infer the latent states using past histories (memories). Such histories of observations have exponentially many possibilities, leading to well-known hardness results in the worst case, both computationally [60, 61, 64, 78] and statistically [40].

Despite these prohibitive hardness results in theory, empirical reinforcement learning (RL) research has achieved remarkable success in solving various tasks under partial observability. This discrepancy arises because previous hardness results are proven for generic worst-case scenarios, while real-world tasks typically possess benign structures that enable efficient learning. Consequently, this raises two crucial questions for RL theory:

What structural conditions enable tractable partially observable RL?

and

How to design algorithms to solve those structured problems?

This paper provides a concise overview of the core ideas and algorithms behind recent advancements in addressing these questions. Specifically, we focus on the

Qinghua Liu (email: qinghual@princeton.edu) is a postdoc researcher at Microsoft Research, NYC. Chi Jin (email: chij@princeton.edu) is an assistant professor at Princeton University.

widely used model—the partially observable Markov decision process (POMDP) [5]—and introduce clear and practical structural conditions that allow for sample-efficient learning of POMDPs. We also present simple and generic algorithms (e.g., variants of Maximum Likelihood Estimation (MLE)) for solving these structured POMDPs. Finally, we unify most seemingly unrelated tractable subclasses of POMDPs through a unified framework called well-conditioned PSR [54]. This unified view not only allows for solving different POMDP subclasses using a simple generic algorithm via unified analysis but also reveals new tractable POMDP subclasses not covered by previous works.

Due to page limit, this survey focuses on the learning-from-interaction setting, which is prevalent in practice. In this setting, an agent learns to solve a decision-making problem by repeatedly interacting with an *unknown* environment and receiving feedback over time. Other important settings, such as offline partially observable RL [e.g., 57, 66, 75, 81] are not covered in this article. Furthermore, the aim of this survey is to present a brief, high-level overview of recent advancements in partially observable RL. Consequently, we will focus on presenting a few key results in the most fundamental setup. We refer interested readers to works [e.g., 2, 26, 27, 55, 74, 79] for results in more complex settings.

Roadmap To begin with, we formally define POMDP and set up basic notations in Section 2. Then we review the classic hardness results for learning POMDPs in Section 3. To circumvent the hardness results, Section 4 introduces sufficient conditions for identifying POMDP parameters via tensor decomposition. However, those conditions are too strong to be practical and parameter identification is unnecessary if the goal is merely to find a good policy. Therefore, Section 5 demonstrates that all prior conditions except for the so-called observable condition (a.k.a. weakly revealing condition) can be removed and a simple generic algorithm–Optimistic MLE sample-efficiently learns any POMDP under the observable condition. In Section 6, we further introduces other tractable subclasses of POMDPs which are closely related to but not fully captured by the observable condition, such as latent MDPs and decodable POMDPs. Finally, Section 8 unifies most of these seemingly different POMDP subclasses under a generic framework named well-conditioned PSR and solves them by Optimistic MLE in a unified manner. Along the way, this new framework reveals new tractable subclasses not captured by prior results such as observable POMDPs with continuous observations. We conclude the paper with discussions on future directions in Section 9.

2. PRELIMINARY

For a positive integer n, we let $[n] = \{1, \ldots, n\}$. We consider episodic nonstationary partially observable Markov decision processes (POMDP) [5]. These processes generalize the standard Markov decision processes by making agents observe a "noisy function" of the state of a controlled Markov process. Formally, such a POMDP is specified by a tuple $(\mathscr{S}, \mathscr{A}, \mathscr{O}; H, \mu_1, \mathbb{T}, \mathbb{O}; r)$. Here $\mathscr{S}, \mathscr{A}$ and $\mathscr{O}$ denote the space of state, action and observation respectively, with respective cardinalities $|\mathscr{S}| = S$, $|\mathscr{A}| = A$ and $|\mathscr{O}| = O$; H denotes the length of each episode; $\mu_1 \in \Delta_S$ denotes the distribution of the initial state where Δ_S is the $(S-1)$-dimensional probability simplex defined over the state space $\mathscr{S}$; $\mathbb{T} = \{\mathbb{T}_{h,a}\}_{(h,a)\in[H-1]\times\mathscr{A}}$ denotes the collection of transition matrices where $\mathbb{T}_{h,a}$ is the $S \times S$ *transition matrix* of action a at stage h such that $\mathbb{T}_{h,a}(\cdot \mid s)$ gives the distribution of the next state if the agent takes action a at state s and stage h; $\mathbb{O} = \{\mathbb{O}_h\}_{h\in[H]}$ denotes the collection of *emission matrices* of size $O \times S$ so that $\mathbb{O}_h(\cdot \mid s)$ gives the distribution over observations at stage h conditioned on the current hidden state being s; and and $r = \{r_h\}_{h\in[H]}$ are the known reward functions from $\mathscr{O} \times \mathscr{A}$ to $[0, 1]$ such that the agent will receive reward $r_h(o, a)$ after she observes $o \in \mathscr{O}$ and executes action a at step h. [1]

In a POMDP, the states are always hidden from the agent: she can only see the observations and her own actions. At the beginning of each episode, the environment samples an initial state s_1 from μ_1. At each stage $h \in [H]$, the agent first observes o_h that is sampled from $\mathbb{O}_h(\cdot \mid s_h)$, the observation distribution of hidden state s_h at stage h. Then the agent takes action a_h, and receives reward $r_h(o_h, a_h)$ that is computed from o_h and a_h. After this, the environment transitions to s_{h+1}, whose distribution follows $\mathbb{T}_{h,a_h}(\cdot \mid s_h)$. The current episode terminates immediately after a_H is taken. We use $\tau_h = (o_1, a_1, \ldots, o_h, a_h)$ to denote a trajectory from stage 1 to stage h.

A policy $\pi = \{\pi_h : \mathcal{T}_h \to \Delta_A\}_{h=1}^{H}$ is a collection of H functions where $\mathcal{T}_h = (\mathscr{O} \times \mathscr{A})^{h-1} \times \mathscr{O}$ denotes the set of all length-h histories. Given a policy π, we use V^π to denote its value, which is defined as the expected total reward received under policy π:

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

[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 latent states through rewards beyond observations. We remark that most 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 τ_H, and is only received at the end of each episode.

where the expectation is with respect to the randomness of the transitions, observations and the policy. Since the state, action, observation spaces and the horizon are all finite, there always exists an optimal policy $\pi^\star$ that achieves the optimal value $V^\star := \sup_\pi V^\pi$. Different from MDPs, the optimal policies in POMDPs are in general history-dependent instead of only depending on the current observation, which makes not only learning, but already computing a near-optimal policy in known POMDPs more challenging than doing the same in MDPs.

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.

Notation 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, $\sigma_k(\mathbf{B})$ to denote its k^{th} largest singular value, and $\mathbf{B}^\dagger$ to denote its Moore-Penrose inverse. For a vector $\mathbf{b} \in \mathbb{R}^m$, we use $\operatorname{diag}(\mathbf{b})$ to denote a diagonal matrix with $[\operatorname{diag}(\mathbf{b})]_{ii} = \mathbf{b}_i$.

3. FUNDAMENTAL LIMITS FOR PLANNING AND LEARNING

It has long been established that generic POMDPs are hard to solve, both computationally and statistically.

Computational hardness On the computational front, [64] demonstrate that computing the optimal policy (planning) from a given POMDP model is PSPACE-complete. This is intuitive considering that the optimal policy of a POMDP is typically history-dependent and requires exponential memory for storage. Furthermore, [78] prove that even identifying the optimal memoryless policy, a policy that selects actions based solely on the current-step observation, remains NP-hard. This result is intriguing because, unlike an optimal history-dependent policy, an optimal memoryless policy only occupies at most linear space for storage. The proof involves demonstrating that any 3-SAT problem can be reduced to a POMDP of comparable size.

Statistical hardness Regarding the statistical complexity of learning POMDPs, [40] show that learning a good policy requires an exponential number of samples (interactions with the environment) in the worst case, even with unlimited computational power, when the model parameters are unknown. Their construction is as follows:

There are two states, s_g and s_b. The agent starts from s_g at the beginning of each episode. At each stage h, there is an *unknown* special action $a_h^\star$ such that the agent will transition to s_g *if and only if* it takes $a_h^\star$ at state s_g, and transition to s_b otherwise. The observations are designed such that the agent always receives the same dummy observation containing zero reward during the first $H-1$ stages, regardless of the current state and action. However, in the final stage, the agent will receive an observation containing reward one if it is in state s_g and an observation with zero reward otherwise.

It is straightforward to verify that the only optimal policy is to sequentially execute the special action sequence $(a_1^\star, \ldots, a_{H-1}^\star)$ from stage 1 to stage $H-1$, which leads to a reward of one in the final stage. Nevertheless, identifying this special action sequence essentially requires blindly attempting all possible action sequences because the only feedback obtained from interaction is the final-stage observation, which merely indicates whether the special action sequence was followed precisely or not.

Takeaway In summary, solving generic POMDPs is infeasible in the worst case. This starkly contrasts with the empirical success of RL in solving tasks with partial observation. We believe the key reason behind this distinction is that practical tasks typically exhibit benign structures that are amenable to commonly used RL algorithms. Consequently, the remainder of this paper will focus on surveying the structural conditions that render POMDPs tractable.

4. LEARNING POMDP VIA PARAMETER IDENTIFICATION

Model-free methods such as Q-learning or policy optimization have been prevalent in fully observable RL. However, their applicability in partially observable RL is significantly limited. This is because a value function or history-based policy typically requires an exponential number of parameters for representation, which in turn could cost an exponential number of samples for learning. Consequently, most existing research in partially observable RL has focused on model-based approaches. These approaches aim to learn the POMDP parameters (transition matrices, observation matrices, and initial state distribution), the number of which scales only polynomially with the number of states, actions, and observations. Once a good parameter estimate is obtained, it can be leveraged to compute a near-optimal policy.

4.1 Tensor decomposition

Guo et al. [28] and Azizzadenesheli et al. [9] identified the following conditions as jointly sufficient for parameter identification in stationary POMDPs where the transition and emission matrices are consistent across different stages $h \in [H]$. For notational simplicity, we will omit the h dependence in $\mathbb{O}_h$ and $\mathbb{T}_h$ throughout this section.

CONDITION 4.1. There exist $\alpha > 0$ such that

(a) **full-rank transition**:

$$\min_{a\in\mathscr{A}} \sigma_S(\mathbb{T}_a) \geq \alpha,$$

(b) **full-rank observation**:

$$\sigma_S(\mathbb{O}) \geq \alpha,$$

(c) **full-coverage intial state distribution**:

$$\min_{s\in[S]} \mu_1[s] \geq \alpha.$$

The algorithms proposed in [9, 28] leverage the following tensor decomposition relation: Suppose an agent consistently takes a fixed action a, irrespective of their observations. Then,

$$\begin{aligned}(2)\qquad &\mathbb{P}(o_1 = \cdot, o_2 = \cdot, o_3 = \cdot)\\ &= (\mathbb{O}\mathrm{diag}(\mu_1)\mathbb{T}_a^\top) \otimes \mathbb{O} \otimes \mathbb{T}_a \in \mathbb{R}^{O\times O\times O},\end{aligned}$$

where $\otimes$ denotes tensor product. It is well established in the tensor decomposition literature [e.g., 48] that we can recover matrices $A, B, C \in \mathbb{R}^{n\times d}$ from tensor $T = A \otimes B \otimes C$ if $\mathrm{rank}(A) = \mathrm{rank}(B) = \mathrm{rank}(C) = d$. In the context of decomposition (2), this implies that we can recover $\mathbb{O}$, $\mathbb{T}_a$ and μ_1 from $\mathbb{P}(o_1 = \cdot, o_2 = \cdot, o_3 = \cdot)$ if $\mathrm{rank}(\mathbb{O}) = \mathrm{rank}(\mathbb{T}_a) = \|\mu_1\|_0 = S$.

Condition 4.1 can be interpreted as a robust version of this full-rank requirement. It guarantees the robust recovery of $\mathbb{O}$, $\mathbb{T}_a$ and μ_1 from an imperfect empirical estimation of $\mathbb{P}(o_1 = \cdot, o_2 = \cdot, o_3 = \cdot)$, utilizing off-the-shelf tensor decomposition methods [e.g., 3]. Guo et al. [28] and Azizzadenesheli et al. [9] provide the following guarantee for learning POMDPs via tensor decomposition.

THEOREM 4.2 (tensor decomposition). *Suppose Condition 4.1 holds. Tensor decomposition learns a model estimate $\hat{\theta} = (\hat{\mathbb{T}}, \hat{\mathbb{O}}, \hat{\mu}_1)$ satisfying*

$$\max_a \|\hat{\mathbb{T}}_a - \mathbb{T}_a\|_1 \leq \varepsilon, \quad \|\hat{\mathbb{O}} - \mathbb{O}\|_1 \leq \varepsilon, \quad \|\hat{\mu}_1 - \mu_1\|_1 \leq \varepsilon$$

within

$$\mathrm{poly}(S, A, O, \alpha^{-1}, \varepsilon^{-1}, \log \delta^{-1})$$

samples with probability at least $1 - \delta$.

After obtaining the model estimate $\hat{\theta}$, we proceed to employ *brute-force planning* to compute its optimal policy $\hat{\pi}$. This planning process, however, incurs a time complexity that is exponential in $\min\{S, H\}$. The suboptimality of $\hat{\pi}$ can be controlled as follows: Given a trajectory $s_1, o_1, a_1, \ldots, s_H, o_H, a_H$, we denote the subsequence proceeding s_h as $< s_h$ and the subsequence following s_h as $> s_h$. We similarly define $< o_h$ ($> o_h$) and $< a_h$ ($>_h$). For any policy π, we have

$$\begin{aligned}&\frac{1}{H}\left|V^\pi(\theta^\star) - V^\pi(\hat{\theta})\right| \leq \sum_{\tau_H}\left|\mathbb{P}^\pi_{\theta^\star}(\tau_H) - \mathbb{P}^\pi_{\hat{\theta}}(\tau_H)\right|\\ &\leq \sum_{(s,o,a)_{1:H}}\Bigg|(\mu_1(s_1) - \hat{\mu}(s_1))\mathbb{P}^\pi_{\theta^\star}(> s_1 \mid s_1)\\ &\quad + \sum_{t=1}^{H}\mathbb{P}^\pi_{\hat{\theta}}(< o_t)\Big(\mathbb{O}(o_t \mid s_t) - \hat{\mathbb{O}}(o_t \mid s_t)\Big)\mathbb{P}^\pi_{\theta\star}(> o_t \mid\leq o_t)\\ &\quad + \sum_{t=1}^{H-1}\mathbb{P}^\pi_{\hat{\theta}}(< s_{t+1})\Big(\mathbb{T}_{a_t}(s_{t+1} \mid s_t) - \hat{\mathbb{T}}_{a_t}(s_{t+1} \mid s_t)\Big)\\ &\quad \times \mathbb{P}^\pi_{\theta\star}(> s_{t+1} \mid\leq s_{t+1})\Bigg|\\ &\leq \|\mu_1(s_1) - \hat{\mu}(s_1)\|_1 + H\|\mathbb{O} - \hat{\mathbb{O}}\|_1\\ &\quad + (H-1)\max_a \|\mathbb{T}_a - \hat{\mathbb{T}}_a\|_1,\end{aligned}$$

where the final inequality utilizes the fact that

$$\begin{aligned}&\sum_{<o_t}\mathbb{P}^\pi_{\hat{\theta}}(< o_t) = \sum_{<s_{t+1}}\mathbb{P}^\pi_{\hat{\theta}}(< s_{t+1})\\ &= \sum_{>o_t}\mathbb{P}^\pi_{\theta^\star}(> o_t \mid\leq o_t) = \sum_{>s_{t+1}}\mathbb{P}^\pi_{\theta^\star}(> s_{t+1} \mid\leq s_{t+1}) = 1.\end{aligned}$$

4.2 Maximum likelihood estimation

Although the tensor decomposition step is computationally efficient, the subsequent brute-force planning step incurs exponential time complexity. Consequently, the overall algorithm remains computationally inefficient. This implies that the overall complexity would not be significantly altered even if we substituted the intricate tensor decomposition step with a computationally expensive but conceptually simpler alternative. In fact, during the preparation of this survey, we realized that tensor decomposition can be readily replaced with the classical maximum likelihood estimation (MLE) algorithm.

Simple algorithm via MLE For each fixed $a \in \mathscr{A}$, we sample N trajectories $\mathcal{D}_a := \{\tau_H^{(i)}\}_{i\in[N]}$ uniformly at random by consistently executing action a. Subsequently, we compute the model estimate via MLE over the collected trajectories $\{\mathcal{D}_a\}_{a\in\mathscr{A}}$:

$$\hat{\theta} \in \operatorname*{argmax}_\theta \sum_a \sum_{\tau_H\in\mathcal{D}_a} \log \mathbb{P}_\theta(\tau_H).$$

Finally, we employ brute-force planning to compute the optimal policy for the estimated model $\hat{\theta}$.

Proof sketch As a consequence of the classical MLE analysis [e.g., 23], we have that with high probability, the estimate $\hat{\theta} = (\hat{\mathbb{T}}, \hat{\mathbb{O}}, \hat{\mu}_1)$ satisfies

$$\max_a \left\| (\hat{\mathbb{O}}\mathrm{diag}(\hat{\mu}_1)\hat{\mathbb{T}}_a^\top) \otimes \hat{\mathbb{O}} \otimes \hat{\mathbb{T}}_a - (\mathbb{O}\mathrm{diag}(\mu_1)\mathbb{T}_a^\top) \otimes \mathbb{O} \otimes \mathbb{T}_a \right\|_1 \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{\mathrm{poly}(S,A,O)}{N}}\right),$$

which, in conjunction with Condition 4.1, further implies that

$$\max_a \|\hat{\mathbb{T}}_a - \mathbb{T}_a\|_1,\ \|\hat{\mathbb{O}} - \mathbb{O}\|_1,\ \|\hat{\mu}_1 - \mu_1\|_1 \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{\mathrm{poly}(S,A,O,\alpha^{-1})}{N}}\right)$$

using existing analysis for tensor decomposition. The remaining proof follows the same as in Section 4.1.

4.3 Limitations

We discuss the limitations of the results presented in section.

Strong assumptions The parameterization of a POMDP is generally non-unique. There can exist multiple sets of parameters that represent the same underlying POMDP. In such cases, parameter identification is inherently impossible, rendering algorithms reliant on it inapplicable. In this section, the uniqueness of parameterization is ensured by imposing strong assumptions that are arguably restrictive in both theoretical and practical sense. For instance, prior works [e.g., 7, 8, 11, 34] have demonstrated that MDPs (the fully observable counterpart of POMDPs) can be efficiently learned without any assumptions on the state transitions and the initial distribution. In contrast, the approach discussed here require both full-rank transition matrices and an initial state distribution that covers all states, conditions that are rarely satisfied in practice.

In Section 5, we address this limitation by shifting our focus from parameter identification to system dynamics identification. This alternative approach requires no assumption on state transitions or initial state coverage at all.

No exploration A central challenge in RL is how to strategically explore the state-action space to gather informative data for computing high-quality policies. However, the results presented in this section bypass this challenge by assuming that the initial state distribution adequately covers all states. This assumption often does not hold in many common RL tasks, such as video games and robotics. Addressing exploration is particularly challenging in partially observable RL, where states remain hidden from the learner. This concealment makes it difficult to determine whether new states have been visited or whether sufficient exploration of all reachable states has been achieved.

In Section 5.3, we combine MLE with the principle of optimism in the face of uncertainty [6] to tackle the exploration challenge central to RL.

Computational inefficiency The algorithms presented in this section rely on brute-force planning, which results in a computational complexity that scales exponentially with $\min\{S, H\}$. We will improve the computational complexity of planning for certain structured POMDPs in Section 5.5.

5. LEARNING POMDP WITH EXPLORATION

Motivated by the discussion in Section 4.3, we shift our focus from estimating parameters to estimating system dynamics, that is, the set of trajectory distributions $\{\mathbb{P}(o_{1:H} = \cdot \mid a_{1:H}) : a_{1:H} \in \mathscr{A}^H\}$. This enables us to predict the outcome for any given policy, which in turn suffices for computing a near-optimal policy. Furthermore, we will introduce algorithms that leverage the principle of optimism in the face of uncertainty to address the challenge of exploration.

5.1 Undercomplete POMDP

The question now becomes how to efficiently identify the system dynamics without identifying the underlying parameters. The first step is to cast a POMDP as an observable operator model (OOM) [32]. Notably, for any POMDP with full-rank observation matrices (i.e., $\mathrm{rank}(\mathbb{O}_h) = S$ for $h \in [H]$), its system dynamics can be represented using the following OOM formulation:

(3)
$$\mathbb{P}(o_{1:H} \mid a_{1:H}) = \mathbf{1}^\top \mathbf{B}_H(o_H, a_H) \cdots \mathbf{B}_1(o_1, a_1)\mathbf{b}_0$$

where

$$\begin{cases} \mathbf{B}_h(o,a) = \mathbb{O}_{h+1}\mathbb{T}_{h,a}\mathrm{diag}(\mathbb{O}_h(o \mid \cdot))\mathbb{O}_h^\dagger, \\ \mathbf{b}_0 = \mathbb{O}_1\mu_1. \end{cases}$$

Algorithm OOM-UCB. Drawing inspiration from previous works on learning hidden Markov models (HMMs) [31], Jin et al. [35] observe that the OOM operators can be estimated using certain linear constraints: $\mathbf{B}_h(a,o)\mathbf{P}_h(a) = \mathbf{Q}_h(o,a)$ where $\mathbf{P}$ and $\mathbf{Q}$ are probability matrices that can be estimated from empirical observations. Jin et al. [35] further derive confidence intervals to quantify the uncertainty associated with these estimators. These confidence intervals are then combined with the principle of optimism in the face of uncertainty [6] to address the exploration challenge central to RL. This leads to a new algorithm, OOM-UCB, with the following theoretical guarantee under the weakly revealing condition.

CONDITION 5.1 (weakly revealing condition). There exists $\alpha > 0$ such that $\min_{h\in[H]} \sigma_S(\mathbb{O}_h) \geq \alpha > 0$.

The intuition behind Condition 5.1 is the same as that of the single-step observable condition, which we will elaborate on in Section 5.2.

THEOREM 5.2 (OOM-UCB). *Suppose Condition 5.1 holds. The OOM-UCB algorithm learns an ε-optimal policy within*

$$\mathrm{poly}(S, A, O, H, \alpha^{-1}, \log(\varepsilon\delta)^{-1})/\varepsilon^2$$

samples with probability at least $1-\delta$.

Theorem 5.2 is significant in several respects. It is the first result demonstrating the feasibility of sample-efficiently learning a rich class of POMDPs without imposing any assumptions on the hidden state transitions and distributions, as was the case in prior works [9, 28]. Moreover, it addresses the long-standing open problem of how to achieve sample-efficient exploration in POMDPs.

Nevertheless, Theorem 5.2 still exhibits two limitations. First, the weakly revealing condition is applicable only to undercomplete POMDPs where the number of observations is at least as large as the number of hidden states, i.e., $S \leq O$. This excludes all overcomplete POMDPs, as $S > O$ implies $\sigma_S(\mathbb{O}_h) = 0$. Second, OOM-UCB is a relatively specialized algorithm that, to our knowledge, is challenging to extend to solve other tractable subclasses of POMDPs.

In the upcoming subsections, we will first extend the weakly revealing condition to address both undercomplete and overcomplete POMDPs. Subsequently, we will introduce a simple yet generic algorithm capable of sample-efficiently learning any POMDP that satisfies this generalized condition.

5.2 Observable Condition

To begin, we note that the weakly revealing condition is, in fact, *equivalent* to the following single-step observable condition [21] up to a multiplicative factor of $\sqrt{\max\{O, S\}}$ (see Section H.3 in [52]).

CONDITION 5.3 (single-step observable condition). There exists $\alpha > 0$ such that for any two state distributions $\mu, \nu \in \Delta_S$,

$$\min_{h\in[H]} \mathrm{TV}(\mathbb{O}_h\mu, \mathbb{O}_h\nu) \geq \alpha \mathrm{TV}(\mu, \nu).$$

The single-step observable condition mandates that distinct hidden state distributions induce distinct observation distributions. Intuitively, it stipulates that current-step observations should divulge a certain amount of information about the current hidden states. Consequently, it naturally excludes pathological instances used in previous works to establish hardness results, wherein observations provide no information about the hidden states. Importantly, Condition 5.1 does *not* imply that we can decode the hidden states from observations. For example,

$$\mathbb{O}_h = \begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{bmatrix}$$

satisfies Condition 5.1 with $\alpha = 1/3$, yet a learner can never perfectly infer the current hidden states from their observations and actions.

Unfortunately, due to its equivalence with the weakly revealing condition, the observable condition never holds in overcomplete POMDPs. To address this limitation, a natural extension is to require that observations over the next consecutive m steps jointly reveal some information about the hidden states, rather than relying on a single-step observation. To formalize this extension, we define the m-step observation-action matrices

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

as follows: For an observation sequence $\mathbf{o}$ of length m, initial state s and action sequence $\mathbf{a}$ of length $m-1$, we let $[\mathbb{G}_h]_{(\mathbf{a},\mathbf{o}),s}$ denote the probability of receiving $\mathbf{o}$ provided that the action sequence $\mathbf{a}$ is executed from state s and step h: for all $(\mathbf{a},\mathbf{o}) \in \mathscr{A}^{m-1} \times \mathscr{O}^m$ and $s \in \mathcal{S}$

$$\begin{aligned}[\mathbb{G}_h]_{(\mathbf{a},\mathbf{o}),s} &= \\ \mathbb{P}(o_{h:h+m-1} &= \mathbf{o} \mid s_h = s, a_{h:h+m-2} = \mathbf{a}).\end{aligned} \tag{4}$$

Analogous to the single-step case, the multi-step observable condition [52, 54] guarantees that the observable sequence over the next m consecutive steps provides sufficient information to distinguish any two state distributions, given a sufficiently large number of observations.

CONDITION 5.4 (multi-step observable condition). There exists $\alpha > 0$ and $m \in \mathbb{N}$ such that for any two state distributions $\mu, \nu \in \Delta_S$,

$$\min_{h\in[H-m+1]} \mathrm{TV}(\mathbb{G}_h\mu, \mathbb{G}_h\nu) \geq \alpha \mathrm{TV}(\mu, \nu),$$

where $\mathbb{G}_h$ is the m-step emission matrix defined in (4).

Note that Condition 5.3 is a special case of Condition 5.4 with $m = 1$. Additionally, for tabular POMDPs, Condition 5.3 is equivalent to assuming that $\min_h \sigma_S(\mathbb{G}_h) > \alpha'$ for some $\alpha' > 0$.

5.3 Algorithm: Optimistic MLE

The question now becomes how to efficiently learn multi-step observable POMDPs. To our knowledge, previous algorithms are tailored specifically to the undercomplete structure of POMDPs and cannot handle overcomplete POMDPs without introducing additional, potentially restrictive, technical assumptions. To address this challenge, Liu et al. [52] proposed a simple yet versatile algorithm called Optimistic Maximum Likelihood Estimation (Optimistic MLE), which can sample-efficiently learn (multi-step) observable POMDPs, as well as several other classes of structured POMDPs, in a unified manner. In this subsection, we present the Optimistic MLE algorithm and its guarantees for observable POMDPs. The results pertaining to learning other tractable classes of POMDPs can be found in Section 8.

To condense notations, we use $\theta = (\mathbb{T}, \mathbb{O}, \mu_1)$ to denote the model parameters of a POMDP and use Θ to denote the collections of all possible model parameters θ that correspond to POMDPs with S states, A actions, and O observations. Furthermore, to emphasize the dependence on θ, we will use $V^\pi(\theta)$ to denote the value of a policy π, and $\mathbb{P}^\pi_\theta(\tau)$ to denote the probability of observing a trajectory τ under policy π, when the underlying POMDP is parameterized by θ.

Algorithm description We provide the pseudocode of Optimistic MLE in Algorithm 1. As can be seen from this pseudocode, in each episode k there are two main steps:

- Optimistic planning (Lines 3-4): Find the POMDP model θ^k with the highest optimal value in the confidence set $\mathcal{B}^k$ and follow the associated optimal policy π^k to collect a trajectory τ^k. [2]
- Confidence set update (Line 5): Add the newly collected policy-trajectory pair into the data buffer. Then update the confidence set to include those models whose log-likelihood is "close" to the maximum possible one on the data collected thus far. Specifically, the confidence set is formulated as follows:

$$\Big\{\hat{\theta} \in \Theta : \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_{\hat{\theta}}(\tau) \geq \max_{\theta'\in\Theta} \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_{\theta'}(\tau) - \beta\Big\} \bigcap \mathcal{B}^1,$$

  where $\mathcal{B}^1$ is the initial confidence set that contains all α-observable POMDP models of a given size.

In contrast to the standard maximum likelihood estimation (MLE) approach, which retains only the model with the highest likelihood, the confidence set above encompasses all observable models with a sufficiently high likelihood. The size of this confidence set is controlled by the parameter $\beta \geq 0$. In particular, if $\beta = 0$, the confidence set collapses to the solutions of MLE. In [52], the selection of β is guided by the magnitude of the "statistical noise" introduced by various random events. By analyzing this noise, one can choose an appropriate value of β to ensure that the true POMDP model is consistently included in the resulting confidence set with high probability.

Theoretical Guarantees Liu et al. [52] prove that under a suitable choice of β, Optimistic MLE can learn any observable POMDP within a polynomial number of samples, as stated in the following theorem:

THEOREM 5.5 (Optimistic MLE). *Suppose Condition 5.4 holds and we choose $\beta = \tilde{\mathcal{O}}\left(H(S^2A + SO)\right)$ in Algorithm 1. Then Optimistic MLE learns an ε-optimal policy within*

$$\mathrm{poly}(S, A^m, O, H, \alpha^{-1}, \log(\varepsilon\delta)^{-1})/\varepsilon^2$$

samples with probability at least $1 - \delta$.

A natural question arises: is the exponential dependence on m in Theorem 5.5 necessary? Liu et al. [52] address this by providing an $A^{\Omega(m)}$ lower bound, which precludes the possibility of an upper bound that is polynomial in m.

Since the work of Liu et al. [52], the sample complexity bound of Optimistic MLE, as well as the lower bound for learning observable POMDPs, have been significantly improved through more refined analysis and novel constructions. Currently, the best upper [16] and lower [18] bounds are

$$\tilde{\mathcal{O}}\left(\frac{S^2OA^m(1 + SA/O)H^3}{\alpha^2\varepsilon^2}\right)$$

and

$$\Omega\left(\frac{(S^{3/2} \vee SA)O^{1/2}A^{m-1}H}{\alpha^2\varepsilon^2}\right)$$

respectively. Closing the gap between these bounds remains an open problem.

5.4 Proof sketch for Optimistic MLE

Below we sketch the proof for undercomplete POMDPs ($m = 1$). The proof for overcomplete POMDPs ($m > 1$) follows similarly with minor modification.

Step 1: bound the regret by the error of operator estimates By analyzing the relaxed MLE condition, one can prove that the groundtruth POMDP model $\theta^\star$ is contained in confidence set $\mathcal{B}^k$ for all $k \in [K]$ with high probability. Now, recall that we choose the model estimate and the behavior policy optimistically in Algorithm 1, i.e., $(\theta^k, \pi^k) = \operatorname{argmax}_{\hat{\theta}\in\mathcal{B}^k,\pi} V^\pi_{\hat{\theta}}$. As a result, we have

[2] For single-step observable POMDPs, Optimistic MLE directly samples τ^k from π^k, while for multi-step observable POMDPs τ^k is sampled from a composition of π^k and random actions.

Algorithm 1 Optimistic Maximum Likelihood Estimation (Optimistic MLE)

```
Initialize: $\mathcal{B}^1 = \{\hat\theta \in \Theta : \hat\theta$ satisfies Condition 5.4$\}$, $\mathcal{D} = \{\}$
for $k = 1, \ldots, K$ do
    compute $(\theta^k, \pi^k) = \mathrm{argmax}_{\hat\theta \in \mathcal{B}^k, \pi} V^\pi(\hat\theta)$
    execute policy $\pi^k$ to collect a trajectory $\tau^k := (o_1^k, a_1^k, \ldots, o_h^k, a_h^k)$
    add $(\pi^k, \tau^k)$ into $\mathcal{D}$ and update
    (5)    $\mathcal{B}^{k+1} = \left\{\hat\theta \in \Theta : \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_{\hat\theta}(\tau) \ge \max_{\theta' \in \Theta} \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_{\theta'}(\tau) - \beta\right\} \bigcap \mathcal{B}^1$
```

$V^\star = \max_\pi V^\pi(\theta^\star) \le \max_{\hat\theta\in\mathcal{B}^k,\pi} V^\pi(\hat\theta) = V^{\pi^k}(\theta^k)$ for all $k \in [K]$. From this, we get

$$\begin{aligned}\sum_{t=1}^{k} V^\star(\theta^\star) - V^{\pi^t}(\theta^\star) &\le \sum_{t=1}^{k} V^{\pi^t}(\theta^t) - V^{\pi^t}(\theta^\star)\\ &\le H \sum_{t=1}^{k}\sum_{\tau_H} |\mathbb{P}^{\pi^t}_{\theta^t}(\tau_H) - \mathbb{P}^{\pi^t}_{\theta^\star}(\tau_H)|.\end{aligned} \tag{6}$$

Therefore, to prove Theorem 1, it suffices to bound the cumulative error in estimating the probability of the individual trajectories, cf. the RHS of (6). By using the OOM representations in (3), it turns out that we can further bound the RHS of (6) by the error in estimating each observable operator:

$$\begin{aligned}&\frac{H\sqrt{S}}{\alpha} \sum_{t=1}^{k}\sum_{h=1}^{H}\sum_{\tau_h} \Big\| \left(\mathbf{B}_h(o_h,a_h) - \mathbf{B}^t_h(o_h,a_h)\right)\\ &\times \mathbf{b}(\tau_{h-1}) \Big\|_1 \times \pi^t(\tau_h) + \text{lower order terms}\end{aligned} \tag{7}$$

where $\mathbf{B}$ and $\mathbf{B}^t$ denotes the operator under $\theta^\star$ and θ^t respectively, and $\mathbf{b}(\tau_h) := \left(\prod_{h'=1}^{h} \mathbf{B}_{h'}(o_{h'}, a_{h'})\right) \mathbf{b}_0$ is the "belief vector" associated with trajectory $\tau_h = (o_1, a_1, \ldots, o_h, a_h)$.

Step 2: derive constraints for the operator estimates from Optimistic MLE As a result of the classic MLE analysis [e.g., 23], we have that with high probability: for all $(k,h) \in [K] \times [H]$

$$\sum_{t=1}^{k-1} \left(\sum_{\tau_h} \left| \mathbb{P}^{\pi^t}_{\theta^k}(\tau_h) - \mathbb{P}^{\pi^t}_{\theta^\star}(\tau_h) \right| \right)^2 = \mathcal{O}(\beta). \tag{8}$$

In brief, this means the model estimate in the k^{th} iteration, that is θ^k, can be used to predict the behavior of the policies followed *before* the k^{th} iteration to a certain accuracy. To proceed, we represent the probabilities in (8) by products of operators using (3) and perform further algebraic transformations, which eventually leads to the following error bound for operator estimation: for all $(k,h) \in [K] \times [H]$

$$\begin{aligned}&\sum_{t=1}^{k-1}\sum_{\tau_h} \left\| \left(\mathbf{B}_h(o_h,a_h) - \mathbf{B}^k_h(o_h,a_h)\right) \mathbf{b}(\tau_{h-1}) \right\|_1\\ &\times \pi^t(\tau_h) = \mathcal{O}\left(\frac{\sqrt{S\beta k}}{\alpha}\right).\end{aligned} \tag{9}$$

Intuitively, the constraints above imply the operator estimates in the k^{th} iteration are close to the true operators when being projected onto the belief vectors that are reweighted by the historical policies. However, a careful examination shows that (9) cannot be directly used to control (7), because (7) involves the operator error of θ^t reweighted by π^t that is the behavior policy in the *same* iteration, which is different from (9).

Step 3: bridge Step 1 and 2 via ℓ_1-norm eluder dimension By further algebraic transformations, we reduce the problem of bounding (7) by (9) to proving the following algebraic inequality.

Proposition 5.6. *Let $\{w_{k,j}\}_{(k,j)\in[K]\times[m]}$ and $\{x_{k,i}\}_{(k,i)\in[K]\times[n]}$ be d-dimensional vectors satisfying: for all $k \in [K]$*

$$\begin{cases} \sum_{t=1}^{k-1}\sum_{j=1}^{m}\sum_{i=1}^{n} |w_{k,j}^\top x_{t,i}| \le \sqrt{k} \\ \sum_{j=1}^{m} \|w_{k,j}\|_2 \le 1 \\ \sum_{i=1}^{n} \|x_{k,i}\|_2 \le 1. \end{cases}$$

Then we have for all $k \in [K]$

$$\sum_{t=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n} |w_{t,j}^\top x_{t,i}| = \tilde{\mathcal{O}}(d\sqrt{k}).$$

At a high level, the precondition and the target in Proposition 5.6 correspond to (9) and (7), respectively. While the inequality above resembles the classic elliptical potential lemma for linear bandits (e.g., see [46]), a naive application of existing techniques for proving the elliptical potential lemma results in a bound that scales suboptimally with k and linearly with n. Such a bound is useless, as n is equal to $(OA)^H$ in their proof. To overcome this challenge, Liu et al. [52] developed a new technique, termed ℓ_1-norm eluder dimension, which yields a bound exhibiting optimal scaling with k and logarithmic scaling with n.

5.5 Improved computational complexity

So far, we have identified a clean and mild condition—the observable condition—as sufficient for sample-efficient learning of POMDPs. We have also introduced a simple and generic algorithm, OMLE, for learning such POMDPs. However, all the results presented thus far pertain only to the statistical efficiency of learning POMDPs and exhibit exponential computational complexity in the worst case. In this subsection, we delve into the question of whether the observable condition confers any computational advantage for learning POMDPs.

Belief stability Golowich et al. [25] made the following intriguing observation for single-step observable POMDPs:

LEMMA 5.7. *Suppose Condition 5.3 holds. Consider two arbitrary belief states b_h, $b'_h \in \Delta_S$ at stage h. Suppose we draw $(o_{h:h+t}, a_{h:h+t})$ from the POMDP by following an arbitrary policy π starting from b_h at stage h. Let b_{h+t} denote the posterior distribution for state s_{h+t} under prior b_h and new observations $(o_{h:h+t}, a_{h:h+t})$. Similarly, let b'_{h+t} denote the counterpart under prior b'_h. Then*

$$\mathbb{E}\left[\left\|b_{h+t} - b'_{h+t}\right\|_1\right] \le \mathcal{O}\left((1-\alpha^4)^t S\right).$$

This result indicates that the dependence of the belief state at stage $h+t$ on that of stage h diminishes exponentially with the number of new observations received after stage h. Importantly, it implies that observations and actions significantly prior to the current stage have a negligible impact on the current belief state. Consequently, observations and actions from the $\tilde{\Theta}(\alpha^{-4})$ most recent stages provide a good approximation for the current belief state.

Short-memory planning and learning Leveraging the belief stability of observable POMDP, Golowich et al. [25] developed Algorithm 2 for planning with *known* model parameters. This algorithm approximates an observable POMDP with an MDP by considering the $\tilde{\Theta}(\alpha^{-4})$ most recent observations and actions in the POMDP as the state in the MDP.

THEOREM 5.8. *Suppose Condition 5.3 holds and choose $n = \Theta(\log(SH/\varepsilon)/\alpha^4)$. Algorithm 2 outputs an ε-optimal policy within $H(OA)^n$ time.*

Golowich et al. [25] also established a computational lower bound, demonstrating that under the Exponential Time Hypothesis, no algorithm can produce ε-suboptimal policies within time $(SAHO)^{o(\log(SAHO/\varepsilon)/\alpha)}$. Therefore, both the exponential dependence on α^{-1} and the quasi-polynomial dependence on other parameters are inherent to the computational complexity of planning in observable POMDPs. The lower bound also implies that the observable condition alone is insufficient for planning in POMDPs within polynomial time. An interesting future direction is to explore what additional structural conditions are necessary to achieve polynomial-time planning.

When the POMDP model is *unknown*, Golowich et al. [24] proposed an algorithm whose sample and computation complexity are both upper bounded by

$$(OA)^{\mathcal{O}(L)} \log(1/\delta)$$

where

$$L = \min\left\{\frac{\log(HSO/\varepsilon\alpha)}{\alpha^4}, \frac{\log^2(HSO/\varepsilon\alpha)}{\alpha^2}\right\}.$$

The above result scales exponentially with the reciprocal of the observable parameter α and quasi-polynomially with all other parameters. This is clearly suboptimal in terms of sample efficiency, as the previous Optimistic MLE algorithm achieves sample efficiency that is polynomial in all relevant parameters under the same observable condition.

6. OTHER TRACTABLE POMDPS

In this section, we discuss several other classes of tractable POMDPs that cannot be captured by the observable condition.

6.1 Latent MDPs

The Latent Markov Decision Process (Latent MDP [42]), also known as multitask RL [13] or multi-modal MDP [15], is a setting that falls between the MDP and POMDP paradigms.

Latent MDP A latent MDP is defined by a context distribution $\mathbf{w} \in \Delta_L$ and L MDPs that share the same state space $\mathscr{S}$ and action space $\mathscr{A}$. We denote the transition probability measure, the reward function and the initial state distribution of the l-th MDP by $\mathbb{T}_l$, R_l and μ_l respectively. The interaction protocol in a latent MDP is as follows: At the beginning of each episode, a latent context l is sampled from $\mathbf{w}$ without being revealed to the learner. Then the initial state s_1 is sampled from μ_l. At each stage h of this episode, the learner observes s_h, picks action a_h, and receives reward $r_{l,h}(s_h, a_h)$ in order. After that, the next state s_{h+1} is sampled from $\mathbb{T}_{l,h}(\cdot \mid s_h, a_h)$. We emphasize that the learner cannot directly observe the latent variable, and the latent variable is resampled only

Algorithm 2 SHORT-MEMORY PLANNING

```
Set $Q_{H+1} \equiv 0$. Denote $\nu(h) := \max\{1, h-n+1\}$, $z_h := (o_h, a_h)$ and
        $\mathbf{f}(z_{\nu(h):h}) := \mathbb{P}\left(o_{h+1} = \cdot \mid \text{prior } s_{\nu(h)} \sim \text{unif}(\mathscr{S}), \text{ observe } z_{\nu(h):h}\right)$
for $h = H, \ldots, 1$ do
    for $(z_{\nu(h):h}) \in (\mathscr{O} \times \mathscr{A})^{h+1-\nu(h)}$ do
        $Q_h(z_{\nu(h):h}) \leftarrow r_h(z_h) + \mathbb{E}_{o_{h+1} \sim \mathbf{f}(z_{\nu(h):h})}\left[\max_{a_{h+1}} Q_{h+1}(z_{\nu(h+1):h+1})\right]$
        $\pi_h(z_{\nu(h):h-1}, o_h) \leftarrow \operatorname{argmax}_{a_h} Q_h(z_{\nu(h):h})$
return $\pi$
```

at the beginning of each episode. The objective is to find a (potentially non-Markovian) policy that approximately maximizes the average cumulative reward.

It is easy to verify that a latent MDP with state space $\mathscr{S}$ can be formulated as a POMDP with hidden state $(l, s) \in [L] \times \mathscr{S}$ and observation space $\mathscr{S}$, if we view the combination of the latent context l and state s in the latent MDP as the hidden state in the POMDP. From this perspective, a latent MDP is a special POMDP where half of the hidden state (the MDP state) is always directly observable while the other half (the latent context) is never observable but remains fixed inside each episode.

Fundamental limits While latent MDPs are POMDPs with seemingly benign structures, they still exhibit similar computational and statistical hardness results to general POMDPs. Steimle et al. [70] prove that planning in latent MDPs with known model parameters is NP-hard in the worse case. Furthermore, Kwon et al. [42] construct a novel class of hard latent MDP instances which require at least $\Omega((SA)^L)$ samples to learn when the model is unknown. As a result, additional structural conditions are needed to render latent MDPs tractable.

Sample-efficient results Kwon et al. [42, 44] derived sample-efficient results for learning structured latent MDPs. First, Kwon et al. [42] prove that if the latent context is revealed to the learner after each episode during training, then an ε-optimal policy can be learned within $\tilde{\mathcal{O}}(LAH^4S^2/\varepsilon^2)$ samples. This result is a special case of a later work [47], which proved that any POMDP can be learned sample-efficiently without additional assumptions if the trajectory of hidden states is revealed at the end of each episode.

For the standard setting without context revealing, Kwon et al. [42] prove that if a latent MDP can be represented as an observable POMDP which further satisfies additional assumptions, then it can be sample-efficiently learned via tensor decomposition methods similar to those in [9, 28]. However, we note that these additional assumptions can be circumvented, as Theorem 5.5 demonstrates that the observable condition alone suffices for Optimistic MLE to sample-efficiently learn any POMDP, of which the latent MDP is a special case.

Finally, a recent insightful work by Kwon et al. [44] prove that if the number of latent contexts L is small (i.e. $L = \mathcal{O}(1)$), then latent MDPs can be sample-efficiently solved without *any* assumption. This result strictly generalizes previous work [41, 43] on learning reward-mixing MDPs, which are special cases of latent MDPs where different contexts share the same state transition function but have distinct reward functions. Note that this positive result does not contradict the previous $\Omega((SA)^L)$ lower bound as it considers the regime of $L = \mathcal{O}(1)$. The algorithm employed in [44] is a reward-free variant of Optimistic MLE [54].

6.2 Decodable POMDPs

Decodable POMDPs are those where the latent state can be precisely inferred from a short, recent history. Formally, they are defined by the following decodability condition [20].

CONDITION 6.1 (m-step decodability). There exists an *unknown* decoder $\phi^\star = \{\phi_h^\star\}_{h\in[H]}$ such that we have $s_h = \phi_h^\star(z_h)$ where $z_h = ((o, a)_{m(h):h-1}, o_h)$ with $m(h) = \max\{1, h-m+1\}$.

Decodable POMDP generalize the model of block MDP [19, 33], which satisfies the above condition with $m = 1$.

Relation to observable POMDP At first glance, decodable POMDP seems quite distinct from observable POMDP, which requires *future* observations and actions to reveal a certain amount of information about the current hidden state (Condition 5.4). It is also straightforward to prove by construction that neither class strictly contains the other (e.g., see Lemma D.4 in Liu et al. [54]). Nevertheless, these two settings are, in fact, deeply connected upon closer examination. From the perspective of history, both imply that the most recent history of a certain length acts as a good proxy state, by the definition of decodable POMDP and the belief stability property of observable

POMDP (Section 5.5). Furthermore, from the perspective of the future, we will show in Section 8.3 that these two settings, along with several others, can be unified by the framework of well-conditioned PSR, where the space of futures has certain low *linear* dimension.

Sample complexity One naive approach to learning decodable POMDP is to view the m-step recent observations and actions as the state and apply any existing MDP algorithms. The main caveat is that the resulted sample complexity scales linearly with the number of possible m-step recent histories, which is $(OA)^m$. This is a highly inefficient for settings with a large observation space. To address this issue, Efroni et al. [20] prove that if the learner is provided with a decoder class Φ containing the groundtruth decoder $\phi^\star$ a priori, then a modified GOLF algorithm [36] can learn an ε-optimal policy within

$$\tilde{O}\left(\frac{H^3A^mS}{\varepsilon^2}\cdot\log|\Phi|\right)$$

samples, which avoids the O^m dependence at the mild cost of scaling logarithmically with the cardinality of the decoder class. In particular, we always have $\log|\Phi| \leq \tilde{O}(S^2AH + SOH)$ for tabular decodable POMDPs. Efroni et al. [20] further provide lower bounds demonstrating that the A^m factor in the sample complexity is unavoidable in the worst case.

6.3 Linear Quadratic Gaussian

Linear quadratic Gaussian (LQG) [4] is one of the most fundamental problems in control theory. It involves a linear dynamic system partially driven by Gaussian noise and equipped with a quadratic loss function. Formally, an LQG system is defined as:

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{w}_t\\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{v}_t\end{aligned}\tag{10}$$

where $\mathbf{x}\in\mathbb{R}^{d_x}$ is the hidden state, $\mathbf{u}\in\mathbb{R}^{d_u}$ is the control input (action), $\mathbf{y}\in\mathbb{R}^{d_y}$ is the observation, $\mathbf{w}\sim\mathcal{N}(\mathbf{0},\Sigma_w)$ is the Gaussian noise in state transition and $\mathbf{v}\sim\mathcal{N}(\mathbf{0},\Sigma_v)$ is the Gaussian noise in observation. The initial state $\mathbf{x}_0$ is sampled from $\mathcal{N}(\mathbf{0},\Sigma_0)$. The standard objective in LQG is to identify a control strategy that selects $\mathbf{u}_t$ conditioned on $\mathbf{y}_{1:t}$ and $\mathbf{u}_{1:t-1}$ in order to minimize

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{y}_t^\top\mathbf{Q}\mathbf{y}_t + \mathbf{u}_t^\top\mathbf{R}\mathbf{u}_t\right).\tag{11}$$

Comparison to tabular POMDP Evidently, LQGs are POMDPs with infinitely many states, actions, and observations, which might initially appear intractable. However, LQGs are, in fact, considerably more tractable than tabular POMDPs due to their favorable linear structure and Gaussian randomness. Regarding computational complexity, the optimal control policy for an LQG is specified by the classical Kalman gain matrix [38] and the feedback gain matrix, both of which can be efficiently computed from the LQG parameters. This stands in stark contrast to the well-known difficulty of planning in tabular POMDPs. Furthermore, when the LQG parameters are unknown, they can be efficiently estimated by simply selecting *random* inputs (e.g., $\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$) to generate data and subsequently performing linear regression to derive estimates. In comparison, the naive random exploration strategy fails trivially even in tabular MDPs, the fully observable counterpart of POMDPs.

Non-asymptotic results for LQG There is a rich line of works [e.g., 45, 58, 62, 68, 71, 73, 82] on learning LQG with finite-sample guarantees. Most of them are based on estimating the so-called Markov parameters [56]

$$\mathbf{G} := [\mathbf{CB},\mathbf{CAB},\ldots,\mathbf{CA}^{m-1}\mathbf{B}]\in\mathbb{R}^{d_y\times md_u},$$

where m is a parameter to be chosen. Once a good estimate $\hat{\mathbf{G}}$ of $\mathbf{G}$ is obtained, there are two common ways to proceed: (1) Proper learning [e.g., 58]: Employ the Ho-Kalman algorithm [30, 63] to estimate $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ from $\hat{\mathbf{G}}$ up to a linear transformation, under standard assumptions in control theory. Subsequently, utilize these estimates to further estimate the Kalman gain matrix and the feedback gain matrix, which together specify the optimal control policy. (2) Improper learning [e.g., 45, 68]: Combine $\hat{\mathbf{G}}$ with an online learning algorithm to directly minimize the objective in (11) in an online fashion. This approach does not explicitly yield a time-invariant controller but can handle both stochastic and adversarial noise.

One widely adopted approach [62] to estimate $\mathbf{G}$ involves first taking purely random actions and then regressing $\mathbf{y}_t$ on $(\mathbf{u}_{t-1},\mathbf{u}_{t-1},\ldots,\mathbf{u}_{t-m})$, leveraging the following linear relationship derived from (10):

$$\begin{aligned}\mathbf{y}_t = \sum_{\tau=0}^{m-1}\mathbf{CA}^\tau\mathbf{B}\mathbf{u}_{t-\tau-1} + \sum_{\tau=0}^{m-1}\mathbf{CA}^\tau\mathbf{w}_{t-\tau-1}\\ + \mathbf{v}_t + \mathbf{CA}^m\mathbf{B}\mathbf{x}_{t-m},\end{aligned}$$

where the second and the third terms are zero-mean Gaussian noise independent of input control $\mathbf{u}$, and the final term vanishes under the choice of $m=\tilde{\Theta}(1/(1-\rho(\mathbf{A})))$ and the standard assumption that spectral radius $\rho(\mathbf{A})<1$. Crucially, this approach requires that the control inputs be non-vanishing and independent of observations. In scenarios where the control inputs are vanishing or dependent, Tsiamis and Pappas [73] and Lale et al. [45] proposed regressing $\mathbf{y}_t$ over both previous inputs $(\mathbf{u}_{t-1},\mathbf{u}_{t-1},\ldots,\mathbf{u}_{t-m})$ and previous observations $(\mathbf{y}_{t-1},\mathbf{y}_{t-1},\ldots,\mathbf{y}_{t-m})$.

7. MULTI-AGENT PARTIALLY OBSERVABLE RL

So far, we have focused solely on single-agent decision-making within the framework of POMDPs. In practice, however, many interesting tasks, such as games [14, 59, 67] and economic models [39], involve multiple decision-makers, each with their own reward function and observation of the environmental states. To model such tasks, [53] considered a general setting called the Partially Observable Markov Game (POMG), a multi-agent generalization of the POMDP.

7.1 Partially Observable Markov Game

A POMG comprises n agents. Each agent i possesses its own reward function $\{r_{h,i}\}_{h\in[H]}$, observation space $\mathscr{O}_i$ and action space $\mathscr{A}_i$. We denote the joint observation space by $\mathscr{O} := \prod_i \mathscr{O}_i$ and the joint action space by $\mathscr{A} = \prod_i \mathscr{A}_i$. In each stage $h \in [H]$, each agent i only observes her own observation $o_{h,i} \in \mathscr{O}_i$ where $o_{h,1}, \ldots, o_{h,n}$ are sampled from the *joint* observation distribution $\mathbb{O}(\cdot \mid s_h) \in \Delta_{|\mathscr{O}|}$. Subsequently, each agent i selects her own action $a_{h,i}$ and receives her own reward $r_{h,i}(o_{h,i}, a_{h,i})$. Then the environment transitions to the next state $s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_{h,1}, \ldots, a_{h,n})$. We emphasize that each agent can *not* observe other agents' observations and actions. As different agents have different reward functions, the learning objective is no longer to maximize any single reward function. Instead, the goal is to find certain types of equilibrium.

For clarity, we focus on learning Nash equilibrium in this section, although similar results can be obtained for other types of equilibria using analogous algorithm design and analysis.

DEFINITION 7.1 (Nash equilibrium). Let $\pi = \pi_1 \times \cdots \times \pi_n$ be a joint policy where the i-th agent follows policy π_i. We say π is an ε-approximate Nash if

$$\max_{i\in[n]} \left(V_i^{\pi} - \max_{\tilde{\pi}_i} V_i^{\tilde{\pi}_i \times \pi_{-i}} \right) \geq \varepsilon$$

where V_i^{π} denotes the i-th agent's expected total reward under policy π and $V_i^{\tilde{\pi}_i \times \pi_{-i}}$ denotes that under policy $\pi_1 \times \cdots \times \pi_{i-1} \times \tilde{\pi}_i \times \pi_{i+1} \times \cdots \times \pi_n$.

Intuitively, a joint policy is an ε-approximate Nash equilibrium if no agent can improve her expected return by more than ε through a unilateral change in her own policy.

Relation to other models We note that the POMG strictly generalizes several settings that have been extensively studied in prior works. First, it is a multi-agent extension of POMDP and a partially observable extension of Markov game [51] (also known as stochastic game [65]). Second, the decentralized POMDP model (dec-POMDP) [10] is a special case of POMG where all agents share the same reward function. Finally, imperfect-information extensive-form games (IIEFGs) [76] are POMGs with tree-structured state transition.

7.2 Observable POMG

Since POMG encompass POMDP, they naturally inherit all the hardness results proven for POMDP. Therefore, similar to POMDP, a natural question to ask for POMG is what structural conditions enable sample-efficient learning. Liu et al. [53] addressed this question by demonstrating that a game-theoretic variant of Optimistic MLE can sample-efficiently learn various types of equilibria in POMGs when the *joint* observation matrices satisfy Condition 5.3.

Their algorithm simply replaces the optimistic planning step (Line 3, Algorithm 1) for computing π^k with the following: Let Π_i denote the collection of all the deterministic policies for agent i, and $\Pi := \prod_{i\in[n]} \Pi_i$. Construct an n-player normal-form game where:

- Player i's strategy space is equal to Π_i.
- Player i's payoff under joint strategy $\pi \in \Pi$ is equal to $\max_{\theta\in\mathcal{B}^k} V_i^{\pi}(\theta)$.

Then choose π^k to be an Nash equilibrium of this normal-form game, which is a rank-1 mixture of deterministic policies in Π. In essence, we first compute the optimistic value estimate under each joint deterministic policy for each player, and then use these estimates to compute a Nash equilibrium.

THEOREM 7.2 (Observable POMG). *Suppose the joint observation matrices $\{\mathbb{O}_h\}_{h\in[H]}$ satisfy Condition 5.3. A game-theoretic variant of Optimistic MLE learns ε-approximate equilibria within*

$$\mathrm{poly}(S, \prod_i |\mathscr{A}_i|, \prod_i |\mathscr{O}_i|, H, \alpha^{-1}, \log \delta^{-1})/\varepsilon^2$$

samples with probability at least $1-\delta$.

Importantly, the above result only requires that all agents' observations *jointly* reveal some information about the hidden state, a condition considerably weaker than enforcing the observable condition on each agent's observations individually. Liu et al. [53] also derive similar results for learning other types of equilibria such as correlated equilibria and coarse correlated equilibria.

To demonstrate the generality of observable POMGs, Liu et al. [53] show that any IIEFG with perfect recall, a setting extensively studied in game theory literature, can be cast as an observable POMG of similar size with $\alpha = 1$. Finally, we remark that Liu et al. [53] also generalized the above result to multi-step observable POMGs where the multi-step joint observation matrix satisfies Condition 5.4.

8. LEARNING PREDICTIVE STATE REPRESENTATION

In preceding sections, we introduced two distinct structural conditions—the observable condition and the decodable condition—that facilitate tractable learning of POMDPs using Optimistic MLE and modified GOLF, respectively. In this section, we present a unifying framework—well-conditioned predictive state representations (well-conditioned PSRs)—which seamlessly integrates these two conditions. Moreover, well-conditioned PSRs also encompass many other interesting settings not captured by previous work, such as observable POMDPs with continuous observations or POMDPs with a few known core action sequences. Finally, we demonstrate that OMLE can solve any well-conditioned PSR in a unified manner, obviating the need of designing specialized algorithms for each distinct setting.

There are several concurrent works [16, 54, 80, 83] that investigate learning PSRs under similar structural conditions. For clarity and coherence, we will mostly follow the conventions in Liu et al. [54], and discuss its distinctions from the other works at the end of Section 8.2.

8.1 Predictive state representation

Consider the most general formulation of sequential decision making (SDM), where the system dynamics are specified by the conditional probability $\mathbb{P}_\theta(o_{1:H} \mid a_{1:H})$ with $\theta \in \Theta$ representing some *unknown* parameter.

System-dynamic matrices. We can fully specify an SDM by a collection of system-dynamic matrices $\{\mathbb{D}_h\}_{h\in[H]}$: For each fixed step h, we refer to an observation-action sequence in previous steps up to h, denoted as

$$\tau_h = (o_{1:h}, a_{1:h}),$$

as a **history**, and refer to an observation-action sequence in future steps, denoted as

$$\omega_h = (o_{h+1:m}, a_{h+1:m})$$

for any $m \in [h+1, H]$, as a **future** (or test). Let $\mathcal{T}_h$ represent the set of all possible histories at step h, and let Ω_h represent the set of all possible futures. We can then 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

$$[\mathbb{D}_h]_{\tau_h,\omega_h} := \mathbb{P}_\theta(o_{1:m}|a_{1:m}) \tag{12}$$

where $\tau_h = (o_{1:h}, a_{1:h})$ and $\omega_h = (o_{h+1:m}, a_{h+1:m})$.

Definition 8.1 (rank of SDM). The **rank** of an SDM is simply defined as $\max_{h\in[H]} \operatorname{rank}(\mathbb{D}_h)$ that is the maximal rank among the system-dynamic matrices $\{\mathbb{D}_h\}_{h\in[H]}$.

As an illustrative example, any POMDP with S hidden states is an SDM with rank at most S. Conversely, there exist low-rank SDMs that cannot be modeled by any POMDP with a state space of polynomial size [32].

PSR and core tests. Predicative State Representation (PSR) is proposed by [50, 69] as a generic approach to modeling 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 $d \geq r$, there always exist d columns (denoted as $\mathcal{Q}_h$) of matrix $\mathbb{D}_h$, such that the submatrix restricted to these columns $\mathbb{D}_h[\mathcal{Q}_h]$ satisfies $\operatorname{rank}(\mathbb{D}_h[\mathcal{Q}_h]) = r$. These d columns correspond to d futures $\mathcal{Q}_h = \{q_1, \ldots, q_d\}$, which are called **core tests**. We further define the set of **core action sequences** $\mathcal{Q}_h^{\mathrm{A}}$, which is the set of unique actions sequences within the set of core tests $\mathcal{Q}_h$. We know immediately that $|\mathcal{Q}_h^{\mathrm{A}}| \leq |\mathcal{Q}_h|$ and any rank-r system-dynamic matrix $\mathbb{D}_h$ admits at least one set of core action sequences with size $|\mathcal{Q}_h^{\mathrm{A}}| \leq r$.

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

$$\mathbb{D}_h = \mathbb{D}_h[\mathcal{Q}_h] \cdot \mathbf{M}_h^\top \tag{13}$$

where

$$\mathbb{D}_h[\mathcal{Q}_h] \in \mathbb{R}^{|\mathcal{T}_h|\times d}, \ \mathbf{M}_h \in \mathbb{R}^{|\Omega_h|\times d}$$

For any history τ_h and future ω_h, we denote the τ_h^{th} row of $\mathbb{D}_h[\mathcal{Q}_h]$ by $\boldsymbol{\psi}(\tau_h)$ and the ω_h^{th} row of $\mathbf{M}_h$ by $\mathbf{m}(\omega_h)$. Then Equation (13) can be equally written as

$$[\mathbb{D}_h]_{\tau_h,\omega_h} = \langle \boldsymbol{\psi}(\tau_h), \mathbf{m}(\omega_h)\rangle. \tag{14}$$

We provide a schematic illustration of the above decomposition in Figure 1. This implies an important property: for any history τ_h,

$$\boldsymbol{\psi}(\tau_h) := (\overline{\mathbb{P}}(\tau_h, q_1), \ldots, \overline{\mathbb{P}}(\tau_h, q_d))$$

serves as a sufficient statistics for the history τ_h in predicting the the probabilities of all futures conditioned on τ_h. Here for simplicity of notation, we are using $\overline{\mathbb{P}}(o_1, a_1, \ldots, o_h, a_h)$ to denote $\mathbb{P}(o_{1:h} \mid a_{1:h})$.

Following the convention in the PSR literature, we assume that all models parameterized by $\theta \in \Theta$ share at least one *common* set of core action sequences $\{\mathcal{Q}_h\}_{h\in[H-1]}$ that is *known* to the learner a priori. We will see that this assumption indeed holds in many settings of interest, such as observable POMDPs, decodable POMDPs, and others. Additionally, both mappings $\mathbf{m}$ and $\boldsymbol{\psi}$ are *unknown* to the learner.

[3] For clarity of presentation, we represent $\mathbb{D}_h$ as a matrix here, which requires $|\Omega_h|$ or $|\mathscr{O}|$ to be finite. However, this framework readily extends to settings with infinite observation spaces. See Appendix A in [54] for further details.

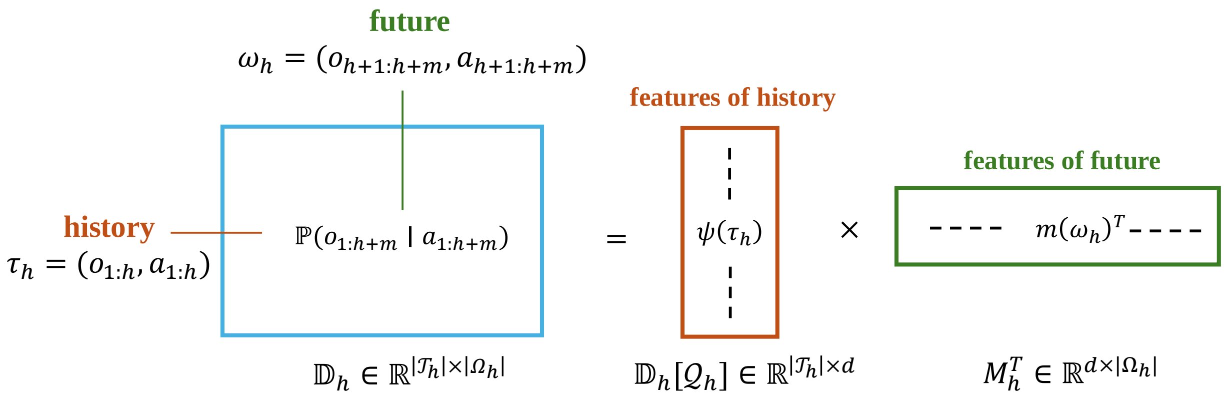
future
$\omega_h = (o_{h+1:h+m}, a_{h+1:h+m})$
features of history
features of future
history
$\tau_h = (o_{1:h}, a_{1:h})$
$\mathbb{P}(o_{1:h+m} \mid a_{1:h+m})$
$=$
$\psi(\tau_h)$
$\times$
$m(\omega_h)^T$
$\mathbb{D}_h \in \mathbb{R}^{|\mathcal{T}_h|\times|\Omega_h|}$
$\mathbb{D}_h[\mathcal{Q}_h] \in \mathbb{R}^{|\mathcal{T}_h|\times d}$
$M_h^T \in \mathbb{R}^{d\times|\Omega_h|}$

Fig 1: PSR decomposition of the system-dynamic matrix.

8.2 Well-conditioned PSR

Since PSR generalizes POMDP, it inherits all the hardness results associated with learning POMDP. Notably, Liu et al. [54] prove that even when the observation space, action space, and sets of core tests are all constant-sized, finding a near-optimal policy still requires an exponential number of samples in the worst case. At a high level, this hardness stems from the following: The definition of PSRs requires that for each stage h, the core tests $\mathcal{Q}_h$ satisfy $\text{rank}(\mathbb{D}_h[\mathcal{Q}_h]) = \text{rank}(\mathbb{D}_h) := r$. However, this requirement alone does not prevent the submatrix $\mathbb{D}_h[\mathcal{Q}_h]$ from being extremely close to a matrix of strictly lower rank. In other words, the matrix $\mathbb{D}_h[\mathcal{Q}_h]$ can be highly ill-conditioned. This ill-conditioning can lead to the linear weights $\mathbf{m}(\omega_h)$ in the decomposition $\mathbb{P}(\tau_h, \omega_h) = \mathbf{m}(\omega_h)^\top \boldsymbol{\psi}(\tau_h)$ becoming extremely large. Consequently, it results in a lack of robustness in predicting the probability $\mathbb{P}(\tau_h, \omega_h) = \mathbf{m}(\omega_h)^\top \boldsymbol{\psi}(\tau_h)$ when the vector $\boldsymbol{\psi}(\tau_h)$ needs to be estimated, as the estimation error of $\boldsymbol{\psi}(\tau_h)$ can be arbitrarily amplified by huge $\mathbf{m}(\omega_h)$.

To rule out such hard instances, core tests are required to not only guarantee $\text{rank}(\mathbb{D}_h[\mathcal{Q}_h]) := r$, but also ensure $\mathbb{D}_h[\mathcal{Q}_h]$ to be "well-conditioned" in certain sense. Liu et al. [54] enforce such a condition by assuming an upper bound on the magnitude of linear weight vectors.

CONDITION 8.2 (γ-well-conditioned PSR). 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 future linear weights $\mathbf{m}$[4] satisfy

$$(15) \qquad \max_{\substack{\mathbf{x}\in\mathbb{R}^{|\mathcal{Q}_h|} \\ \|\mathbf{x}\|_1 \le 1}} \sum_{\omega_h \in \Omega_h} \pi(\omega_h) \cdot |\mathbf{m}(\omega_h)^\top \mathbf{x}| \le \frac{1}{\gamma} .$$

Intuitively, the parameter γ^{-1} above quantifies the extent to which the future weight vectors $\{\mathbf{m}(\omega_h)\}_{\omega_h\in\Omega_h}$ can magnify 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, as otherwise, the previously mentioned hard instances would come into play. In Section 8.3, we will demonstrate that many common partially observable RL problems naturally fall into the category of γ-well-conditioned PSRs with a moderate value of γ, e.g., observable POMDPs and multistep decodable POMDPs.

THEOREM 8.3 (Liu et al. [54]). *A simple variant of* Optimistic MLE *can learn an ε-optimal policy for any γ-well conditioned PSR within*

$$\text{poly}(r, \gamma^{-1}, \max_h |\mathcal{Q}_h^A|, \log \mathcal{N}_\theta, A, H, \log(\delta^{-1}\varepsilon^{-1}))/\varepsilon^2$$

samples with probability at least $1-\delta$.

The result in Theorem 8.3 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, we highlight two important points:

- The sample complexity does *not* depend on the size of the core tests, but rather only on the size of the core action sequences.
- The sample complexity is completely *independent* of the size of the observation space.

Both empower the result to handle problems with *continuous* observations.

Finally, we note that several concurrent works [16, 80, 83] have derived similar results for learning PSRs. Among these, Chen et al. [16] obtained the sharpest sample complexity bound. Furthermore, Chen et al. [16]

[4] For each future ω_h, there generally exist infinitely many choices of $\mathbf{m}(\omega_h)$ that satisfies Equation (14). However, Liu et al. [54] only enforce Equation (15) for two specific, natural choices of $\mathbf{m}(\omega_h)$. This is a significantly weaker assumption than requiring it to hold for all possible choices of linear weights.

employed similar techniques used in the analysis of Optimistic MLE to demonstrate that well-conditioned PSRs can be incorporated into the decision-estimation coefficient framework [22], a general framework for model-based RL. However, none of these works provide efficient guarantees for learning observable POMDPs with continuous observations.

8.3 Examples of well-conditioned PSR

Liu et al. [54] demonstrate that both observable POMDPs and decodable POMDPs are well-conditioned PSRs with rank equal to the number of states, a mild well-conditioned parameter, and a moderate number of core action sequences.

PROPOSITION 8.4. *Any m-step α-observable POMDP can be represented as a well-conditioned PSR with*

- *rank $r = S$,*
- *core action sequences $\mathcal{Q}_h^A = \mathscr{A}^{\min\{m-1,H-h\}}$,*
- *and well-condioned parameter $\gamma = \mathcal{O}(\alpha/S)$.*

PROPOSITION 8.5. *Any m-step α-decodable POMDP can be represented as a well-conditioned PSR with*

- *rank $r = S$,*
- *core action sequences $\mathcal{Q}_h^A = \mathscr{A}^{\min\{m,H-h\}}$,*
- *and well-condioned parameter $\gamma = 1$.*

By combining Proposition 8.4 and 8.5 with Theorem 8.3, we immediately obtain polynomial sample-efficiency guarantees for learning observable POMDPs and decodable POMDPs under a unified framework via the same algorithm. Importantly, these results directly apply to POMDPs with continuous observations as long as the log-covering number of Θ is finite, e.g. POMDPs with Gaussian emission (Section 5.1.2 in [54]). In contrast, the previous result (Theorem 5.5) is inapplicable to POMDPs with infinitely many observations.

More examples Finally, we remark that Liu et al. [54] also introduce several other interesting POMDP classes that fall within the framework of well-conditioned PSRs. For instance, one such class is decodable POMDP with a large state space, where the number of hidden states in the sample complexity bound can be replaced by the rank of the transition matrices, which could be significantly smaller in certain settings [20]. Another interesting example is POMDP with a few core action sequences. In this setting, there exists a small set of *known* exploratory action sequences, potentially of different lengths, such that any two state mixtures can be distinguished from the observation distributions induced by at least one exploratory action sequence. This setting can be viewed as a refinement of m-step observable POMDPs when more prior knowledge is available, in the sense that the set of core action sequences $\mathcal{Q}_h$ contains only a few selected sequences instead of all those of length $m-1$. Consequently, we could have a more compact set of core action sequences with a size much smaller than A^{m-1}.

9. FUTURE DIRECTIONS

In this section, we outline several directions that we believe warrant further exploration in future research.

Polynomial-time algorithms None of the algorithms discussed thus far achieve polynomial computational complexity. Golowich et al. [25] established a lower bound demonstrating that learning observable POMDPs necessitates super-polynomial time in the worst case. It remains an open question as to what additional structural conditions are required to enable learning or planning in POMDPs with polynomial time complexity.

Multiagency Many real-world tasks involve both partial observability and multi-agency. Liu et al. [53] derived promising preliminary results for learning partially observable Markov games (Theorem 7.2). However, the sample complexity in their work scales polynomially with the size of the joint action and observation spaces, which, in turn, scales exponentially with the number of agents. This issue, known as the curse of multiagency [37], is problematic even for a moderate number of agents. An interesting question to investigate is for which subclass of POMGs we can circumvent this exponential dependence. For example, we believe that this is possible for fully cooperative POMGs, i.e., dec-POMDPs.

Model-free approach Most algorithms discussed in this survey are model-based. In practice, however, model-free approaches such as Q-learning and actor-critic algorithms are typically employed to solve partially observable tasks. It remains unclear how to develop simple and generic model-free methods for learning POMDPs, with accompanying theoretical guarantees. Moreover, it would be desirable if the designed methods could be readily combined with RNNs or Transformers to address partially observable problems of practical interest.

Core test discovery In the section on PSRs, we assumed that the core tests (or action sequences) are known to the learner a priori. However, this is not always the case in general. Therefore, an intriguing question in both theory and practice is how to automatically discover useful core tests from interactions in a data-driven manner. Additionally, how can we identify the minimal set of core action sequences that suffice for our learning purposes? We seek a compact set of core action sequences because the sample complexity for learning PSRs typically grows polynomially with the number of such sequences.

Minimax rate and instance-dependent bounds It remains an open problem to close the gap between the best upper bound [17] and the best lower bound [18] for learning observable POMDPs. Furthermore, all current results presented thus far are instance-independent. It would be valuable to derive more fine-grained, instance-dependent results for learning POMDPs and PSRs.

REFERENCES

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

[2] Awni Altabaa and Zhuoran Yang. On the role of information structure in reinforcement learning for partially-observable sequential teams and games. *arXiv preprint arXiv:2403.00993*, 2024.

[3] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

[4] Karl J Åström. *Introduction to stochastic control theory*. Courier Corporation, 2012.

[5] Karl Johan Åström. Optimal control of markov processes with incomplete state information i. *Journal of mathematical analysis and applications*, 10:174–205, 1965.

[6] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[7] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

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

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

[10] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4): 819–840, 2002.

[11] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

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

[13] Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. *arXiv preprint arXiv:1309.6821*, 2013.

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

[15] Iadine Chades, Josie Carwardine, Tara Martin, Samuel Nicol, Régis Sabbadin, and Olivier Buffet. Momdps: a solution for modelling adaptive management problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 267–273, 2012.

[16] Fan Chen, Yu Bai, and Song Mei. Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms. *arXiv preprint arXiv:2209.14990*, 2022.

[17] Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022.

[18] Fan Chen, Huan Wang, Caiming Xiong, Song Mei, and Yu Bai. Lower bounds for learning in revealing pomdps. In *International Conference on Machine Learning*, pages 5104–5161. PMLR, 2023.

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

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

[21] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. The value of observation for monitoring dynamic systems. In *IJCAI*, pages 2474–2479, 2007.

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

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

[24] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in observable pomdps, without computationally intractable oracles. *Advances in neural information processing systems*, 35: 1458–1473, 2022.

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

[26] Jiacheng Guo, Minshuo Chen, Huan Wang, Caiming Xiong, Mengdi Wang, and Yu Bai. Sample-efficient learning of pomdps with multiple observations in hindsight. *arXiv preprint arXiv:2307.02884*, 2023.

[27] Jiacheng Guo, Zihao Li, Huazheng Wang, Mengdi Wang, Zhuoran Yang, and Xuezhou Zhang. Provably efficient representation learning with tractable planning in low-rank pomdp. In *International Conference on Machine Learning*, pages 11967–11997. PMLR, 2023.

[28] Zhaohan Daniel Guo, Shayan Doroudi, and Emma Brunskill. A PAC RL algorithm for episodic POMDPs. In *Artificial Intelligence and Statistics*, pages 510–518. PMLR, 2016.

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

[30] BL Ho and Rudolf E Kálmán. Effective construction of linear state-variable models from input/output functions: Die konstruktion von linearen modeilen in der darstellung durch zustandsvariable aus den beziehungen für ein-und ausgangsgrößen. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

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

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

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

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

[35] Chi Jin, Sham M Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete POMDPs. *NeurIPS*, 2020.

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

[37] Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

[38] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[39] David M Kreps. *Game theory and economic modelling*. Oxford University Press, 1990.

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

[41] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Reinforcement learning in reward-mixing mdps. *Advances in Neural Information Processing Systems*, 34:2253–2264, 2021.

[42] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34: 24523–24534, 2021.

[43] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Reward-mixing mdps with few latent contexts are learnable. In *International Conference on Machine Learning*, pages 18057–18082. PMLR, 2023.

[44] Jeongyeol Kwon, Shie Mannor, Constantine Caramanis, and Yonathan Efroni. Rl in latent mdps is tractable: Online guarantees via off-policy evaluation. *arXiv preprint arXiv:2406.01389*, 2024.

[45] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.

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

[47] Jonathan Lee, Alekh Agarwal, Christoph Dann, and Tong Zhang. Learning in pomdps is sample-efficient with hindsight observability. In *International Conference on Machine Learning*, pages 18733–18773. PMLR, 2023.

[48] Sue E Leurgans, Robert T Ross, and Rebecca B Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993.

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

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

[51] Michael L Littman. Markov games as a framework for multiagent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[52] Qinghua Liu, Alan Chung, Csaba Szepesvari, and Chi Jin. When is partially observable reinforcement learning not scary? In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5175–5220. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/liu22f.html.

[53] Qinghua Liu, Csaba Szepesvári, and Chi Jin. Sample-efficient reinforcement learning of partially observable markov games. *Advances in Neural Information Processing Systems*, 35:18296–18308, 2022.

[54] Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. 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*, pages 363–376, 2023.

[55] Xiangyu Liu and Kaiqing Zhang. Partially observable multiagent rl with (quasi-) efficiency: the blessing of information sharing. In *International Conference on Machine Learning*, pages 22370–22419. PMLR, 2023.

[56] Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[57] Miao Lu, Yifei Min, Zhaoran Wang, and Zhuoran Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. *arXiv preprint arXiv:2205.13589*, 2022.

[58] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

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

[62] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

[63] Samet Oymak and Necmiye Ozay. Revisiting ho–kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4):1914–1928, 2021.

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

[65] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[66] Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *International Conference on Machine Learning*, pages 20057–20094. PMLR, 2022.

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

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

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

[70] Lauren N Steimle, David L Kaufman, and Brian T Denton. Multi-model markov decision processes. *IISE Transactions*, 53 (10):1124–1139, 2021.

[71] Yi Tian, Kaiqing Zhang, Russ Tedrake, and Suvrit Sra. Can direct latent model learning solve linear quadratic gaussian control? In *Learning for Dynamics and Control Conference*, pages 51–63. PMLR, 2023.

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

[73] Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

[74] Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Provably efficient reinforcement learning in partially observable dynamical systems. *Advances in Neural Information Processing Systems*, 35:578–592, 2022.

[75] Masatoshi Uehara, Haruka Kiyohara, Andrew Bennett, Victor Chernozhukov, Nan Jiang, Nathan Kallus, Chengchun Shi, and Wen Sun. Future-dependent value-based off-policy evaluation in pomdps. *Advances in Neural Information Processing Systems*, 36, 2024.

[76] J v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

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

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

[79] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Embed to control partially observed systems: Representation learning with provable sample efficiency. *arXiv preprint arXiv:2205.13476*, 2022.

[80] Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

[81] Yuheng Zhang and Nan Jiang. On the curses of future and history in future-dependent value functions for off-policy evaluation. *arXiv preprint arXiv:2402.14703*, 2024.

[82] Yang Zheng, Luca Furieri, Maryam Kamgarpour, and Na Li. Sample complexity of linear quadratic gaussian (lqg) control for output feedback systems. In *Learning for dynamics and control*, pages 559–570. PMLR, 2021.

[83] Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2022.