Designing Near-Optimal Partially Observable Reinforcement Learning

Ming Shi
Dept. of ECE
The Ohio State University
Columbus, OH
shi.1796@osu.edu

Yingbin Liang
Dept. of ECE
The Ohio State University
Columbus, OH
liang.889@osu.edu

Ness Shroff
Dept. of ECE and CSE
The Ohio State University
Columbus, OH
shroff.11@osu.edu

***Abstract*—Partially observable Markov decision processes (POMDPs) have been widely applied in various real-world applications. However, existing results have shown that learning in POMDPs is intractable in the worst case. The main challenge lies in the lack of latent state information. For example, in wireless channel scheduling, due to energy and security constraints, it is usually difficult or impossible for the user to know the conditions/states of all channels. Thus, a key fundamental question here is: how much online state information (OSI) is sufficient to achieve tractability? In this paper, we make the first effort to establish fundamental conditions and methods for bridging the gap between partially observable reinforcement learning and networking with incomplete state information. Specifically, we establish a lower bound that reveals a surprising hardness result: unless we have full OSI, we need an exponentially scaling sample complexity to obtain an ϵ-optimal policy solution for POMDPs. Nonetheless, motivated by the structures of practical systems, we identify important subclasses of POMDPs that are tractable, even with only partial OSI. For two subclasses of POMDPs with partial OSI, we provide new algorithms that are proved to be near-optimal by establishing new regret upper and lower bounds.**

***Index Terms*—reinforcement learning, partial observability, sample complexity, regret analysis, wireless channel scheduling**

I. Introduction

We investigate partially observable Markov decision processes (POMDPs) in reinforcement learning (RL) systems, where an agent interacts with the environment sequentially *without observing* the latent state (e.g., complete channel conditions). The goal is to achieve a large cumulative reward.

Consider wireless channel scheduling as an example [1]–[3]. At each time, a user must choose one of the channels for transmission. The conditions of these channels evolve along the time. However, due to energy and security constraints, the user may not know the complete conditions of all channels. Others examples include autonomous tanks that typically do not have a global view of traffic conditions due to limited reception [4], AI-trained robots that receive noisy observations of the battle field due to sensory noise [5], and so forth (Fig. 1).

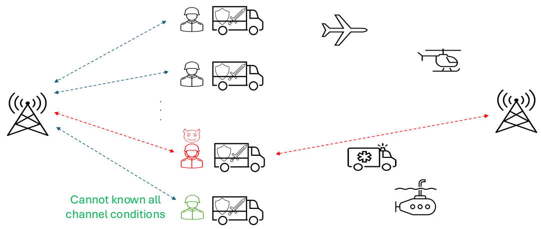


Fig. 1: Due to energy and security constraints, the user (e.g., the edge device used by the soldier) cannot know the conditions of all wireless channels during communication

Existing information-theoretical results have shown that learning in partially observable MDPs is intractable in the worst case [6] and PSPACE-complete [7]. This is in contrast to fully observable MDPs, where many efficient algorithms have been developed. The challenge of POMDPs lies in the lack of latent state information, such that the Markov property that simplifies fully observable MDPs does not hold any more.

To resolve the intractability issue, recent work has exploited hindsight state information [8], [9], where full state information becomes available at the end of each episode (i.e., after a certain number of time slots). This is motivated by the fact that, although the true latent state is unknown before the agent takes an action, some information may become available in hindsight. However, this assumption on *full* hindsight information may not hold, e.g., the conditions of the wireless channels that were not probed will still remain unknown to the user.

This issue motivates us to study *partial* state information based on query. We call it *partial "Online State Information" (OSI)*. For example, in wireless channel scheduling [1]–[3], partial OSI corresponds to the conditions of probed channels.

To model such partial OSI concretely, we consider vector-structured states [10]–[13]. Specifically, the state is given by a d-dimensional vector with each element representing a feature, such as the condition of one channel. Then, partial OSI means that at each step, a subset of $\tilde{d}$ ($1 \leq \tilde{d} < d$) elements in the state-vector will be revealed to the agent after her query, e.g., conditions of the $\tilde{d}$ channels that are probed by the user.

Our contributions: The key fundamental open question is that *with such partial OSI, can POMDPs be tractable?* In this paper, we provide in-depth answers to this open question.

(1) We establish a new lower bound in Theorem 1 that reveals a surprising hardness result: unless we have *full* OSI, we need an exponentially scaling sample complexity of $\tilde{\Omega}(\frac{A^H}{\epsilon^2})$ to find an ϵ-optimal policy for POMDPs, where A and H are the number of actions and the episode length, respectively. This

result indicates a sharp gap between POMDPs with *partial* OSI and POMDPs with *full* OSI (or full hindsight information). This may seem somewhat counter-intuitive, because by combining partial OSI from multiple steps (e.g., by querying each state-element one-by-one), one may construct full information of a state, and thus enjoy a similar tractability as that with full OSI. However, in Sec. III, we carefully design a worst-case instance, such that with only polynomial complexity, partial OSI at each step and even a combination of it from multiple steps are not sufficient to achieve an ϵ-optimal solution.

Then, one may ask: *Is there any subclass that is tractable with only partial OSI?* To push the boundary further along this axis, we identify two intriguing *tractable subclasses* of POMDPs with only partial OSI. They are motivated by structures of practical systems for networking and communications.

(2) In Sec. IV, we identify a tractable subclass with partial OSI, where additional partial noisy observations for the elements (e.g., channel conditions) in the state-vector that are not queried/probed are available after query. For this subclass, we provide a new near-optimal algorithm. Our algorithm design and regret analysis involve a non-trivial generalization of the observable operator method [14], [15] to handle *the non-trivial complexity structure of the partial noisy observations under the adaptively queried partial OSI*. Our result explains the fundamental value of channel condition estimation during transmission, e.g., utilizing past and side information.

(3) In Sec. V, we identify another tractable subclass of POMDPs with partial OSI, where the transitions of the state-elements are independent of each other. In the algorithm design, we use *adversarial importance weights* (for addressing in-episode biases), *heterogeneous decay parameters* (for addressing across-episode biases), and *query-based Q-value functions* (for addressing parameter-related inconsistent learning rates). Our result explains the fundamental value of probing capability in wireless communications. In addition, our theoretical analysis shows that the regret can be further reduced as the query/probing capability $\tilde{d}$ increases.

II. Problem Formulation

In this section, we introduce the problem formulation for POMDPs with partial online state information (OSI).

A. The Traditional Episodic POMDP

Episodic POMDPs (Fig. 2a) are usually modelled by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, H, \Delta_1, \mathbb{P}, \mathbb{O}, r)$, where $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{O}$ denote the state space with S states, the action space with A actions and the observation space with O observations, respectively; H denotes the number of steps in an episode; $\Delta_1 : \mathcal{S} \rightarrow [0,1]$ determines the randomness of the initial state at the beginning of an episode; $\mathbb{P} = \{\mathbb{P}_h : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0,1]\}_{h=1}^{H-1}$ and $\mathbb{O} = \{\mathbb{O}_h : \mathcal{O} \times \mathcal{S} \rightarrow [0,1]\}_{h=1}^{H}$ denote the *unknown* transition and emission probability measures, respectively; and $r = \{r_h : \mathcal{O} \times \mathcal{A} \rightarrow [0,1]\}_{h=1}^{H}$ denotes the reward function. Specifically, an online agent interacts with the environment in K episodes. At each step h of an episode k, the agent receives a global noisy observation o_h^k generated according to the emission probability $\mathbb{O}_h(\cdot|s_h^k)$, where s_h^k is the *unknown* true latent state. Next, the agent takes an action a_h^k and receives the reward $r_h(o_h^k, a_h^k)$. Then, the environment transits to the next state s_{h+1}^k, which is drawn according to the transition probability $\mathbb{P}_h(\cdot|s_h^k, a_h^k)$. The goal of the agent is to find a policy that achieves a high expected cumulative reward.

B. Partial OSI

Based on the traditional POMDP introduced above, we provide a concrete formulation for partial OSI. Specifically, we consider the vector-structured states [10], [11], [13], [16]. Each state s is represented by a d-dimensional feature vector $\vec{\phi}(s) = [\phi_1(s), ..., \phi_d(s)]^{\mathrm{T}} \in \tilde{\mathbb{S}}^d$, where $\tilde{\mathbb{S}}$ is the universal set of the values for each sub-state (i.e., element) in $\vec{\phi}(s)$, and $[\cdot]^{\mathrm{T}}$ denotes the transpose of a vector. We use $|\tilde{\mathbb{S}}|$ to denote the cardinality of the set $\tilde{\mathbb{S}}$. In each episode k, the agent interacts with the environment as follows: (Step-i) According to a *query policy* π_q^k, the agent actively queries $\tilde{d}$ sub-states of the unknown latent state s_h^k, where $1 \leq \tilde{d} < d$; (Step-ii) The queried sub-states $\{\phi_i(s_h^k)\}_{\{i \in \hat{i}_h^k\}}$, i.e., the *partial OSI*, are revealed to the agent, where $\hat{i}_h^k$ denotes the indices of the queried sub-states; (Step-iii) According to an *action policy* π_a^k, the agent chooses one sub-state $i_h^k \in \hat{i}_h^k$ and takes an action $a_h^k \in \mathcal{A}$; (Step-iv) The agent receives a *reward* $r_h(\phi_{i_h^k}(s_h^k), a_h^k)$, where $r_h : \tilde{\mathbb{S}} \times \mathcal{A} \rightarrow [0,1]$.

Motivating example: Consider wireless channel scheduling in military. Here, each *sub-state* $\phi_i(s)$ represents the condition, e.g., busy or idle, of one wireless channel. At each step, the user first actively probes the conditions of channels $\hat{i}_h^k$, and then observes the conditions of the sensed channels, i.e., the *partial OSI*. However, due to energy and security constraints, the agent cannot sense all the channels. Finally, she transfers packets using one sensed channel $i_h^k \in \hat{i}_h^k$, and receives a *reward* associated with this chosen channel i_h^k and *action* a_h^k.

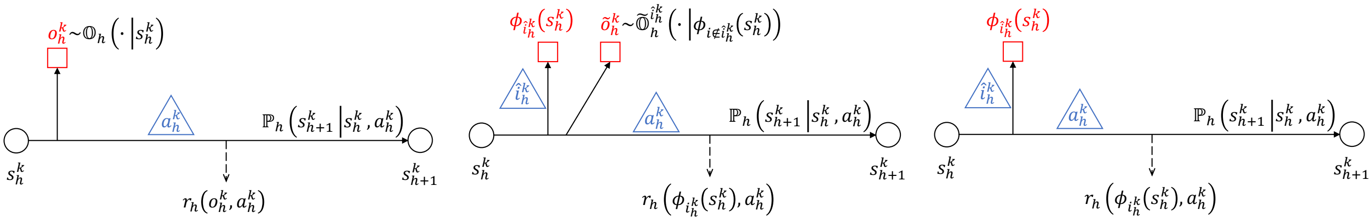
C. Performance Metric

In episode k, the agent queries sub-states according to a *query policy* π_q^k, e.g., it determines which wireless channels to probe. After receiving the partial OSI $\phi_{\hat{i}_h^k}(s_h^k)$ at each step, the agent takes an action according to an *action policy* π_a^k, e.g., it determines which wireless channel to use for communication. Moreover, before the partial OSI for step h is revealed in episode k, the feedback revealed to the agent is $\Phi_h^k = (\phi_{\hat{i}_1^k}(s_1^k), a_1^k, ..., \phi_{\hat{i}_{h-1}^k}(s_{h-1}^k), a_{h-1}^k) \in \hat{\Phi}_h$, where $\hat{\Phi}_h$ denotes the feedback space. After the partial OSI has been revealed, the feedback revealed is $\Phi_h^{k,'} = \{\Phi_h^k \cup \phi_{\hat{i}_h^k}(s_h^k)\} \in \hat{\Phi}_h^{'}$. We use the V-value $V^{\pi^k} \triangleq \mathbb{E}_{\{\pi_q^k, \pi_a^k, \mathbb{P}, \Delta_1\}}[\sum_{h=1}^{H} r_h(\phi_{i_h^k}(s_h^k), a_h^k)]$ to denote the expected total reward in episode k.

We take the regret as the performance metric, which is the difference between the expected cumulative reward of the online joint policies and that of the optimal policy, i.e.,

$$Reg^{\pi^{1:K}}(K) \triangleq \sum_{k=1}^{K} \left[V^{\pi^*} - V^{\pi^k}\right], \tag{1}$$

where $\pi^* \triangleq \arg\sup_{\{\pi_q, \pi_a\}} V^{\pi}$ denotes the optimal policy. *To the best of our knowledge, we are the first to provide near-optimal regrets for partially observable RL with partial OSI.*


$o_h^k \sim \mathbb{O}_h(\cdot | s_h^k)$
a_h^k
$\mathbb{P}_h(s_{h+1}^k | s_h^k, a_h^k)$
s_h^k
s_{h+1}^k
$r_h(o_h^k, a_h^k)$
$\phi_{\hat{i}_h^k}(s_h^k)$
$\tilde{o}_h^k \sim \tilde{\mathbb{O}}_h^{\hat{i}_h^k}(\cdot | \phi_{i \notin \hat{i}_h^k}(s_h^k))$
$\hat{i}_h^k$
a_h^k
$\mathbb{P}_h(s_{h+1}^k | s_h^k, a_h^k)$
s_h^k
s_{h+1}^k
$r_h(\phi_{\hat{i}_h^k}(s_h^k), a_h^k)$
$\phi_{\hat{i}_h^k}(s_h^k)$
$\hat{i}_h^k$
a_h^k
$\mathbb{P}_h(s_{h+1}^k | s_h^k, a_h^k)$
s_h^k
s_{h+1}^k
$r_h(\phi_{\hat{i}_h^k}(s_h^k), a_h^k)$

(a) Traditional POMDPs (b) With partial OSI and observations (c) With partial OSI and independency

Fig. 2: A sketch of one step: the squares represent the feedback, and the triangles represent the actions and/or queries

III. Perils of not Having Full OSI

In this section, we answer the long-standing open question: *whether POMDPs with online state information are tractable without full OSI (i.e., $\tilde{d} = d$)?* In Theorem 1 below, we establish a new lower bound that reveals a *surprising hardness result*: unless we have full OSI, we need an exponential sample complexity to find an ϵ-optimal policy for POMDPs, where a policy π is ϵ-optimal if $V^{\pi} \geq V^{\pi^*} - \epsilon$.

Theorem 1. ***(Intractability for not having full OSI)*** *For POMDPs with only partial online state information introduced in Sec. II-B, there exist hard instances, such that with a probability $p \geq 1/3$, any algorithm needs at least $\Omega(A^H/\epsilon^2)$ samples to find an ϵ-optimal policy.*

Theorem 1 demonstrates the hardness of POMDPs without full OSI: a polynomially scaling sample complexity Poly(A, H, S, K) is impossible for finding an ϵ-optimal policy. Theorem 1 may be counter-intuitive, because by combining partial OSI from multiple steps (e.g., querying each sub-state one-by-one), one may construct full information of a state, and thus enjoy similar tractability as that with full OSI. Below, we provide a worst-case instance and key proof ideas. Please see our technical report [17] for the complete proof.

A. Our Key Proof Ideas for Theorem 1

The important parts in our proof are to design special state representations and transitions, such that partial OSI cannot help the agent to improve her statistical knowledge about the true latent state. Towards this end, we construct a worst-case instance with four states, i.e., $s(1)$, $s(2)$, $s(3)$ and $s(4)$.

Idea I (Semi-correlated state representations): Our first idea is to construct the states, such that by observing $\tilde{d} = 1$ sub-state, it is impossible to infer the true state. Specifically, we let $\vec{\phi}(s(1)) = [x_1, x_2]^{\mathrm{T}}$, $\vec{\phi}(s(2)) = [x_3, x_4]^{\mathrm{T}}$, $\vec{\phi}(s(3)) = [x_1, x_4]^{\mathrm{T}}$ and $\vec{\phi}(s(4)) = [x_3, x_2]^{\mathrm{T}}$, where $x_1, ..., x_4$ are sub-states.

We now introduce the high-level idea for constructing such state representations. Let us consider states $s(1)$ and $s(2)$ as a group of states, and we call it group a. Similarly, we call states $s(3)$ and $s(4)$ group b. Thus, under our construction of the state representations, each state in group a (i.e., $s(1)$ and $s(2)$) must contain one and only one common sub-state as that in each state of group b (i.e., $s(3)$ and $s(4)$). This is why we call it "*semi-correlated*". For example, the first sub-states of both state $s(1)$ and state $s(3)$ are x_1. This means that, by only querying the first sub-state $i = 1$, the agent cannot know whether she is in a state from group a or group b.

However, whether a combination of partial OSI from multiple steps would be enough? To answer this question, we construct special state transitions using our idea II below.

Idea II (Closed-loop state transitions): Our second key idea is to construct closed-loop state transitions. Specifically, in each episode, the agent starts from state $s_1 = s(1)$. At step $h = 1$, (i) if action $a(1)$ is chosen, the state will transition to $s(1)$ and $s(2)$ with the same probability (wsp); (ii) if action $a(2)$ is chosen, the state will transition to $s(3)$ and $s(4)$ wsp. At step $h = 2$, (i) if $a(1)$ is chosen, both states $s(1)$ and $s(2)$ will transition to $s(3)$ and $s(4)$ wsp; (ii) if $a(2)$ is chosen, they will transition to $s(1)$ and $s(2)$ wsp. Similar for step $h = 4$.

Then, together with the semi-correlated state representations, even when the partial OSI about the first and second sub-states from multiple steps are combined, such a *closed-loop* state transition still prevents the agent from knowing which group of states she is in. For example, at step $h = 1$ of two episodes, the agent can keep taking action $a(1)$ and query the first and second sub-states one-by-one. Then, the partial OSI at step $h = 2$ could be $\phi_1(s_2^k) = x_1$ (i.e., the first sub-state of $s(1)$) and $\phi_2(s_2^{k+1}) = x_4$ (i.e., the second sub-state of $s(2)$). However, the first and second sub-states of $s(3)$ are also x_1 and x_4. Thus, such a combination of partial OSI (i.e., $\phi_1(s_2^k) = x_1$ and $\phi_2(s_2^{k+1}) = x_4$) is not enough for the agent to distinguish between visiting $(s(1), s(2))$ and visiting $s(3)$.

Idea III (Group-based reward functions): Up to here, only with partial OSI, the agent cannot improve her statistical knowledge. Thus, she can only rely on the statistical relation between the sequence of actions that is taken and the reward that is received. Hence, to create difficulties, (i) we let the rewards r_h at steps $h = 1, 2, 3$ be 0; (ii) if the final state is in *group* b, the reward at step $h = 4$ follows Bernoulli distribution with mean $\frac{1}{2}$; (iii) if the final state is in *group* a, the reward at step $h = 4$ follows Bernoulli distribution with a slightly higher mean equal to $\frac{1}{2} + \epsilon$. In this way, the optimal policy will take action sequence $(a(1), a(2), a(1))$ for all episodes, so that she can remain in group a and enjoy an expected total reward equal to $\frac{1}{2} + \epsilon$ in every episode. In contrast, the online agent has to try every sequence of actions to figure out which sequence provides larger reward with high probability. Since there are A^H action sequences, according to the Hoeffding's inequality, we can show that the sample complexity for achieving an ϵ-optimal policy is $\Omega(A^H/\epsilon^2)$.

Algorithm 1 Optimistic MLE with Partial OSI

```
Initialization: $\Theta^0 = \{\theta \in \Theta : \min_{\{h,\hat{i}\}} \sigma_{\tilde{S}}(\tilde{\mathbb{O}}_h^{\hat{i}}) \geq \alpha\}$.
for $k = 1 : K$ do
    Step-1: Estimate the problem model $\hat{\theta} \triangleq (\hat{\mathbb{P}}, \hat{\tilde{\mathbb{O}}}, \hat{\Delta}_1)$:
        $\Theta^k = \Theta^0 \cap \Big\{\hat{\theta} \in \Theta^0 : \sum_{\tau=1}^{k-1} \log P_{\hat{\theta}}^{\pi^\tau}(\Gamma^\tau) \geq \max_{(\mathbb{P}',\tilde{\mathbb{O}}',\Delta_1') \in \Theta^0} \sum_{\tau=1}^{k-1} \log P_{\mathbb{P}',\tilde{\mathbb{O}}',\Delta_1'}^{\pi^\tau}(\Gamma^\tau) - \beta\Big\}$.    (2)
    Step-2: Update the joint policy $\pi^k \triangleq \arg\max_{\pi:\hat{\theta}\in\Theta^k} \mathbb{E}_{\{\pi_q,\pi_a,\Delta_1,\hat{\theta}\}}[\sum_{h=1}^{H} r_h(\phi_{i_h^k}(s_h^k), a_h^k)]$.
    for $h = 1 : H$ do
        Step-3: Query the partial OSI $\phi_{\hat{i}_h^k}(s_h^k)$ according to the query policy $\pi_{q,h}^k$. Collect partial noisy observation $\tilde{o}_h^k$. Specify one sub-state $i_h^k$ and take an action $a_h^k$ according to the action policy $\pi_{a,h}^k$.
    end for
end for
```

IV. Tractability and Optimality under Partial OSI and Partial Noisy Observations

In the next two sections, we answer the key open question: *Is there any subclass that is tractable with partial OSI?* To this end, we identify two intriguing tractable subclasses under partial OSI motivated by channel estimation, and provide new near-optimal algorithms. The tractable subclass with flexible query capability that we study in this section is as follows.

Subclass 1. ***(POMDPs with partial OSI and partial noisy observations)*** *See Fig. 2b. At each step h of an episode k: (Step-i) the agent actively queries sub-states $\hat{i}_h^k$; (Step-ii) The partial OSI $\phi_{\hat{i}_h^k}(s_h^k)$ is revealed; (Step-iii) The agent receives the partial noisy observation $\tilde{o}_h^k$ for the other $d - \tilde{d}$ sub-states that are not queried, where $\tilde{o}_h^k$ is generated according to the partial emission probability $\tilde{\mathbb{O}}_h^{\hat{i}_h^k}\left(\cdot \middle| \{\phi_i(s_h^k)\}_{\{i \notin \hat{i}_h^k\}}\right)$. The partial emission matrix $\tilde{\mathbb{O}}_h^{\hat{i}} \in \mathbb{R}^{O \times |\tilde{\mathbb{S}}|^{d-\tilde{d}}}$ satisfies the* partially *revealing condition: there exists a constant $\alpha > 0$, such that $\sigma_{\tilde{S}}(\tilde{\mathbb{O}}_h^{\hat{i}}) \geq \alpha$ for any sub-states $\hat{i}$ and step h, where $\tilde{S} = |\tilde{\mathbb{S}}|^{d-\tilde{d}}$ and $\sigma_{\tilde{S}}(\cdot)$ denotes the $\tilde{S}$-th largest singular value of a matrix. Namely, $\min_{\{h,\hat{i}\}} \sigma_{\tilde{S}}(\tilde{\mathbb{O}}_h^{\hat{i}}) \geq \alpha$ holds; (Step-iv) The agent chooses a sub-state $i_h^k \in \hat{i}_h^k$, takes an action a_h^k, and receives a reward $r_h(\phi_{i_h^k}(s_h^k), a_h^k)$; (Step-v) The next state s_{h+1}^k is drawn according to the joint transition probability $\mathbb{P}_h(\cdot|s_h^k, a_h^k)$.*

We make two claims. (i) In contrast to standard POMDPs, the partial noisy observation $\tilde{o}_h^k$ in Subclass 1 depends on the query policy, whose outputs further affect the action policy. These two new dependencies require non-trivial developments in the algorithm design and regret analysis. (ii) In wireless channel scheduling, the partial noisy observation could be collected using condition estimation and prediction, e.g., by utilizing past and side information. Our results below show the fundamental value of such observations in communications.

A. Optimistic MLE with Partial OSI (OMLE-POSI)

We develop a near-optimal algorithm for Subclass 1, called Optimistic Maximum Likelihood Estimation with Partial OSI (OMLE-POSI). See Algorithm 1. The new challenge here is: how to design the query policy, such that the combination of partial OSI and partial noisy observations guarantee the existence of a near-optimal OMLE solution? To overcome this difficulty, our algorithm involves a non-trivial generalization of the standard observable operator method [14], [15].

Idea-I (Partial-information based bonus term): In contrast to the full noisy observation or full hindsight state information in standard POMDPs, only *partial* noisy observation is available in our case. To make it worse, it is affected by the adaptive query of the agent. Hence, when applying maximum likelihood estimation in Step-1 of Algorithm 1, we design *a new bonus term* $\beta = O\left((|\tilde{\mathbb{S}}|^{2d}A + |\tilde{\mathbb{S}}|^{d-\tilde{d}}O)\ln(|\tilde{\mathbb{S}}|^{d}AOHK)\right)$, which depends on the size of the non-queried sub-state space $|\tilde{\mathbb{S}}|^{d-\tilde{d}}$. $\Gamma^\tau \triangleq \{\phi_{\hat{i}_1^\tau}(s_1^\tau), \tilde{o}_1^\tau, a_1^\tau, ..., \phi_{\hat{i}_H^\tau}(s_H^\tau), \tilde{o}_H^\tau, a_H^\tau\}$ denotes the feedback that includes partial noisy observations $\tilde{o}_{1:H}^\tau$. This new bonus term captures the new complexity of partial-information space $\tilde{\mathbb{O}}_h^{\hat{i}}$, and satisfied the optimism-in-the-face-of-uncertainty principle [18] under incomplete information.

Idea-II (Bilevel query-and-action optimization): In contrast to standard POMDPs, the action policy $\pi_{a,h}$ here relies on the output of a query policy $\pi_{q,h}$. Thus, the query $\hat{i}_h^k$ and action a_h^k cannot be simply mapped to a single decision space. As a result, in the value iteration step (Step-2 of Algorithm 1), the reward maximization becomes a bilevel optimization problem.

Theorem 2. ***(Regret)*** *For POMDPs with the partial OSI and partially revealing condition, with probability $1-\delta$, the regret $Reg^{\text{OMLE-POSI}}(K)$ of OMLE-POSI can be upper-bounded by,*

$$\tilde{O}\left(|\tilde{\mathbb{S}}|^{2d-\tilde{d}}OAH^4\sqrt{K(|\tilde{\mathbb{S}}|^{2d}A + |\tilde{\mathbb{S}}|^{(d-\tilde{d})/2}O)/\alpha^2}\right).$$

Theorem 2 shows that (i) the regret depends polynomially on A, H and $|\tilde{\mathbb{S}}|$; (ii) The regret further decreases exponentially as $\tilde{d}$ increases; (iii) The regret of OMLE-POSI depends on $\sqrt{K}$, which is tight. *To the best of our knowledge, this is the first such near-optimal result for POMDPs with partial OSI.* Note that in our proof of Theorem 2 (in our technical report [17]), the main difficulty is the non-trivial complexity structure of partial noisy observations under adaptively queried partial OSI. Indeed, directly applying the observable operator method will result in a regret that does not decrease with $\tilde{d}$.

Note that the only parameter that the above regret does not have a polynomial dependency on is d. Below, we provide a lower bound, which shows the necessity of such a dependency.

Theorem 3. ***(Lower bound)*** *For POMDPs with partial OSI and partially revealing condition, the regret of any algorithm π can be lower-bounded as follows,*

$$Reg^{\pi}(K) \geq \tilde{\Omega}\left(\sqrt{AH} \cdot |\tilde{\mathbb{S}}|^{d/2} \cdot \sqrt{K}\right). \quad (3)$$

Our key proof idea (in [17]) is to construct a special state transition, such that even with partial OSI, all combinations of sub-states must be explored to achieve a sub-linear regret.

V. Tractability and Optimality under Partial OSI and Independent Sub-States

In this section, we discuss another tractable subclass with relatively restricted query capability in each episode.

Subclass 2. ***(POMDPs with partial OSI and independent sub-states)*** *See Fig. 2c. In each episode k: (Step-i) the agent actively queries sub-states $\hat{i}^k$ (not for each step h, i.e., relatively restricted query capability); (Step-ii) After the partial OSI $\phi_{\hat{i}^k}(s_h^k)$ is revealed at each step, the agent chooses a sub-state $i_h^k \in \hat{i}^k$, takes an action a_h^k, and then receives a reward $r_h(\phi_{i_h^k}(s_h^k), a_h^k)$; (Step-iii) The next state s_{h+1}^k is drawn according to the transition probability $\mathbb{P}_h(\cdot|s_h^k, a_h^k) = \prod_{i=1}^d \mathbb{P}_{h,i}(\phi_i(\cdot)|\phi_i(s_h^k), a_h^k)$, where the product form indicates that the sub-states have independent transition kernels.*

This subclass is motivated by practical applications, e.g., wireless channel scheduling. Due to probing energy and delay, the user may not *probe frequently*. Here, we do not assume additional partial noisy observations for non-queried sub-states. However, *without partial OSI in Step-ii of Subclass 2*, learning under independent sub-states could still be intractable.

Proposition 1. ***(Intractability for not having partial OSI)*** *There exist POMDPs with independent sub-states, such that learning an ϵ-optimal policy requires $\tilde{\Omega}(A^H/\epsilon^2)$ samples.*

A. Optimistic-In-Pessimistic-Out Learning (OIPOL)

We develop new near-optimal algorithms for Subclass 2. The new challenge here is: *how to query partial OSI to avoid intractability in Proposition 1 and achieve optimality?* To overcome this difficulty, our algorithms involve three interesting and important ideas. Due to page limits, we focus on introducing our new algorithm OIPOL for the case with $\tilde{d} > 1$ (see Algorithm 2). We use $mod(k, x)$ to denote the remainder when k is divided by x, and let $\kappa = \lceil (d-1)/(\tilde{d}-1) \rceil$.

Idea-I (Adversarial importance weights for addressing in-episode biases): Note that the query $\hat{i}$ could cause *errors* in V-values. Additionally, these errors could result in *non-stationary in-episode biases* for future decisions (although the state-transition and reward are stationary) [19]. Hence, in contrast to existing POMDP solutions that maintain a confidence set, a more conservative solution is required.

Step-1 and Step-2 of (2): at the beginning of every κ episodes, OIPOL updates the global weights and probabilities for each sub-state i according to a new exponential weighting:

$$w^k(i) = w^{k-\kappa}(i) \cdot e^{\frac{(d-1)\eta_1}{d(\tilde{d}-1)} \sum_{\tau=k-\kappa}^{k-1} \sum_{h=1}^{H} \hat{r}_h^\tau(\phi_i(s_h^\tau), a_h^\tau)},$$
$$\text{and } p^k(i) = (1-\eta_1) w^k(i) / \sum_{i'=1}^{d} w^k(i') + \eta_1/d, \qquad (4)$$

and then chooses a leading sub-state according to $p^k(i)$. We note that (i) η_1 is the first key decay parameter (see the second one η_2 in (5)). With a smaller η_1, the global weight increases more slowly, and thus the algorithm behaves more pessimistically. (ii) The estimated reward $\hat{r}_h^\tau(\phi_i(s_h^\tau), a_h^\tau)$ is $r_h^\tau(\phi_i(s_h^\tau), a_h^\tau) - r_h^\tau(\phi_{\tilde{i}^{\lfloor k/\kappa \rfloor}}(s_h^\tau), a_h^\tau)$ if $i \in \hat{i}^\tau$, and is 0 otherwise. Removing the *common* leading sub-state reward $r_h^\tau(\phi_{\tilde{i}^{\lfloor k/\kappa \rfloor}}(s_h^\tau), a_h^\tau)$ is the critical idea for eliminating the in-episode bias. (iii) The first term in $p^k(i)$ captures how important the sub-state $\phi_i(s)$ is, and the second term is a uniform distribution for exploiting different sub-states.

Algorithm 2 Optimistic-In-Pessimistic-Out Learning

```
for $k = 1 : K$ do
    if $mod(k, \kappa) = 1$ then
        Step-1: Update the global weights $w^k(i)$ and probabilities $p^k(i)$ according to (4).
        Step-2: Choose a leading sub-state $\tilde{i}^{\lceil k/\kappa \rceil}$, i.e., the leader, according to the global probability $p^k(i)$.
        Step-3: Initialize the local weight $\tilde{w}^k(i)$ according to the global weight $w^k(i)$, i.e., $\tilde{w}^k(i) = w^k(i)$.
    end if
    Step-4: Choose $\tilde{d} - 1$ supporting sub-states, i.e., the follower, uniformly randomly from the sub-states that have not yet been chosen in most-recent $\kappa$ episodes, i.e., from $\lfloor \frac{k-1}{\kappa} \rfloor \cdot \kappa + 1$ to $(\lfloor \frac{k-1}{\kappa} \rfloor + 1) \cdot \kappa$.
    Step 5: According to (5), update the local weights $\tilde{w}^k(i)$ and probabilities $\tilde{p}^k(i)$ for sub-state $i$ queried.
    Step-6: Choose the rewarding sub-state $i_h^k$ according to the updated local probability $\tilde{p}^k(i)$.
    for $h = H : 1$ do
        Step-7: Update $Q$-values according to (6).
    end for
    for $h = 1 : H$ do
        Step-8: Take an action $a_h^k$ that maximizes the updated $Q$-value function, and collect the partial OSI.
    end for
end for
```

Idea-II (Heterogeneous decay parameters for addressing across-episode biases): Note that sub-optimal queries $\hat{i}$ at the beginning of episodes result in *unavoidable across-episode biases* for choosing rewarding actions at each step. Hence, in contrast to gradient descents that use homogeneous learning rates, a heterogeneous solution is required.

Step-5 and Step-6 of Algorithm 2: At the beginning of episode k, OIPOL updates the local weights $\tilde{w}^k(i)$ and probabilities $\tilde{p}^k(i)$ for sub-states $\phi_i(s)$ in query set $\hat{i}^k$ (formed by the leading sub-state $\tilde{i}^k$ and $\tilde{d}-1$ supporting sub-states):

$$\tilde{w}^k(i) = \tilde{w}^{k-1}(i) \cdot e^{\frac{\eta_2}{\tilde{d}} \sum_{h=1}^{H} r_h^{k-1}(\phi_i(s_h^{k-1}), a_h^{k-1})}, \text{ and}$$
$$\tilde{p}^k(i) = (1-\eta_2)\tilde{w}^k(i) / \sum_{i' \in \hat{i}^k} \tilde{w}^k(i') + \eta_2/\tilde{d}, \qquad (5)$$

and chooses the rewarding sub-state i_h^k according to $\tilde{p}^k(i)$. We note that (i) to make the algorithm optimistic enough *in* the episode, the value of the *decay parameter* η_2 should be larger than the value of η_1. Theorem 4 below provides a sufficient condition on how much η_2 should be larger than η_1. (ii) The factor $(d-1)/(\tilde{d}-1)$ in (4) does not appear in (5), because the local weight is updated for the sub-states in $\hat{i}^k$. (iii) The denominator in the first term of $\tilde{p}^k(i)$ only includes $i \in \hat{i}^k$.

Idea-III (Query-based Q-value functions for addressing parameter-related inconsistent learning rates): The remaining question is how to get the correct factor (i.e., $(d-1)\eta_1/d(\tilde{d}-1)$) for $w^k(i)$? We choose $\tilde{d}-1$ supporting sub-states uniformly randomly from the sub-states that have not yet been queried in most-recent episodes (i.e., Step-3 and Step-4 in Algorithm 2). Then, conditioned on the leading sub-state, each sub-state is chosen with probability $\frac{\tilde{d}-1}{d-1}$, which results in the factor $\eta_1/(d\frac{\tilde{d}-1}{d-1}) = \frac{(d-1)\eta_1}{d(\tilde{d}-1)}$.

Step-7 and Step-8 of Algorithm 2: In order to address the inconsistent-learning-rate issue due to heterogeneous decay parameters, we construct query-based Q-value functions that follow an *optimism-in-face-of-partial-OSI* principle,

$$\begin{aligned} Q_h^k(\phi_i(s), a) = \min\{ & r_h(\phi_i(s), a) + [\mathbb{P}_h^k V_{h+1}^k](\phi_i(s), a) \\ & + O(\sqrt{H^2/\mathcal{N}_h^k(\phi_i(s), a)}), H\}, \text{ for all } i \in \hat{i}^k, \end{aligned} \tag{6}$$

where $\mathbb{P}_h^k(\phi_i(s')|\phi_i(s), a) = \frac{\mathcal{N}_h^k(\phi_i(s), a, \phi_i(s'))}{\mathcal{N}_h^k(\phi_i(s), a)}$ is the estimated transition kernel, $\mathcal{N}_h^k(\phi_i(s), a)$ and $\mathcal{N}_h^k(\phi_i(s), a, \phi_i(s'))$ are the number of times $(\phi_i(s), a)$ and $(\phi_i(s), a, \phi_{\hat{i}}(s'))$ have been visited at step h up to episode k, respectively, and $V_h^k(\phi_i(s)) = \max_a Q_h^k(\phi_i(s), a)$. Finally, OIPOL takes an action to maximize $Q_h^k(\phi_i(s), a)$.

Theorem 4. ***(Regret)*** *For POMDPs with partial online state information and independent sub-states, by choosing $\eta_1 = \tilde{O}(1/\sqrt{K})$ and $\eta_2 = \frac{16(d-1)}{\tilde{d}-1}\eta_1$, with probability $1-\delta$, the regret $Reg^{OIPOL}(K)$ can be upper-bounded by*

$$\tilde{O}\left(H^{\frac{5}{2}} |\tilde{\mathbb{S}}|^2 A \sqrt{\frac{dK \ln d}{\tilde{d}-1}} \left(\ln \frac{H^2 |\tilde{\mathbb{S}}| AK}{\delta} \right)^2 \right). \tag{7}$$

Theorem 4 shows that (i) the regret of OIPOL depends polynomially on all problem parameters; (ii) The regret of OIPOL decreases further as the query capability $\tilde{d}$ increases; (iii) The dependency on K is $\tilde{O}(\sqrt{K})$, which is tight. *To the best of our knowledge, this is the first such near-optimal result for POMDPs with partial OSI.* Further, our regret analysis (in [17]) includes new technical developments to handle the correlations (i) between the action and query policies, and (ii) between the in-episode and across-episode biases.

VI. Conclusion

In this paper, we establish a lower bound that reveals a *surprising hardness* result: unless we have full OSI, we need an exponentially scaling sample complexity to obtain an ϵ-optimal policy for POMDPs. Nonetheless, motivated by practical wireless communications, we identify two intriguing tractable subclasses of POMDPs with only *partial* OSI, e.g., probed channel conditions. We provide new RL algorithms, which are proved to be near-optimal by establishing regret upper and lower bounds. Our solutions resolve the open problem when applying RL to real-world networking, and could serve as a key foundation for future work in this direction.

Acknowledgment

This work has been supported in part by NSF grants NSF AI Institute (AI-EDGE) CNS-2112471, CNS-2312836, CNS-2106933, CNS-2106932, CNS-2312836, CNS-1955535, CNS-1901057, and RINGS-2148253, by Army Research Office under Grant W911NF-21-1-0244 and was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0225. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc networks: A POMDP Framework," *IEEE Journal on selected areas in communications*, vol. 25, no. 3, pp. 589–600, 2007.

[2] Y. Chen, Q. Zhao, and A. Swami, "Joint Design and Separation Principle for Opportunistic Spectrum Access in the Presence of Sensing Errors," *IEEE Transactions on Information Theory*, vol. 54, no. 5, 2008.

[3] W. Ouyang, S. Murugesan, A. Eryilmaz, and N. B. Shroff, "Exploiting Channel Memory for Joint Estimation and Scheduling in Downlink Networks—A Whittle's Indexability Analysis," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1702–1719, 2015.

[4] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held *et al.*, "Towards Fully Autonomous Driving: Systems and Algorithms," in *2011 IEEE intelligent vehicles symposium (IV)*. IEEE, 2011, pp. 163–168.

[5] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell *et al.*, "Solving Rubik's Cube with a Robot Hand," *arXiv preprint arXiv:1910.07113*, 2019.

[6] A. Krishnamurthy, A. Agarwal, and J. Langford, "PAC Reinforcement Learning with Rich Observations," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[7] C. Papadimitriou and J. Tsitsiklis, "The Complexity of Markov Decision Processes," *Mathematics of operations research*, vol. 12, no. 3, 1987.

[8] S. R. Sinclair, F. V. Frujeri, C.-A. Cheng, L. Marshall, H. D. O. Barbalho, J. Li, J. Neville, I. Menache, and A. Swaminathan, "Hindsight Learning for MDPs with Exogenous Inputs," in *International Conference on Machine Learning*. PMLR, 2023, pp. 31 877–31 914.

[9] J. Lee, A. Agarwal, C. Dann, and T. Zhang, "Learning in POMDPs is Sample-Efficient with Hindsight Observability," in *International Conference on Machine Learning*. PMLR, 2023, pp. 18 733–18 773.

[10] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably Efficient Reinforcement Learning with Linear Function Approximation," in *Conference on Learning Theory*. PMLR, 2020, pp. 2137–2143.

[11] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement Learning: Theory and Algorithms," *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pp. 10–4, 2019.

[12] M. Shi, Y. Liang, and N. Shroff, "A Near-Optimal Algorithm for Safe Reinforcement Learning Under Instantaneous Hard Constraints," in *International Conference on Machine Learning*. PMLR, 2023, pp. 31 243–31 268.

[13] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang, "Model-Based RL with Value-Targeted Regression," in *International Conference on Machine Learning*. PMLR, 2020, pp. 463–474.

[14] Q. Liu, A. Chung, C. Szepesvári, and C. Jin, "When Is Partially Observable Reinforcement Learning Not Scary?" in *Conference on Learning Theory*. PMLR, 2022, pp. 5175–5220.

[15] H. Jaeger, "Observable Operator Models for Discrete Stochastic Time Series," *Neural computation*, vol. 12, no. 6, pp. 1371–1398, 2000.

[16] M. Shi, X. Lin, and S. Fahmy, "Competitive Online Convex Optimization with Switching Costs and Ramp Constraints," *IEEE/ACM Transactions on Networking*, vol. 29, no. 2, pp. 876–889, 2021.

[17] M. Shi, Y. Liang, and N. Shroff, "Theoretical Hardness and Tractability of POMDPs in RL with Partial Online State Information," *arXiv preprint arXiv:2306.08762*, 2023.

[18] M. G. Azar, I. Osband, and R. Munos, "Minimax Regret Bounds for Reinforcement Learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 263–272.

[19] M. Shi, X. Lin, and L. Jiao, "Power-of-2-Arms for Bandit Learning with Switching Costs," in *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2022, pp. 131–140.