Achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ Regret in Average-Reward POMDPs with Known Observation Models

Alessio Russo
Politecnico di Milano

Alberto Maria Metelli
Politecnico di Milano

Marcello Restelli
Politecnico di Milano

Abstract

We tackle average-reward infinite-horizon POMDPs with an unknown transition model but a known observation model, a setting that has been previously addressed in two limiting ways: (i) frequentist methods relying on suboptimal stochastic policies having a minimum probability of choosing each action, and (ii) Bayesian approaches employing the optimal policy class but requiring strong assumptions about the consistency of employed estimators. Our work removes these limitations by proving convenient estimation guarantees for the transition model and introducing an optimistic algorithm that leverages the optimal class of deterministic belief-based policies. We introduce modifications to existing estimation techniques providing theoretical guarantees separately for each estimated action transition matrix. Unlike existing estimation methods that are unable to use samples from different policies, we present a novel and simple estimator that overcomes this barrier. This new data-efficient technique, combined with the proposed *Action-wise OAS-UCRL* algorithm and a tighter theoretical analysis, leads to the first approach enjoying a regret guarantee of order $\mathcal{O}(\sqrt{T \log T})$ when compared against the optimal policy, thus improving over state of the art techniques. Finally, theoretical results are validated through numerical simulations showing the efficacy of our method against baseline methods.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

Reinforcement Learning (RL) (Sutton and Barto, 2018) tackles the sequential decision-making problem of an agent interacting with an unknown or partially known environment with the goal of maximizing the long-term sum of rewards. The RL agent should trade-off between *exploring* the environment to learn its structure and *exploiting* the estimates to compute a policy that maximizes the reward. This problem has been successfully addressed in past works under the MDP formulation (Bartlett and Tewari, 2009; Jaksch et al., 2010; Zanette and Brunskill, 2019). MDPs assume full observability of the state space but this assumption is often violated in many real-world scenarios such as robotics or finance, where only a partial observation of the environment is available. In this case, it is more appropriate to model the problem using Partially-Observable MDPs (Sondik, 1978).

Further challenges emerge when tackling POMDPs since (i) the estimation problem turns into identifying the latent parameters of the model, (ii) the planning problem is known to be computationally intractable even for known models (Mossel and Roch, 2005).
Different approaches have been devised to tackle the estimation problem (Guo et al., 2016; Xiong et al., 2022). In the finite horizon setting, Jin et al. (2020) present a sample-efficient algorithm for the undercomplete POMDP setting, where the number of observations is larger than the number of states, while in the average-reward setting, Azizzadenesheli et al. (2016) provide guarantees on the regret by introducing a model-based approach that leverages *spectral decomposition* (Anandkumar et al., 2014) techniques to estimate the latent model while employing memoryless policies.

This paper considers an average-reward POMDP setting with a known observation model but an unknown transition model that the agent needs to learn. The assumption of having partial knowledge of the environment has been variously addressed in the past, both in the bandit setting (Maillard and Mannor, 2014;

Russo et al., 2024b) and in the POMDP setting (Jafarnia Jahromi et al., 2022; Russo et al., 2024a). Relying on the knowledge of the observation model can be of interest in many real-world scenarios. Sometimes this knowledge is available from scratch. For example, in robotics, the properties of the sensors of a robot are available beforehand, while in other scenarios the observation model can be learned offline from simulation (Thananjeyan et al., 2021; Marco et al., 2017) or from historical data, while the transition model of the environment needs to be learned in an online fashion. This assumption is also reasonable in the case of non-stationary environments where the change is local and associated only with the transition model while the previous knowledge of the observation model can be retained. Similar motivations hold as well for problems dealing with *Transfer Learning*.

Contribution We report here the main contributions of our work:

- We present the *Action-wise* OAS procedure that estimates the transition model under the assumption of knowing the observation model. We show that this technique also handles samples collected under different policies.
- Under some technical assumptions, we prove estimation guarantees separately for the transition matrix associated with each action.
- We introduce the *Action-wise* OAS-UCRL algorithm, an optimistic approach that makes use of the presented estimation method and employs the optimal class of deterministic belief-based policies.
- By using new theoretical results, we prove that this algorithm enjoys a $\mathcal{O}(\sqrt{T \log T})$ regret guarantee when compared against the optimal POMDP policy, thus improving over state-of-the-art results.

2 RELATED WORKS

In recent years, significant progress has been made in understanding the fully observable RL setting. Some studies focused on the episodic scenario with finite horizon (Jin et al., 2018; Azar et al., 2017), while others have examined the non-episodic undiscounted setting (Jaksch et al., 2010; Ortner and Ryabko, 2012; Bartlett and Tewari, 2009). In contrast, Partially Observable MDPs have received relatively less attention, also due to the inherent difficulty of this setting both from the learning and the planning perspective.

Learning in POMDPs A POMDP instance is considered hard when the observation model does not contain enough information to allow learning the underlying transition model. These pathological cases can be avoided by assuming that the observation model is full-rank or, stated differently, when the minimum singular value α of the observation model is positive, namely $\alpha > 0$. Instances belonging to this class are called **α-weakly revealing** and can be learned efficiently.

Within this class of tractable problems, some works focused on the **episodic** setting such as Jin et al. (2020) and Liu et al. (2022a). The first one considers the undercomplete case, with the number of states less or equal to the number of observations ($S \leqslant 0$): the authors do not focus on regret but introduce an algorithm with optimal sample complexity ($1/\epsilon^2$) for learning an ϵ-optimal policy. Differently, Liu et al. (2022a) provide an approach based on a *Maximum Likelihood Estimation* technique and show regret guarantees of order $\widetilde{\mathcal{O}}(\sqrt{T})$ for the undercomplete case. They also consider the more difficult overcomplete setting ($S > O$) for which they prove a guarantee of order $\widetilde{\mathcal{O}}(T^{2/3})$. In the subsequent work of Chen et al. (2023), these regret guarantees were shown to be tight with respect to the horizon T in both the undercomplete and overcomplete settings.

A second stream of works focuses on the **non-episodic average reward** setting. In particular, Azizzadenesheli et al. (2016) and Xiong et al. (2022) consider the standard POMDP setting where neither the observation nor the transition model are known and they both employ *Spectral Decomposition* (SD) techniques to retrieve the model parameters. Azizzadenesheli et al. (2016) consider the class of memoryless stochastic policies[1] with each action having a minimum probability $\iota > 0$ of being chosen: they show a regret guarantee of order $\widetilde{\mathcal{O}}(\sqrt{T})$ under this class of stochastic policies. Concerning the work of Xiong et al. (2022), they present the SEEU algorithm which alternates between purely exploratory and purely exploitative phases. Their algorithm reaches a $\mathcal{O}(T^{2/3})$ regret when compared against the optimal class of belief-based policies.

On the other hand, both Jafarnia Jahromi et al. (2022) and Russo et al. (2024a) assume, as in our setting, to have **knowledge of the observation model**. Jafarnia Jahromi et al. (2022) develop the PSRL-POMDP algorithm, a Bayesian approach that jointly learns the model parameters and exploits the available knowledge. They prove a Bayesian regret of order $\mathcal{O}(T^{2/3})$

[1] In a memoryless policy, the next action is chosen only based on the last received observation.

when compared against the optimal policy, however they do not provide guarantees for the employed model estimator: the obtained result on the regret only holds by assuming to have a consistent estimator. Differently, Russo et al. (2024a) develop the OAS estimation approach and prove consistency for it. They consider the powerful class of belief-based policies but focus on those having a minimum probability $\iota > 0$ of choosing each action. Under these conditions, they reach a regret order of $\widetilde{\mathcal{O}}(\sqrt{T})$ when compared against this class of stochastic policies.
We refer the reader to Appendix A and to Table 1 for a detailed comparison with the works mentioned above. It characterizes the different works in terms of adopted assumptions, applied estimation procedures, and algorithm properties.

Stemming from the OAS estimation procedure introduced in Russo et al. (2024a), we adopt similar ideas and present a new estimation approach that overcomes the limiting assumption on the minimum action probability. This aspect allows us to successfully compare the newly devised regret minimization algorithm against the optimal class of POMDP policies.

3 PRELIMINARIES

In this section, we present the adopted notation and the necessary background that will be useful to understand what will follow.

Partially Observable MDP A Partially Observable Markov Decision Process (POMDP) (Åström, 1965) is defined by a tuple $\mathcal{Q} \coloneqq (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathbb{T}, \mathbb{O}, \boldsymbol{\nu}, r)$ with $\mathcal{S}$ being a finite state space ($|\mathcal{S}| =: S$), $\mathcal{A}$ a finite action space ($|\mathcal{A}| =: A$) and $\mathcal{O}$ a finite observation space ($|\mathcal{O}| =: O$). $\mathbb{T} = \{\mathbb{T}_a\}_{a\in\mathcal{A}}$ denotes a collection of transition matrices $\mathbb{T}_a$ with size $S \times S$. Each transition matrix is such that $\mathbb{T}_a(\cdot|s) \in \Delta(\mathcal{S})$[2] defines the distribution of the next state when the agent takes action a in state s. $\mathbb{O} = \{\mathbb{O}_a\}_{a\in\mathcal{A}}$ denotes the set of observation matrices of size $O \times S$ such that $\mathbb{O}_a(\cdot|s) \in \Delta(\mathcal{O})$ represents the distribution over observations when the agent takes action a conditioned on the hidden state s. $\boldsymbol{\nu} \in \Delta(\mathcal{S})$ denotes the distribution over the initial state, while $r : \mathcal{O} \to [0, 1]$ is the known reward function mapping each observation to a finite reward such that $r(o)$ is the reward received when the agents observe $o \in \mathcal{O}$. In a POMDP, states are hidden and the agent can only see its own actions and the received observations. The interaction proceeds as follows. At each step t, the agent chooses an action a_t and gets an observation o_t which is conditioned on the hidden state s_t according to the law $\mathbb{O}_{a_t}(\cdot|s_t)$. Finally, the action taken will make the POMDP transition into a new hidden state s_{t+1} according to the distribution $\mathbb{T}_{a_t}(\cdot|s_t)$.

Policies in POMDPs A policy $\pi := (\pi_t)_{t=0}^{\infty}$ defines the set of decision rules characterizing the interaction of an agent with the environment. Deterministic policies map a history $h \in \mathcal{H}_t$ into actions $\pi_t : \mathcal{H}_t \to \mathcal{A}$ such that $a = \pi_t(h)$ defines the action chosen when history $h \in \mathcal{H}_t$ is observed. When interacting with a POMDP the history can be defined as $h := (a_j, o_j)_{j=0}^{t} \in \mathcal{H}_t$. We denote by $\mathcal{H}$ the space of histories of arbitrary length.

From POMDP to Belief MDP In a POMDP, it is always possible to build a belief vector $b_t \in \mathcal{B}$ (with $\mathcal{B} \coloneqq \Delta(\mathcal{S})$) by using the knowledge of the transition and observation matrices $\mathbb{T}$ and $\mathbb{O}$ respectively, and from the observed history at time $t-1$, $h_{t-1} := (a_j, o_j)_{j=0}^{t-1}$. Thus, at time t it holds $b_t(s) := P(S_t = s|h_{t-1})$. From the agent's point of view, a POMDP can be seen as a belief MDP (Krishnamurthy, 2016). The update rule of the belief b_{t+1} is determined by Bayes's theorem as follows:

$$b_{t+1}(s) = \frac{\sum_{s'\in\mathcal{S}} b_t(s')\mathbb{O}_{a_t}(O_t = o_t|S_t = s')\mathbb{T}_{a_t}(s|s')}{\sum_{s''\in\mathcal{S}} \mathbb{O}_{a_t}(O_t = o_t|S_t = s'')b_t(s'')}. \tag{1}$$

Given an initial belief b, the average reward of the infinite-horizon belief MDP is defined as: $\rho_b^\pi := \limsup_{T\to\infty}(1/T)\mathbb{E}[\sum_{t=0}^{T-1} r(o_t)|b_0 = b)]$. If the underlying MDP is weakly communicating, Bertsekas (1995) showed that $\rho^* := \sup_\pi \rho_b^\pi$ is independent of the initial belief b and the following Bellman optimality equation can be defined:

$$\rho^* + v(b) = \max_{a\in\mathcal{A}} \left[g(b, a) + \int_{\mathcal{B}} P(\, db'|b, a)v(b') \right], \tag{2}$$

with $g(b, a)$ representing the expected instantaneous reward obtained when taking action a under belief b such that $g(b, a) = \sum_{s\in\mathcal{S}} \sum_{o\in\mathcal{O}} b(s)\mathbb{O}_a(o|s)r(o)$. Ultimately, $v : \mathcal{B} \to \mathbb{R}$ defines the bias function quantifying the cumulative deviation of rewards with respect to ρ^* when starting from a belief b (Mahadevan, 1996).

4 PROBLEM FORMULATION

We tackle the average-reward infinite-horizon POMDP setting described in Section 3. In particular, we focus on the class of *undercomplete* POMDPs (Jin et al., 2020), where the number of states is less than or equal

[2] We use $\Delta(\mathcal{X})$ to denote the simplex over a finite space $\mathcal{X}$.

to the number of observations ($S \leqslant O$). As in previous works (Jafarnia Jahromi et al., 2022; Russo et al., 2024a), we assume knowledge of the observation model $\mathbb{O} = \{\mathbb{O}_a\}_{a\in\mathcal{A}}$, while we learn the transition model $\mathbb{T} = \{\mathbb{T}_a\}_{a\in\mathcal{A}}$.

We consider the class of belief-based policies mapping the space $\mathcal{B}$ of belief over the states to actions, such that $\pi : \mathcal{B} \to \mathcal{A}$. We denote by $\mathcal{P}$ the set of such belief-based policies.

In the following, we introduce the assumptions that we enforce for our setting.

Assumption 4.1. (**Minimum Value Transition Matrices**) *The smallest value in the transition matrices is* $\epsilon := \min_{s,s'\in\mathcal{S}} \min_{a\in\mathcal{A}} \mathbb{T}_a(s'|s) > 0$.

Despite seeming a strong assumption, this one-step reachability condition is widely used in works addressing partial observability (Zhou et al., 2021; Russo et al., 2024a; Jiang et al., 2023; Xiong et al., 2022). It is used for multiple reasons: first, it ensures geometric ergodicity of the Markov chain induced by the employed policy; second, it plays a key role in the theoretical analysis since allows to bound the error in the estimated belief vector as a function of the error in the estimated transition model (see Lemma D.1 for details). In practical scenarios, this assumption is satisfied in various POMDP applications, such as those involving information gathering (Guo et al., 2016).

Assumption 4.2. (**α-weakly Revealing Condition**) *There exists* $\alpha > 0$ *such that* $\min_{a\in\mathcal{A}} \sigma_S(\mathbb{O}_a) \geqslant \alpha$.

Here, we use $\sigma_S(\mathbb{O}_a)$ to denote the S-th singular value of matrix $\mathbb{O}_a$. This second assumption relates to the identifiability of the POMDP parameters and has been largely used in works tackling the partially observable setting. This condition quantifies the amount of information provided by the observations when inferring the latent states. A positive α value rules out pathological POMDP instances and identifies the tractable subclass of *weakly revealing* POMDPs (Jin et al., 2020; Liu et al., 2022a,b) (see Section 2). This assumption is related to the more typical *full-rank* condition employed in works using spectral decomposition techniques (Azizzadenesheli et al., 2016; Zhou et al., 2021; Hsu et al., 2012). A direct implication of this condition is that $S \leqslant O$, which represents a common scenario in many real-world settings, such as medical applications where the state (physical condition) of a patient generates a large number of different observations (Hauskrecht and Fraser, 2000) or dialogue systems, with a number of observations (words) that is much larger than the possible states (topics) of the conversation (Png et al., 2012).

Learning Objective As defined before, the objective in our setting is to find the belief-based policy maximizing Equation (2).

Xiong et al. (2022) have shown that under assumption 4.1, this equation is always verified. We tackle this problem using a regret minimization approach and we compare our policy against the optimal POMDP policy which plays according to the current belief value. Since determining the optimal policy for the POMDP model is generally computationally intractable (Madani, 1999), in this work we do not focus on solving this planning problem. Instead, we assume access to an optimization oracle able to solve Equation (2), thus maximizing the average reward ρ^* while returning the optimal policy $\pi \in \mathcal{P}$ under a given model.

The total regret over the interaction horizon of length T is thus defined as:

$$\mathcal{R}_T := T\rho^* - \sum_{t=0}^{T-1} r^{\pi}(o_t), \tag{3}$$

with apex π in the reward denoting that observations are obtained while following policy $\pi \in \mathcal{P}$.

5 ACTION-WISE OAS ESTIMATION PROCEDURE

Based on the OAS approach developed in Russo et al. (2024a), we present here the *Action-wise Observation-Aware Spectral* (AOAS) estimation technique that aims at estimating the transition model $\mathbb{T} = \{\mathbb{T}_a\}_{a\in\mathcal{A}}$ characterizing the POMDP[3].

Let us assume to interact with a POMDP instance $\mathcal{Q}$ using a belief-based policy $\pi \in \mathcal{P}$ and to collect samples $\mathcal{D} = \{(a_t, o_t)_{t=0}^n\}$ with $n + 1 = |\mathcal{D}|$ denoting the cardinality of the dataset.

We will then group pairs of samples collected in consecutive timestamps such that from dataset $\mathcal{D}$ we can build a new dataset $\mathcal{G} = \{(a_t, a_{t+1}, o_t, o_{t+1})_{t=0}^{n-1}\}$ having cardinality $n = |\mathcal{G}|$. The tuples of the form (a, a', o, o') with $a, a' \in \mathcal{A}$ and $o, o' \in \mathcal{O}$ will be the core elements employed in our estimation approach.

The estimation of transition matrix $\mathbb{T}_a$ related to action a is done by only considering those tuples in $\mathcal{G}$ whose first element $(a_t)_{t=0}^{n-1}$ coincides with action a, while the remaining part of the tuple (a_{t+1}, o_t, o_{t+1}) is actually employed for estimation. For convenience, we use a vector notation to represent the last three elements of each tuple, such that $\boldsymbol{x}_t \in \mathbb{R}^{AO^2}$ will denote the one-hot encoding of tuple (a_{t+1}, o_t, o_{t+1}).

Let us now denote by $\mathcal{E}(a, n, m)$ the event which holds

[3]Refer to Appendix A.1 for a detailed comparison between the approaches.

true when, in a dataset $\mathcal{G}$ of consecutive samples having cardinality n, the number of tuples having action a as a first element is equal to m. More formally:

$$\mathcal{E}(a,n,m) = \left\{ m = \sum_{t=0}^{n-1} \mathbb{1}\{a_t = a\} \right\}. \tag{4}$$

Here, we use $\mathbb{1}\{\cdot\}$ to denote the indicator function, which is equal to 1 when the condition is satisfied and 0 otherwise.
The elements defined above allow us to present the following distribution of interest $\boldsymbol{d}_{AO^2}^{(a,n,m)} \in \Delta(\mathcal{A} \times \mathcal{O}^2)$ over the tuples (a', o, o'):

$$\boldsymbol{d}_{AO^2}^{(a,n,m)} = \mathbb{E}_{\pi,\boldsymbol{\nu}}\left[\frac{1}{m} \sum_{t=0}^{n-1} \mathbb{1}\{a_t = a\}\, \boldsymbol{x}_t \,\middle|\, \mathcal{E}(a,n,m) \right], \tag{5}$$

where the expectation is with respect to policy $\pi \in \mathcal{P}$ and the initial state distribution $\boldsymbol{\nu} \in \Delta(\mathcal{S})$ of the POMDP, and it is conditioned on the event $\mathcal{E}(a,n,m)$. Here, the subscript AO^2 employed for the presented distribution represents the size of its support.[4]

Since we know that the received observations can be mapped to the underlying latent states using the observation model $\mathbb{O}$, we can define a relation linking the distribution $\boldsymbol{d}_{AO^2}^{(a,n,m)}$ defined on the tuples (a', o, o') with an analogous distribution $\boldsymbol{d}_{AS^2}^{(a,n,m)} \in \Delta(\mathcal{A} \times \mathcal{S}^2)$ defined on the non-observable tuples (a', s, s')[5]. Indeed, we can easily observe that, for each element (a', o, o') of vector $\boldsymbol{d}_{AO^2}^{(a,n,m)}$, we have:

$$\boldsymbol{d}_{AO^2}^{(a,n,m)}(a', o, o') = \sum_{s,s' \in \mathcal{S}} \mathbb{O}_a(o|s)\, \mathbb{O}_{a'}(o'|s')\, \boldsymbol{d}_{AS^2}^{(a,n,m)}(a', s, s').$$

The relation stated above can be defined for all the elements of the considered distributions. Thus, using matrix notation, we have:

$$\boldsymbol{d}_{AO^2}^{(a,n,m)} = \mathbb{B}_a\, \boldsymbol{d}_{AS^2}^{(a,n,m)}, \tag{6}$$

where $\mathbb{B}_a$ is a block diagonal matrix of size $AO^2 \times AS^2$ obtained by aligning along its diagonal the matrices $\{\mathbb{O}_{a,a'}\}_{a' \in \mathcal{A}}$. The different matrices $\mathbb{O}_{a,a'}$ have dimension $O^2 \times S^2$ and are in turn obtained as follows:

$$\mathbb{O}_{a,a'} \coloneqq \mathbb{O}_a \otimes \mathbb{O}_{a'},$$

where $\otimes$ denotes the Kronecker product between matrices $\mathbb{O}_a$ and $\mathbb{O}_{a'}$. We recall that since the observation model is available, we can compute the block diagonal matrix $\mathbb{B}_a$ for any $a \in \mathcal{A}$.

The distribution $\boldsymbol{d}_{AS^2}^{(a,n,m)}$ can be linked to the transition matrix $\mathbb{T}_a$ using the following considerations. First of all, we define a new quantity $\boldsymbol{d}_{S^2}^{(a,n,m)} \in \Delta(\mathcal{S}^2)$ which is obtained by aggregating elements in $\boldsymbol{d}_{AS^2}^{(a,n,m)}$. In particular, each element of this new vector is obtained as:

$$\boldsymbol{d}_{S^2}^{(a,n,m)}(s, s') = \sum_{a' \in \mathcal{A}} \boldsymbol{d}_{AS^2}^{(a,n,m)}(a', s, s'). \tag{7}$$

The final step involves recognizing the proportional relationship between elements in $\mathbb{T}_a$ and $\boldsymbol{d}_{S^2}^{(a,n,m)}$, which leads to the final expression:

$$\mathbb{T}_a(s'|s) = \frac{\boldsymbol{d}_{S^2}^{(a,n,m)}(s, s')}{\sum_{s'' \in \mathcal{S}} \boldsymbol{d}_{S^2}^{(a,n,m)}(s, s'')} \quad \forall s, s' \in \mathcal{S}. \tag{8}$$

For details on the derivation of Equation (8), we refer to Lemma E.6.

Estimation Procedure Having defined the procedure connecting the distribution on the action-observation tuples $\boldsymbol{d}_{AO^2}^{(a,n,m)}$ to the transition model, we show how an estimate of the transition matrix $\mathbb{T}_a$ can be computed. The following holds for any action $a \in \mathcal{A}$. Let a policy π interact with the environment for $n+1$ timestamps generating a dataset $\mathcal{D}$ of samples and let us group consecutive samples obtaining a new dataset $\mathcal{G}$ with cardinality n, as previously described. By denoting with $n(a)$ the number of tuples in $\mathcal{G}$ starting with action a, we can estimate $\boldsymbol{d}_{AO^2}^{(a,n,n(a))}$ as:

$$\widehat{\boldsymbol{d}}_{AO^2}^{(a,n,n(a))} = \frac{1}{n(a)} \sum_{t=0}^{n-1} \mathbb{1}\{a_t = a\}\, \boldsymbol{x}_t. \tag{9}$$

The estimate corresponding to the associated distribution $\widehat{\boldsymbol{d}}_{AS^2}^{(a,n,n(a))}$ over the non-observable tuples (a', s, s') can thus be obtained by inverting Equation (6) as follows:

$$\widehat{\boldsymbol{d}}_{AS^2}^{(a,n,n(a))} = \mathbb{B}_a^{\dagger}\, \widehat{\boldsymbol{d}}_{AO^2}^{(a,n,n(a))}, \tag{10}$$

where $\mathbb{B}_a^{\dagger}$ denotes the Moore-Penrose of matrix $\mathbb{B}_a$. We stress that, by the weakly-revealing assumption (4.2) and the properties of the Kronecker product, this matrix is always invertible since it holds that $\sigma_{\min}(\mathbb{B}_a) \geqslant \alpha^2$.
In a subsequent step, Equation (7) is used to obtain an estimate $\widehat{\boldsymbol{d}}_{S^2}^{(a,n,n(a))}$. Since this estimate may erroneously contain negative terms, we modify the negative ones by setting them to 0, thus obtaining a non-negative vector $\bar{\boldsymbol{d}}_{S^2}^{(a,n,n(a))}$. The newly obtained quantity is then plugged into Equation (8) to compute an

[4] A similar notation will be used as well for other distributions defined throughout the work.

[5] For a formal definition of this distribution, we refer to Appendix B.

estimate $\widehat{\mathbb{T}}_a$ of the action transition matrix.
The pseudocode of the reported approach is presented in Algorithm 1.

5.1 Sample Reuse and Theoretical Guarantees

One of the typical issues affecting the learning of model parameters in the average-reward POMDP setting is the inability to use samples coming from different policies (Azizzadenesheli et al., 2016; Russo et al., 2024a). The objective of this section is to show a simple estimator that is able to overcome this problem.
In particular, let us assume that $(\pi_i)_{i=0}^{k-1}$ policies interact separately with the environment and let us denote with $(\mathcal{G}_i)_{i=0}^{k-1}$ the generated datasets of consecutive samples. Let n_i and $n_i(a)$ indicate respectively the cardinality of the dataset $\mathcal{G}_i$ and the number of tuples from this dataset starting with action a. Based on these quantities, we can define a mixed distribution described as:

$$\boldsymbol{d}_{AO^2}^{(a,k)} = \frac{1}{N_k(a)} \sum_{i=0}^{k-1} n_i(a)\, \boldsymbol{d}_{AO^2}^{(a,n_i,n_i(a))}, \qquad (11)$$

where $N_k(a) = \sum_{i=0}^{k-1} n_i(a)$, while the new quantity $\boldsymbol{d}_{AO^2}^{(a,k)} \in \Delta(\mathcal{A} \times \mathcal{O}^2)$ mixes the per-policy distributions assigning to each of them a weight proportional to $n_i(a)$. An unbiased estimate of this quantity is obtained as:

$$\begin{aligned}\widehat{\boldsymbol{d}}_{AO^2}^{(a,k)} &= \frac{1}{N_k(a)} \sum_{i=0}^{k-1} n_i(a)\, \widehat{\boldsymbol{d}}_{AO^2}^{(a,n_i,n_i(a))} \\ &= \frac{1}{N_k(a)} \sum_{i=0}^{k-1} \sum_{j=0}^{n_i-1} \mathbb{1}\{a_{j,i} = a\}\, \boldsymbol{x}_{j,i} \qquad (12)\end{aligned}$$

where we use $a_{j,i}$ to represent the action at timestamp j referring to dataset $\mathcal{G}_i$, while $\boldsymbol{x}_{j,i}$ denotes the j-th indicator vector from $\mathcal{G}_i$.
By observing Equation (12), we can see that it is equivalent to the estimator defined in (9) when computed on the unique dataset $\mathcal{U} = \bigcup_{i=0}^{k-1} \mathcal{G}_i$ obtained from the union of the different datasets $\mathcal{G}_i$.
As we will see in Lemma 5.1, the number k of different policies influences the guarantees of the estimated transition matrix. However, this aspect is not reflected in the pseudocode of the *Action-wise* OAS procedure where the approach can be used without modifications by simply providing as input the union dataset $\mathcal{U}$.

The mixed distribution defined in (11) and its estimator show how to combine samples from different policies. By employing these quantities in the analysis and using Algorithm 1 on the collected data, we prove a consistent approach for estimating each action transition matrix $\mathbb{T}_a$, as observed from the following result:

Algorithm 1 Action-wise OAS Algorithm

Input: Observation matrix $\{\mathbb{O}_a\}_{a\in\mathcal{A}}$, dataset $\mathcal{G} = \{(a_0, a_1, o_0, o_1), \ldots, (a_{n-1}, a_n, o_{n-1}, o_n)\}$ of consecutive samples, dataset size $n = |\mathcal{G}|$
Create block diagonal matrices $\{\mathbb{B}_a\}_{a\in\mathcal{A}}$ from the observation model $\mathbb{O}$
Set action counters $n(a) = 0 \quad \forall a \in \mathcal{A}$
Define vectors of count $\boldsymbol{c}(a) \in \mathbb{R}^{AO^2} \quad \forall a \in \mathcal{A}$ and set their elements to zero
t = 0
while $t < n$ **do**
  Get tuple $(a_t, a_{t+1}, o_t, o_{t+1})$ from $\mathcal{G}$
  $\boldsymbol{x}_t \leftarrow$ one_hot_encode(a_{t+1}, o_t, o_{t+1})
  $\boldsymbol{c}(a_t) = \boldsymbol{c}(a_t) + \boldsymbol{x}_t$
  $n(a_t) = n(a_t) + 1$
  $t = t + 1$
end while
for $a \in \mathcal{A}$ **do**
  if $n(a) > 0$ **then**
    Compute $\widehat{\boldsymbol{d}}_{AO^2}^{(a,n,n(a))} = \boldsymbol{c}(a)/n(a)$
    Compute $\widehat{\boldsymbol{d}}_{AS^2}^{(a,n,n(a))}$ using Equation (10)
    Compute $\widehat{\boldsymbol{d}}_{S^2}^{(a,n,n(a))}$ using Equation (7)
    Compute positive $\bar{\boldsymbol{d}}_{S^2}^{(a,n,n(a))}$ from $\widehat{\boldsymbol{d}}_{S^2}^{(a,n,n(a))}$
    Compute $\widehat{\mathbb{T}}_a$ from $\bar{\boldsymbol{d}}_{S^2}^{(a,n,n(a))}$ using Equation (8)
  end if
end for

Lemma 5.1. *Let us assume that k policies $(\pi_i)_{i=0}^{k-1}$, each with $\pi_i \in \mathcal{P}$, separately interact with a POMDP instance $\mathcal{Q}$ satisfying Assumptions 4.1 and 4.2. By providing the union dataset $\mathcal{U} = \bigcup_{i=0}^{k-1} \mathcal{G}_i$ to Algorithm 1, with probability at least $1-\delta$, it holds that:*

$$\|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F \leqslant \frac{4\widetilde{G}}{\alpha^2 d_{\min}^{(a)}\,(1-\widetilde{\eta})} \sqrt{\frac{2k\,SA\,\log(2AO^2k/\delta)}{N_k(a)}}$$

where $\widetilde{G} \geqslant 1$ and $\widetilde{\eta} \leqslant 1 - \frac{\epsilon}{1-\epsilon}$ are determined by the deployed policies, while $d_{\min}^{(a)}$ represents the minimum state distribution conditioned on action a.

The parameter $\widetilde{\eta}$ appearing in the bound refers to a contraction coefficient associated with the different Markov chains induced by the policies and its value is always strictly smaller than 1 under Assumption 4.1. Using this assumption, we are also able to bound away from 0 the minimum state distribution $d_{\min}^{(a)}$. Finally, the α term deriving from Assumption 4.2 characterizes the amount of information carried by the observations to infer the underlying states.

Algorithm 2 The Action-wise OAS-UCRL Algorithm

1: **Input:** Observation matrix $\{\mathbb{O}_a\}_{a\in\mathcal{A}}$, confidence level δ, length of initial episode T_0
2: **Initialize:** $t=0$, $k=0$, policy π_0 uniform on actions $\mathcal{A}$, belief b_0 uniform over states $\mathcal{S}$, collected pairs of samples $\mathcal{G}=\varnothing$
3: **while** $t<T$ **do**
4: **if** $k>0$ **then**
5: Compute transition model $\widehat{\mathbb{T}}=\{\widehat{\mathbb{T}}_a\}_{a\in\mathcal{A}}$ from $\mathcal{G}$ using Algorithm 1
6: Build a confidence region $\mathcal{C}_{a,k}(\delta_{a,k})$ around each $\widehat{\mathbb{T}}_a$
7: Define the set $\mathcal{C}_k(\delta_k)$ of admissible POMDPs
8: Get policy π_k from the oracle (Equation 13)
9: **end if**
10: Set $n_k(a)=0 \quad \forall a\in\mathcal{A}$
11: Execute $a_t=\pi_k(b_t)$
12: Observe o_t, get reward $r_t=r(o_t)$
13: Update belief to b_{t+1} using Equation (1)
14: Set $t=t+1$
15: **while** $t<T_0$ or $n_k(\pi_k(b_t))<N_k(\pi_k(b_t))$ **do**
16: Execute $a_t=\pi_k(b_t)$
17: Observe o_t, get reward $r_t=r(o_t)$
18: Update belief to b_{t+1} using Equation (1)
19: Update $n_k(a_{t-1})=n_k(a_{t-1})+1$
20: Add $(a_{t-1},a_t,o_{t-1},o_t)$ to $\mathcal{G}$
21: Set $t=t+1$
22: **end while**
23: Set $N_{k+1}(a)=N_k(a)+n_k(a) \quad \forall a\in\mathcal{A}$
24: Set $k=k+1$
25: **end while**

Furthermore, we remark that Lemma 5.1 remains valid under a weaker condition than Assumption 4.1. In particular, it suffices to impose an ergodicity-like assumption for each action. In Appendix B, we provide a more detailed description of this assumption together with a formal derivation of the results of the Lemma.

6 ACTION-WISE OAS-UCRL ALGORITHM

In this section, we present the *Action-wise* OAS-UCRL (AOAS-UCRL) algorithm which is inspired by the combination of an optimistic approach mimicking the UCRL algorithm (Jaksch et al., 2010) and the *Action-wise* OAS estimation procedure. The algorithm starts with an initial episode $k=0$ of length T_0 where a uniform exploration policy is used to collect samples. At the beginning of each successive episode k, all the samples collected up to that moment are provided to the *Action-wise* OAS Algorithm. The *Action-wise* OAS algorithm returns as output the estimated transition model $\widehat{\mathbb{T}}=\{\widehat{\mathbb{T}}_a\}_{a\in\mathcal{A}}$. The algorithm then proceeds by building a confidence region $\mathcal{C}_{a,k}(\delta_{a,k})$ around every $\widehat{\mathbb{T}}_a$ such that the real transition matrix lies in it with high probability, namely $P(\mathbb{T}_a\in\mathcal{C}_{a,k}(\delta_{a,k}))\geqslant 1-\delta_{a,k}$, with $\delta_{a,k}\coloneqq\delta/(Ak^3)$. These confidence regions together define the confidence region $\mathcal{C}_k(\delta_k)$ of the real POMDP instance $\mathcal{Q}$, for which in turn it holds that $P(\mathcal{Q}\in\mathcal{C}_k(\delta_k))\geqslant 1-\delta_k$, with $\delta_k\coloneqq\delta/k^3$. As specified in Section 4, we assume the existence of an oracle that is able to compute the optimal policy corresponding to the optimistic POMDP contained in $\mathcal{C}_k(\delta_k)$. More formally, the oracle computes:

$$\pi_k=\arg\max_{\pi\in\mathcal{P}}\max_{\widetilde{\mathcal{Q}}\in\mathcal{C}_k(\delta_k)}\rho(\pi,\widetilde{\mathcal{Q}}), \tag{13}$$

where we used $\rho(\pi,\widetilde{\mathcal{Q}})$ to emphasize the dependence of the average reward from policy π and the POMDP instance $\widetilde{Q}$.
The policy π_k returned by the oracle is then used during episode k to interact with the environment. An episode k terminates whenever there is at least an action a such that the number $n_k(a)$ of times it appears as a first element in that episode matches the total number of times $N_k(a)$ it appears as a first element from the beginning of the interaction (line 15). The pseudocode of the approach is reported in Algorithm 2.

We prove the following result for the *Action-wise* OAS-UCRL algorithm. The proof is reported in Appendix C.

Theorem 6.1. *Let us assume to have a POMDP instance $\mathcal{Q}$ satisfying Assumptions 4.1 and 4.2. If the Action-wise OAS-UCRL algorithm is run for T steps, with probability at least $1-2\delta$, it suffers from a total regret:*

$$\mathcal{R}_T\leqslant\mathcal{O}\left(\frac{CD\widetilde{G}}{\alpha^2\widetilde{d}_{\min}}\sqrt{SA^3T\,\log T\,\log O}\right).$$

where $C\coloneqq\frac{4(1-\epsilon)^3}{\epsilon^4}$ and D is a finite constant bounding the span of the bias function (definition in Proposition E.1).

This result is achieved also thanks to a new bound on the belief error presented in Lemma D.1.
To the best of our knowledge, this is the first algorithm enjoying a regret of order $\mathcal{O}(\sqrt{T\log T})$ in the average-reward POMDP setting when compared against the optimal policy, thus improving over state-of-the-art approaches. Indeed, the PSRL-POMDP algorithm (Jafarnia Jahromi et al., 2022) reaches a $\mathcal{O}(T^{2/3})$ regret guarantee, but their result holds under the assumption of employing a consistent estimator, which they do not provide.
Concerning the OAS-UCRL approach (Russo et al., 2024a), it reaches a $\mathcal{O}(\sqrt{T\log T})$ regret guarantee

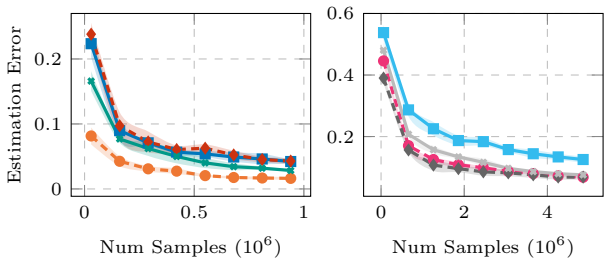

Figure 1: Error in Frobenious norm of the Different Action Transition Matrices under two POMDP Instances (10 runs, 95 %c.i.).

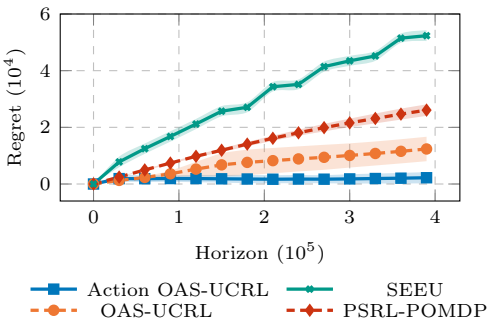

Figure 2: Regret Results on a POMDP Instance with $S = 3$, $A = 4$, $O = 4$ (10 runs, 95 %c.i.).

when compared to the weaker class of stochastic belief-based policies. It is possible to show that, by optimizing the minimum action probability $\iota > 0$ characterizing their policy class, their approach suffers a regret of order $\tilde{\mathcal{O}}(T^{4/5})$ when compared against the optimal policy.

7 NUMERICAL SIMULATIONS

In this section, we will provide numerical simulations testing both the AOAS estimation procedure and the AOAS-UCRL algorithm presented in the previous sections. Further experiments and simulation details are reported in Appendix F.

Estimation Performance This first set of experiments shows the estimation error of the transition model for two different POMDP instances when the *Action-wise* OAS algorithm is employed. In particular, the objective is to show that AOAS is able to reduce the estimation error when using samples collected from different policies. On the left plot of Figure 1, the POMDP has $S = 5$ states, $A = 4$ actions and $O = 8$ observations, while on the right the values are $S = 10$, $A = 4$ and $O = 16$. Each line in the plot represents the estimation error of the associated transition matrix $\mathbb{T}_a$. Samples are collected using belief-based policies that change every 10^4 steps. The implemented policies play the action maximizing the immediate reward given the current belief and their change is determined by varying the transition model used to update the belief vector[6]. To promote the choice of different actions, we employ stochastic policies. We notice that the approach works well also in larger problem instances, such as the one on the right. From details reported in Appendix F about the considered instances, we observe that the actions having higher error are either those that have been chosen less (so fewer samples are available) or those associated with a block diagonal matrix $\mathbb{B}_a$ with low values of $\sigma_{\min}(\mathbb{B}_a)$.

Regret Results This second set of experiments compares the *Action-wise* OAS-UCRL algorithm with different baseline approaches. We exclude from the comparison the SM-UCRL algorithm from Azizzadenesheli et al. (2016) since it employs memoryless policies which are known to yield a linear regret with respect to our oracle. We compare against the SEEU (Xiong et al., 2022) approach but we use a modified version that provides the algorithm with information about the observation model. However, this approach is the one showing the highest regret under the considered instance and this aspect is mainly due to (i) the need for SD approaches of a large number of samples to provide good estimates and (ii) the inherent nature of the algorithm which alternates between purely exploratory and purely exploitative phases.
The comparison proceeds with other baseline algorithms naively developed under the assumption of knowing the observation model. It is possible to see that our solution outperforms the competing alternatives, thus validating the theoretical results. Concerning the PSRL-POMDP algorithm, we chose to implement it using a standard particle filter approach which however lacks estimation guarantees, as reflected in the suffered regret. Regarding the OAS-UCRL algorithm instead, the higher regret with respect to our algorithm can be attributed to the stochasticity of the employed policy (we set the minimum action probability to $\iota = 0.025$) but also to the less accurate estimates of the model parameters since OAS-UCRL only uses samples coming from the last episode for model estimation, thus discarding all previous ones.

In Appendix F, we present two different ablation studies. (i) The first shows the regret performance of Action OAS-UCRL when compared against the OAS-UCRL algorithm run under different values of the minimum action probability ι, (ii) the second one in-

[6]We refer to the transition model adopted in Equation 1 to update the belief, which can be arbitrarily different from the real transition model.

stead explores the impact of the samples reuse strategy adopted by Action OAS-UCRL.

8 CONCLUSIONS AND FUTURE WORKS

We introduced a novel estimation procedure to learn the POMDP parameters assuming the knowledge of the observation model. We showed how this approach can be used to learn the transition matrix associated with each action separately; we proved consistency for it, and we highlighted that it can even be used with samples collected under different policies. After that, we proposed the *Action-wise* OAS-UCRL algorithm which, to the best of our knowledge, is the first to achieve a regret guarantee of order $\widetilde{\mathcal{O}}(\sqrt{T})$ when compared against the optimal policy in the average reward POMDP setting. We reached this result thanks to the new proposed estimation procedure and new tighter theoretical results on the estimated belief error.
In future work, we plan to extend these techniques to a more standard setting where neither the transition nor the observation model is available.

Acknowledgements

This paper is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

References

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, jan 2014.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Anima Anandkumar. Reinforcement learning of pomdps using spectral methods. In *Annual Conference Computational Learning Theory*, 2016.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 1967.

Peter L. Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. 2009.

Dimitri Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. 01 1995.

Fan Chen, Huan Wang, Caiming Xiong, Song Mei, and Yu Bai. Lower bounds for learning in revealing pomdps. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Yohann De Castro, Elisabeth Gassiat, and Sylvain Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden markov models. *IEEE Transactions on Information Theory*, 63(8):4758–4777, 2017. doi: 10.1109/TIT.2017.2696959.

Zhaohan Guo, Shayan Doroudi, and Emma Brunskill. A pac rl algorithm for episodic pomdps. 05 2016.

Milos Hauskrecht and Hamish S. F. Fraser. Planning treatment of ischemic heart disease with partially observable markov decision processes. *Artificial intelligence in medicine*, 2000.

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. ISSN 0022-0000. JCSS Special Issue: Cloud Computing 2011.

Mehdi Jafarnia Jahromi, Rahul Jain, and Ashutosh Nayyar. Online learning for unknown partially observable mdps. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, aug 2010. ISSN 1532-4435.

Bowen Jiang, Bo Jiang, Jian Li, Tao Lin, Xinbing Wang, and Chenghu Zhou. Online restless bandits with unobserved states. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Chi Jin, Sham M. Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Vikram Krishnamurthy. *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambridge University Press, 2016. doi: 10.1017/CBO9781316471104.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary?, 2022a.

Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvári, and Chi Jin. Optimistic mle – a generic model-based algorithm for partially observable sequential decision making, 2022b.

Omid Madani. On the computability of infinite-horizon partially observable markov decision processes. 12 1999.

Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. 1996.

Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. *31st International Conference on Machine Learning, ICML 2014*, 1, 05 2014.

Alonso Marco, Felix Berkenkamp, Philipp Hennig, Angela P. Schoellig, Andreas Krause, Stefan Schaal, and Sebastian Trimpe. Virtual vs. real: Trading off simulations and physical experiments in reinforcement learning with bayesian optimization. *CoRR*, 2017.

Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden markov models. Association for Computing Machinery, 2005.

Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Neural Information Processing Systems*, 2012.

Shaowei Png, Joelle Pineau, and Brahim Chaib-Draa. Building adaptive dialogue systems via bayes-adaptive pomdps. *IEEE Journal of Selected Topics in Signal Processing*, 2012.

Giorgia Ramponi, Amarildo Likmeta, Alberto Maria Metelli, Andrea Tirinzoni, and Marcello Restelli. Truly batch model-free inverse reinforcement learning about multiple intentions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, 2020.

Alessio Russo, Alberto Maria Metelli, and Marcello Restelli. Efficient learning of pomdps with known observation model in average-reward setting, 2024a. URL `https://arxiv.org/abs/2410.01331`.

Alessio Russo, Alberto Maria Metelli, and Marcello Restelli. Switching latent bandits. *Transactions on Machine Learning Research*, 2024b.

Edward J. Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2): 282–304, 1978.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Brijen Thananjeyan, Kirthevasan Kandasamy, Ion Stoica, Michael Jordan, Ken Goldberg, and Joseph Gonzalez. Resource allocation in multi-armed bandit exploration: Overcoming sublinear scaling with adaptive parallelism. PMLR, 2021.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2010.

Yi Xiong, Ningyuan Chen, Xuefeng Gao, and Xiang Zhou. Sublinear regret for learning pomdps, 2022.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312. PMLR, 09–15 Jun 2019.

Xiang Zhou, Yi Xiong, Ningyuan Chen, and Xuefeng Gao. Regime switching bandits. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

K.J Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 1965.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**, we focused on the properties and highlighted the dependency with respect to the sample size.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **No**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **Yes**
   (c) Clear explanations of any assumptions. **Yes**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes/No**, instructions and parameters are provided in Appendix F.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**, see Appendix F.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**, they are reported in the captions of the figures.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Not Applicable**
   (b) The license information of the assets, if applicable. **Not Applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable**
   (d) Information about consent from data providers/curators. **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**

APPENDIX ORGANIZATION

Here is an outline of the appendix.

- Section A presents a comparison of our work with the most relevant related works.
- Section B and Section C are devoted respectively to the proof of Lemma 5.1 and Theorem 6.1.
- Section D contains Lemma D.1, a result that relates the error in the belief vector with the error in the estimated transition model. This bound improves over existing results and is crucial for proving Theorem 6.1.
- Section E is miscellaneous of new and existing useful results that support the theoretical analysis of the work.
- Section F provides further details on the simulations and their results presented in the main paper.

A COMPARISON WITH RELATED WORKS

This section is devoted to a more detailed comparison of our work with state-of-the-art approaches in this setting. In particular, we confront with: the SM-UCRL algorithm (Azizzadenesheli et al., 2016), the SEEU algorithm (Xiong et al., 2022), the PSRL-POMDP algorithm (Jafarnia Jahromi et al., 2022) and the OAS-UCRL algorithm (Russo et al., 2024a) (Section A.1).

- the **SM-UCRL algorithm** (Azizzadenesheli et al., 2016) tackles the POMDP learning problem without assuming knowledge of either the transition or the observation model. It employs standard Spectral Decomposition approaches to learn the model parameters. However, it assumes the class of memoryless policies that are characterized by choosing the next action only conditioning on the last observation seen: these policies are suboptimal in the POMDP setting. Furthermore, they assume that each action is always chosen with a minimum probability $\iota > 0$. The SM-UCRL algorithm works in different episodes and the model parameters are computed at the beginning of each episode using only samples collected during the previous episode, thus discarding all the others. The algorithm reaches a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret when compared against the Optimal Stochastic Memoryless policy.
- the **SEEU algorithm** (Xiong et al., 2022) considers again the standard POMDP setting without having partial knowledge of the model. The algorithm alternates between purely exploratory phases when collected samples are used to estimate the model parameters using Spectral approaches, and exploitative phases where an optimal policy is used on the learned optimistic POMDP. This algorithm shows a $\widetilde{\mathcal{O}}(T^{2/3})$ regret guarantee when compared against the optimal Belief-based policy, which is the optimal class of policy in this setting. Since they consider the class of belief-based policies, they require, as in our case, the one-step reachability assumption 4.1 to obtain regret guarantees for their approach.
- the **PSRL-POMDP algorithm** (Jafarnia Jahromi et al., 2022) considers a POMDP setting with a known observation model but an unknown transition model. They employ a Bayesian approach to update the model parameters at each timestamp. However, they do not provide a consistent approach for model estimation and base their results on the assumption that the employed estimates are consistent, thus obtaining a more accurate estimate as more samples are acquired. Under this assumption, and an analogous assumption on the consistency of the belief estimates, they show a Bayesian regret of order $\mathcal{O}(T^{2/3})$ against the optimal POMDP policy.

A.1 Comparison between the *Action-wise* OAS-UCRL Algorithm and the OAS-UCRL Algorithm of (Russo et al., 2024a)

In this section, we will show the main differences of our *Action-wise* OAS-UCRL with respect to the OAS-UCRL approach described in (Russo et al., 2024a). In particular, we highlight that:

- the OAS-UCRL approach employs the class of stochastic belief-based policies for which each action has a minimum probability $\iota > 0$ of being chosen at each timestamp. This ensures a continual refinement of the estimate of the transition matrix $\mathbb{T}_a$ over time.

- Because of the previous point, in their regret analysis, they compare against the optimal stochastic belief-based policy and they reach a regret guarantee of order $\widetilde{\mathcal{O}}(\sqrt{T})$ with respect to this oracle. However, we improve over their result since we obtain a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret guarantee when compared with the optimal deterministic belief-based policy.
  It is indeed possible to show that, by optimizing their regret result over the minimum action probability ι, the OAS-UCRL algorithm suffers regret $\widetilde{\mathcal{O}}(T^{4/5})$ when compared against the optimal policy. We observe that the regret of the OAS-UCRL approach with respect to the optimal stochastic policy can be bounded by[7] $C\sqrt{T}/\iota^{3/2}$, where C is a constant related to the problem parameters. When compared against the optimal POMDP policy, the regret of the OAS-UCRL algorithm can be expressed as:

$$\mathcal{R}_T \leqslant T\,(A-1)\iota + C\frac{\sqrt{T}}{\iota^{3/2}} \tag{14}$$

  where we introduced an additional term $T\,(A-1)\iota$ representing the regret suffered when choosing the suboptimal action, which happens with probability $(A-1)\iota$. This probability is then multiplied by the total interaction horizon T.
  By optimizing the regret in (14) with respect to the minimum action probability ι, we obtain a final regret order of $\widetilde{\mathcal{O}}(T^{4/5})$ for the OAS-UCRL algorithm under the optimal POMDP policy.

The improvements in terms of regret of the *Action-wise* OAS-UCRL algorithm over the OAS-UCRL counterpart are mainly due to: (i) a tighter analysis of the belief estimation error (see Lemma D.1); (ii) the differences in the employed estimation procedure.
We report here the main differences between the *Action-wise* OAS and the OAS procedures:

- The OAS procedure focuses on estimating the stationary distribution on action-observation pairs[8] induced by the employed policy π while the *Action-wise* OAS procedure estimates distributions defined over a finite amount of samples and conditioned on the event defined in Equation (4). The distribution we consider allows us to obtain estimation guarantees separately for each action transition matrix $\mathbb{T}_a$, while the OAS procedure provides estimation guarantees for the whole transition model, namely it bounds $\sum_{a\in\mathcal{A}} \|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F$. Indeed, estimating the stationary distribution conditioned on action a was not a viable choice since, by removing the minimum action probability assumption used in Russo et al. (2024b), there may be cases where this conditional stationary distribution would not exist.
  This happens for example when a given action a is not played under stationary conditions, namely $\lim_{t\to\infty} P(A_t = a) = 0$, and this prevents us from defining the following stationary conditional distribution:

$$d^{(a)}_{AO^2}(a', o, o') \coloneqq \lim_{t\to\infty} d^{\pi}_t(a', o, o'|a),$$

  with $d^{\pi}_t(a', o, o'|a) \coloneqq P(A_{t+1} = a', O_t = o, O_{t+1} = o'|A_t = a)$. This aspect becomes crucial when the chain does not start from stationarity. Indeed, in this case, we may get some samples from a if, for some initial t, we have $P(A_t = a) > 0$ but we would not be able to use them to define an estimate of $d^{(a)}_{AO^2}$ since this conditional distribution does not exist.
  The scenario detailed above led us to opt for a distribution different from the stationary one.

- We show that the *Action-wise* OAS estimation procedure works also for samples collected under different policies. This improves the sample efficiency of the approach with respect to the OAS method which is instead characterized by only employing samples deriving from a unique distribution.

- As a drawback, our *Action-wise* OAS procedure requires Assumption 4.1 to hold, while the OAS approach only requires the ergodicity of the induced chain, which is a weaker condition than Assumption 4.1.

[7] Here, we disregard logarithmic terms.

[8] They estimate $d^{\pi}_{A^2O^2}(a, a', o, o') := \lim_{t\to\infty} d^{\pi}_t(a, a', o, o')$ with $d^{\pi}_t(a, a', o, o') := P(A_t = a, A_{t+1} = a', O_t = o, O_{t+1} = o'|\pi)$.

Table 1: Table Comparing with the Most Relevant Related Works.

	SM-UCRL	**SEEU**	**PSRL-POMDP**	**OAS-UCRL**	**Action-wise OAS-UCRL**
Knowledge of Observation Model	No	No	Yes	Yes	Yes
Ergodicity of Induced Chain	Yes	Yes	Yes	Yes	Yes
Minimum Transition Probability	No	Yes	No	Yes	Yes
Invertible Transition Model	Yes	Yes	No	No	No
Full-rank Observation Model*	Yes	Yes	Yes**	Yes	Yes
Minimum Action Probability	Yes	No	No	Yes	No
Consistent Transition Model Estimation	No	No	Yes	No	No
Consistent Belief Estimation	No	No	Yes	No	No
Estimation Technique	Spectral Decomp.	Spectral Decomp.	Bayesian Update	OAS	Action-wise OAS
Consistent Estimation	✓	✓	✗	✓	✓
Handles Uniform Policies	✓	✓	✗	✓	✓
Handles Memoryless Policies	✓	✗	✗	✓	✓
Handles Belief-based Policies	✗	✗	✗	✓	✓
Algorithm Type	Optimistic	Alternating Explor-Optimistic	Bayesian	Optimistic	Optimistic
Oracle Policy	Opt. Stochastic Memoryless	Opt. POMDP	Opt. POMDP	Opt. Stochastic POMDP	Opt. POMDP
Regret	$\tilde{\mathcal{O}}(\sqrt{T})$	$\tilde{\mathcal{O}}(T^{2/3})$	$\mathcal{O}(T^{2/3})$	$\tilde{\mathcal{O}}(\sqrt{T})$	$\tilde{\mathcal{O}}(\sqrt{T})$

A.2 Comparison Table

Inspired by Table 1 in Russo et al. (2024a), we present a similar comparison defined in terms of (i) required Assumptions (first sub-table), (ii) properties of used estimation techniques (second sub-table), and (iii) properties of the associated regret-minimization algorithm (third sub-table).

Concerning the first sub-table, the cells with *Yes* denote a required assumption, and viceversa.
Some notes referring to the content of the table:

(*): Saying that a matrix is full-rank corresponds to saying, under the weakly-revealing terminology, that $\alpha > 0$;

(**): For the PSRL-POMDP algorithm, the full-rank observability assumption is reported in terms of the Kullback-Leibler divergence between the probability distributions on the next observation when conditioned on different transition models (see their Assumption 2 for details).

B PROOF OF LEMMA 5.1

In this section, we will provide the proof for Lemma 5.1. This Lemma presents a bound on the estimation error of the action transition matrix $\mathbb{T}_a$ when samples used to make the estimate come from different policies. First of all, we start by reporting the statement.

Lemma 5.1. *Let us assume that k policies $(\pi_i)_{i=0}^{k-1}$, each with $\pi_i \in \mathcal{P}$, separately interact with a POMDP instance $\mathcal{Q}$ satisfying Assumptions 4.1 and 4.2. By providing the union dataset $\mathcal{U} = \bigcup_{i=0}^{k-1} \mathcal{G}_i$ to Algorithm 1, with probability at least $1-\delta$, it holds that:*

$$\|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F \leqslant \frac{4\widetilde{G}}{\alpha^2 d_{\min}^{(a)} (1-\widetilde{\eta})} \sqrt{\frac{2k\, SA\, \log(2AO^2k/\delta)}{N_k(a)}}$$

where $\widetilde{G} \geqslant 1$ and $\widetilde{\eta} \leqslant 1 - \frac{\epsilon}{1-\epsilon}$ are determined by the deployed policies, while $d_{\min}^{(a)}$ represents the minimum state distribution conditioned on action a.

Proof. This proof will be developed in two main parts: (i) the first one is devoted to showing the estimation error of the transition matrix both in the case of samples coming from a unique policy and the case of samples deriving from different policies; (ii) the second one shows how to reach theoretical guarantees on each action transition matrix $\mathbb{T}_a$ starting from guarantees on the action-observation distribution derived in part (i).

First Part of the Proof

Let us now focus on the first part of the proof and let us first consider the case of samples obtained from a unique policy.

Let us assume that a policy $\pi \in \mathcal{P}$ is employed to interact with the environment for $n+1$ steps and a dataset $\mathcal{D} = \{(a_t, o_t)_{t=0}^{n}\}$ is generated. By grouping pairs of samples collected in consecutive timestamps, we define the new dataset $\mathcal{G} = \{(a_t, a_{t+1}, o_t, o_{t+1})_{t=0}^{n-1}\}$ with cardinality $n = |\mathcal{G}|$.

We recall here the definition of the distribution reported in Equation (5), that is:

$$\boldsymbol{d}_{AO^2}^{(a,n,m)} = \mathbb{E}_{\pi,\boldsymbol{\nu}}\left[\frac{1}{m}\sum_{t=0}^{n-1} \mathbb{1}\{a_t = a\}\, \boldsymbol{x}_t \,\middle|\, \mathcal{E}(a,n,m)\right],$$

where the event $\mathcal{E}(a,n,m)$ holds true when the considered dataset $\mathcal{G}$ has size n and the number of tuples in $\mathcal{G}$ having action a as a first element coincides with m.
For completeness, let us also provide the formal definition of the distribution $\boldsymbol{d}_{AS^2}^{(a,n,m)} \in \Delta(\mathcal{A} \times \mathcal{S}^2)$ introduced in the main paper. Starting from the process that generated the dataset $\mathcal{G}$ of consecutive pairs, let us assume to have access to the underlying states $(s_t)_{t=0}^{n}$. As done for $\mathcal{G}$, we can define a dataset $\mathcal{M} = \{(a_t, a_{t+1}, s_t, s_{t+1})_{t=0}^{n-1}\}$ having cardinality $|\mathcal{M}| = n$. Let us also denote with $\boldsymbol{y}_t \in \mathbb{R}^{AS^2}$ the one-hot encoded vector defined over the last three elements (a', s, s') of each tuple. From the defined quantities, the distribution $\boldsymbol{d}_{AS^2}^{(a,n,m)}$ can be defined as:

$$\boldsymbol{d}_{AS^2}^{(a,n,m)} = \mathbb{E}_{\pi,\boldsymbol{\nu}}\left[\frac{1}{m}\sum_{t=0}^{n-1} \mathbb{1}\{a_t = a\}\, \boldsymbol{y}_t \,\middle|\, \mathcal{E}(a,n,m)\right].$$

Having clarified this aspect, we can go back to considering the distribution $\boldsymbol{d}_{AO^2}^{(a,n,m)}$.
Since this quantity contains elements that are all observable, the associated estimator can be computed by simply counting the realizations of observed tuples from dataset $\mathcal{G}$ and then dividing them by the number of samples m, as described in Equation (9):

$$\widehat{\boldsymbol{d}}_{AO^2}^{(a,n,m)} = \frac{1}{m} \sum_{t=0}^{n-1} \mathbb{1}\{a_t = a\}\, \boldsymbol{x}_t.$$

In the related work of Azizzadenesheli et al. (2016) a different estimator is employed but in a similar setting: in particular, in their Theorem 13, they consider estimates derived from samples drawn from a POMDP and the estimates are conditioned to a specific action a, as it is for our case.
In particular, we are able to show that, with probability at least $1-\delta$, we have:

$$\begin{aligned}
\left\| \boldsymbol{d}_{AO^2}^{(a,n,m)} - \widehat{\boldsymbol{d}}_{AO^2}^{(a,n,m)} \right\|_2 &= \left\| \frac{1}{m}\, \mathbb{E}_{\pi,\boldsymbol{\nu}} \left[\frac{1}{m} \sum_{t=0}^{n-1} \mathbb{1}\{a_t = a\}\, \boldsymbol{x}_t \,\middle|\, \mathcal{E}(a,n,m) \right] - \frac{1}{m} \sum_{t=0}^{n-1} \mathbb{1}\{a_t = a\}\, \boldsymbol{x}_t \right\|_2 \\
&\leqslant \sqrt{\left(\frac{G(\pi)}{1-\eta(\pi)} \right)^2 \frac{8 \log\left((AO^2+1)/\delta \right)}{m}} \qquad (15) \\
&\leqslant \frac{G(\pi)}{1-\eta(\pi)} \sqrt{\frac{8 \log\left(2AO^2/\delta \right)}{m}}, \qquad (16)
\end{aligned}$$

where the result in 15 combines both a concentration result on matrix estimates appearing in Tropp (2010) and an analysis on the variance of the samples coming from the Markov chain appearing in Azizzadenesheli et al. (2016) which shows that the Markovian dependency between samples leads to a further term $\frac{G(\pi)^2}{(1-\eta(\pi))^2}$ in the expression of the variance. Here, $1 \leqslant G(\pi) < \infty$ is the geometric ergodicity while $0 \leqslant \eta(\pi) < 1$ is the contraction coefficient, also known as Dobrushin coefficient (Krishnamurthy, 2016). Finally, the last inequality in 16 follows from simple algebraic manipulations.

Remark B.1. As observed in the main paper, we point out that an assumption weaker than Assumption 4.1 can be used in this part of the proof. Indeed, as observed above, the bound in line 15 holds under the geometric ergodicity assumption. For the quantities we estimate, this corresponds in having a policy π that induces a state distribution such that $d_S^{(a,n,m)}(s) > 0$ for each $s \in \mathcal{S}$, where we define the state distribution as:

$$d_S^{(a,n,m)}(s) = \sum_{s' \in \mathcal{S}} d_{S^2}^{(a,n,m)}(s,s'),$$

with $d_{S^2}^{(a,n,m)}(s,s')$ being defined in Equation (7). By assuming that this *action-ergodicity* condition holds for every policy π_i instead of directly using Assumption 4.1, the guarantees of this lemma can be preserved.

Having defined a bound holding for samples coming from a unique distribution, we are ready to extend this result to samples coming from multiple policies. Let us assume that k different policies are employed and let us denote with n_i the cardinality of the generated dataset $\mathcal{G}_i$ and with $n_i(a)$ the number of tuples from $\mathcal{G}_i$ starting with action a. By recalling the definitions of the expected distribution and the related estimator reported respectively in Equations (11) and (12), we can prove the following relation holding with probability at least $1-\delta$:

$$\begin{aligned}
\left\| \boldsymbol{d}_{AO^2}^{(a,k)} - \boldsymbol{d}_{AO^2}^{(a,k)} \right\|_2 &= \left\| \frac{1}{N_k(a)} \sum_{i=0}^{k-1} n_i(a)\, \boldsymbol{d}_{AO^2}^{(a,n_i,n_i(a))} - \frac{1}{N_k(a)} \sum_{i=0}^{k-1} n_i(a)\, \widehat{\boldsymbol{d}}_{AO^2}^{(a,n_i,n_i(a))} \right\|_2 \\
&= \left\| \frac{1}{N_k(a)} \sum_{i=0}^{k-1} n_i(a) \left(\boldsymbol{d}_{AO^2}^{(a,n_i,n_i(a))} - \widehat{\boldsymbol{d}}_{AO^2}^{(a,n_i,n_i(a))} \right) \right\|_2 \\
&\leqslant \frac{1}{N_k(a)} \sum_{i=0}^{k-1} n_i(a) \left\| \boldsymbol{d}_{AO^2}^{(a,n_i,n_i(a))} - \widehat{\boldsymbol{d}}_{AO^2}^{(a,n_i,n_i(a))} \right\|_2 \qquad (17) \\
&\leqslant \frac{1}{N_k(a)} \sum_{i=0}^{k-1} n_i(a)\, \frac{G(\pi_i)}{1-\eta(\pi_i)} \sqrt{\frac{8 \log\left(2AO^2 k/\delta \right)}{n_i(a)}}, \qquad (18)
\end{aligned}$$

where in line 17 we applied the triangle inequality, while in line 18 we bound the estimation error made on each distribution $\boldsymbol{d}_{AO^2}^{(a,n_i,n_i(a))}$ with probability $1-\delta/k$ and apply the union bound.
Let us define the new values $\widetilde{G} \coloneqq \max_i G(\pi_i)$ and $\widetilde{\eta} \coloneqq \min_i \eta(\pi_i)$. This allows us to proceed as follows:

$$\begin{aligned}
\left\|\boldsymbol{d}_{AO^2}^{(a,k)} - \boldsymbol{d}_{AO^2}^{(a,k)}\right\|_2 &\leqslant \frac{1}{N_k(a)} \sum_{i=0}^{k-1} n_i(a)\, \frac{G(\pi_i)}{1-\eta(\pi_i)} \sqrt{\frac{8\log\left(2AO^2k/\delta\right)}{n_i(a)}} \\
&\leqslant \frac{\widetilde{G}}{N_k(a)\,(1-\widetilde{\eta})} \sqrt{8\log\left(2AO^2k/\delta\right)} \sum_{i=0}^{k-1} n_i(a) \sqrt{\frac{1}{n_i(a)}} \qquad (19) \\
&= \frac{\widetilde{G}}{N_k(a)\,(1-\widetilde{\eta})} \sqrt{8\log\left(2AO^2k/\delta\right)} \sum_{i=0}^{k-1} \sqrt{n_i(a)} \\
&\leqslant \frac{\widetilde{G}}{(1-\widetilde{\eta})} \sqrt{\frac{8k\log\left(2AO^2k/\delta\right)}{N_k(a)}} \qquad (20)
\end{aligned}$$

where in line 19 we bound the singular terms in the summation using the definition of $\widetilde{G}$ and $\widetilde{\eta}$ and bring them out of the sum, line 20 is instead obtained by using the definition $N_k(a) = \sum_{i=0}^{k-1} n_i(a)$ and the Cauchy-Schwartz inequality for which it holds that $\sum_{i=0}^{k-1} \sqrt{n_i(a)} \leqslant \sqrt{k \sum_{i=0}^{k-1} n_i(a)}$.
The expression obtained shows that the error of the combined estimator pays a further term $\sqrt{k}$ in the bound, but it scales with $\mathcal{O}(1/\sqrt{N_k(a)})$, analogously as per the single-policy estimator $\widehat{\boldsymbol{d}}_{AO^2}^{(a,n,n_i(a))}$.

Second Part of the Proof

Let us now focus on the second part. This part shares some similarities with the proof of Lemma 5.2 appearing in Russo et al. (2024a). However, the results are applied to different quantities since (i) the distribution employed here is conditioned on a specific action a, (ii) this distribution is defined with respect to a finite amount of samples and (iii) this distribution combines samples coming from different policies.

Having defined a concentration result on the combined estimator $\widehat{\boldsymbol{d}}^{(a,k)}$, we proceed as follows:

$$\begin{aligned}
\left\|\boldsymbol{d}_{AS^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{AS^2}^{(a,k)}\right\|_2 &= \left\|\mathbb{B}_a^{\dagger} \left(\boldsymbol{d}_{AO^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{AO^2}^{(a,k)}\right)\right\|_2 \\
&\leqslant \left\|\mathbb{B}_a^{\dagger}\right\|_2 \left\|\boldsymbol{d}_{AO^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{AO^2}^{(a,k)}\right\|_2 \\
&= \frac{1}{\sigma_{\min}(\mathbb{B}_a)} \left\|\boldsymbol{d}_{AO^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{AO^2}^{(a,k)}\right\|_2 \qquad (21) \\
&\leqslant \frac{1}{\alpha^2} \left\|\boldsymbol{d}_{AO^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{AO^2}^{(a,k)}\right\|_2, \qquad (22)
\end{aligned}$$

where the first equality can be directly derived from Equation (10), while the first inequality follows by the consistency property of matrices. The last inequality derives from the definition of the block diagonal matrix $\mathbb{B}_a$, which is composed of submatrices $\{\mathbb{O}_{a,a'}\}_{a' \in \mathcal{A}}$ for which it holds that $\sigma_{\min}(\mathbb{O}_{a,a'}) \geqslant \alpha^2$ for all $(a,a') \in \mathcal{A}^2$. For the properties of block diagonal matrices, it also follows that $\sigma_{\min}(\mathbb{B}_a) \geqslant \alpha^2$. Combining this last result with the one in (20), we get with probability at least $1-\delta$:

$$\left\|\boldsymbol{d}_{AS^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{AS^2}^{(a,k)}\right\|_2 \leqslant \frac{\widetilde{G}}{\alpha^2\,(1-\widetilde{\eta})} \sqrt{\frac{8k\log\left(2AO^2k/\delta\right)}{N_k(a)}} \qquad (23)$$

Following the steps of the *Action-wise* OAS estimation algorithm (Algorithm 1), the estimated vector $\widehat{\boldsymbol{d}}_{AS^2}^{(a,k)} \in \mathbb{R}^{AS^2}$ is aggregated into the new vector $\widehat{\boldsymbol{d}}_{S^2}^{(a,k)} \in \mathbb{R}^{S^2}$ such that:

$$\widehat{\boldsymbol{d}}_{S^2}^{(a,k)}(s,s') = \sum_{a' \in \mathcal{A}} \widehat{\boldsymbol{d}}_{AS^2}^{(a,k)}(a',s,s') \qquad \forall s,s' \in \mathcal{S}. \qquad (24)$$

Making use of the Aggregation Lemma appearing in Lemma E.3, it is possible to show that the following holds:

$$\left\| \boldsymbol{d}_{S^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{S^2}^{(a,k)} \right\|_2 \leqslant \sqrt{A} \, \left\| \boldsymbol{d}_{AS^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{AS^2}^{(a,k)} \right\|_2 . \tag{25}$$

The next step in the algorithm requires to set to 0 all the negative elements appearing in vector $\widehat{\boldsymbol{d}}_{S^2}^{(a,k)}$. By Assumption 4.1, it can easily be observed that all the elements appearing in the real quantity $\boldsymbol{d}_{S^2}^{(a,k)}$ are positive. For this reason, the elements of the non-negative vector $\bar{\boldsymbol{d}}_{S^2}^{(a,k)}$ obtained by transforming $\widehat{\boldsymbol{d}}_{S^2}^{(a,k)}$ are closer to the real quantities contained in $\boldsymbol{d}_{S^2}^{(a,k)}$. We can thus see that:

$$\left\| \boldsymbol{d}_{S^2}^{(a,k)} - \bar{\boldsymbol{d}}_{S^2}^{(a,k)} \right\|_2 \leqslant \left\| \boldsymbol{d}_{S^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{S^2}^{(a,k)} \right\|_2 . \tag{26}$$

For what follows, we will use notation $\boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot) \in \mathbb{R}^S$ to denote the subvector of dimension S containing the quantities $\boldsymbol{d}_{S^2}^{(a,k)}(s,s')$ for each $s' \in \mathcal{S}$.

We can then write:

$$\begin{aligned}
\|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F &= \sqrt{\sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} \left(\mathbb{T}_a(s,s') - \widehat{\mathbb{T}}_a(s,s') \right)^2} = \sqrt{\sum_{s\in\mathcal{S}} \left\| \mathbb{T}_a(s,\cdot) - \widehat{\mathbb{T}}_a(s,\cdot) \right\|_2^2} \\
&= \sqrt{\sum_{s\in\mathcal{S}} \left\| \frac{\boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot)}{\left\| \boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot) \right\|_1} - \frac{\bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot)}{\left\| \bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot) \right\|_1} \right\|_2^2} && (27) \\
&\leqslant \sqrt{\sum_{s\in\mathcal{S}} \left\| \frac{\boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot)}{\left\| \boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot) \right\|_2} - \frac{\bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot)}{\left\| \bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot) \right\|_2} \right\|_2^2} && (28) \\
&\leqslant \sqrt{\sum_{s\in\mathcal{S}} \frac{4 \left\| \boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot) - \bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot) \right\|_2^2}{\max\left\{ \left\| \boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot) \right\|_2, \left\| \bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot) \right\|_2 \right\}^2}} && (29) \\
&\leqslant \sqrt{\sum_{s\in\mathcal{S}} \frac{4 \left\| \boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot) - \bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot) \right\|_2^2}{\left\| \boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot) \right\|_2^2}} \\
&\leqslant \sqrt{\sum_{s\in\mathcal{S}} \frac{4S \left\| \boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot) - \bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot) \right\|_2^2}{\left(d_{\min}^{(a)} \right)^2}} && (30) \\
&= \sqrt{\frac{4S \left\| \boldsymbol{d}_{S^2}^{(a,k)} - \bar{\boldsymbol{d}}_{S^2}^{(a,k)} \right\|_2^2}{\left(d_{\min}^{(a)} \right)^2}} && (31) \\
&= \frac{2\sqrt{S} \left\| \boldsymbol{d}_{S^2}^{(a,k)} - \bar{\boldsymbol{d}}_{S^2}^{(a,k)} \right\|_2}{d_{\min}^{(a)}} \\
&\leqslant \frac{2\sqrt{S} \left\| \boldsymbol{d}_{S^2}^{(a,k)} - \widehat{\boldsymbol{d}}_{S^2}^{(a,k)} \right\|_2}{d_{\min}^{(a)}}. && (32)
\end{aligned}$$

Equality in line 27 holds for the definition of the estimated matrix $\widehat{\mathbb{T}}_a$[9]. The first inequality in line 28 derives from the relation between norms $\left\| \bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot) \right\|_1 \geqslant \left\| \bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot) \right\|_2$, while line 29 follows from Lemma E.2.

[9]We assume here that the estimated vectors are such that $\|\bar{d}_{S^2}^{(a)}(s,\cdot)\|_1 \neq 0$. However, if this is not the case, instead of

Line 30 follows from Lemma E.5 with $d_{\min}^{(a)}$ representing the minimum state probability conditioned on action a, which is bounded away from 0 thanks to Assumption 4.1. The equality in line 31 is simply obtained by observing that:

$$\sum_{s\in\mathcal{S}} \|\boldsymbol{d}_{S^2}^{(a,k)}(s,\cdot) - \bar{\boldsymbol{d}}_{S^2}^{(a,k)}(s,\cdot)\|_2^2 = \|\boldsymbol{d}_{S^2}^{(a,k)} - \bar{\boldsymbol{d}}_{S^2}^{(a,k)}\|_2^2,$$

holding by the definition of $\boldsymbol{d}_{S^2}^{(a,k)}$ and $\widehat{\boldsymbol{d}}_{S^2}^{(a,k)}$ respectively, while the last inequality simply uses the bound in (26).

By combining the results obtained in (23), Equation (25), and (32), we obtain the final result holding with probability $1-\delta$:

$$\|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F \leqslant \frac{4\,\widetilde{G}}{\alpha^2 d_{\min}^{(a)}(1-\widetilde{\eta})}\sqrt{\frac{2k\ SA\ \log\left(2AO^2k/\delta\right)}{N_k(a)}},$$

which completes the proof. □

C PROOF OF THEOREM 6.1

In this section, we will provide the proof for Theorem 6.1. This theorem makes use of the result in Lemma 5.1 which shows convergence results for samples collected under different policies. The main steps of the proof share similarities with the proof of Theorem 6.1 of Russo et al. (2024a). In particular, we improve over that result by (i) adopting the new *Action-wise* OAS estimator which is able to reuse samples from different episodes, and by (ii) providing a tighter concentration result on the belief error (Lemma D.1).

Notation and Useful Quantities

Before proceeding with the proof, we will need to define some notation that will be useful for what will follow. We define the expected reward of an action a_t assuming to be in state s_t as:

$$\mu(s_t,a_t) = \sum_{o\in\mathcal{O}} r(o)\mathbb{O}_{a_t}(o|s_t) = \boldsymbol{r}^\top \mathbb{O}_{a_t}(\cdot|s_t).$$

Therefore, we can define the expected reward given a belief state b_t at time t when taking action a_t as:

$$g(b_t,a_t) = \sum_{s\in\mathcal{S}} \mu(s,a_t)b_t(s) = \boldsymbol{\mu}(a_t)b_t = \boldsymbol{r}^\top \mathbb{O}_{a_t} b_t, \tag{33}$$

where the last equalities define the expression in matrix notation, with $\boldsymbol{\mu}(a_t)$ being a vector of dimension S containing the quantity $\mu(s,a_t)\ \forall s\in\mathcal{S}$.

We will use $\mathbb{T} = \{\mathbb{T}_a\}_{a\in\mathcal{A}}$ to denote the real transition model and $\mathcal{Q}$ to denote the real POMDP instance.
We will employ $\widehat{\mathbb{T}}_k = \{\widehat{\mathbb{T}}_{a,k}\}$ to denote the transition model estimated by the *Action-wise* OAS estimation procedure at the beginning of episode k, while we will use $\mathbb{T}_k = \{\mathbb{T}_{a,k}\}$ to denote the optimistic transition model returned as output by the oracle and actually used during episode k. Analogously, we will denote the estimated and the optimistic POMDP instances at episode k with $\widehat{\mathcal{Q}}_k$ and $\mathcal{Q}_k$ respectively.

We will denote with t_k the starting time of episode k and each episode k will be thus characterized by the timestamps $[t_k,\, t_k+1,\ldots,\, t_{k+1}-1]$ with $t_{k+1}-1$ defining the last timestamp of episode k. For convenience, during the analysis, we will use variable E_k to characterize the interval associated with the timestamps of episode k from which we exclude the last timestamp of the episode, namely $E_k \coloneqq [t_k,\, t_k+1,\ldots,\, t_{k+1}-2]$. The last sample of each episode k is excluded from E_k since it is not entirely used for estimation by the *Action-wise* OAS procedure.

We will also define a probability distribution defined on the belief space as:

$$U(b_{t+1}|b_t,a) = P_{\mathcal{Q}}(b_{t+1}|b_t,a)$$

nullifying the negative terms of vector $\widehat{\boldsymbol{d}}_{S^2}^{(a)}$, we could simply make the terms positive by a small amount and the result in Equation (26) would still hold.

where the probability is defined with respect to the transition model $\mathbb{T}$ and the observation model $\mathbb{O}$ referred to the POMDP $\mathcal{Q}$. We will use U_k to denote a probability distribution defined with respect to the optimistic POMDP $\mathcal{Q}_k$.

Having defined the used notation, we start by reporting the main result of the Theorem here.

Theorem 6.1. *Let us assume to have a POMDP instance $\mathcal{Q}$ satisfying Assumptions 4.1 and 4.2. If the Action-wise OAS-UCRL algorithm is run for T steps, with probability at least $1-2\delta$, it suffers from a total regret:*

$$\mathcal{R}_T \leqslant \mathcal{O}\left(\frac{CD\widetilde{G}}{\alpha^2 \widetilde{d}_{\min}}\sqrt{SA^3T\,\log T\,\log O}\right).$$

where $C \coloneqq \frac{4(1-\epsilon)^3}{\epsilon^4}$ and D is a finite constant bounding the span of the bias function (definition in Proposition E.1).

Proof. We recall here the definition of regret as reported in (3):

$$\mathcal{R}_T := T\rho^* - \sum_{t=0}^{T-1} r(o_t) = \sum_{t=0}^{T-1}(\rho^* - \mathbb{E}^{\pi}[r(O_t)|\mathcal{F}_{t-1}]) + \sum_{t=0}^{T-1}(\mathbb{E}^{\pi}[r(O_t)|\mathcal{F}_{t-1}] - r(o_t)), \tag{34}$$

where we consider an expectation $\mathbb{E}^{\pi}$ taken w.r.t. the true transition model $\mathbb{T} = \{\mathbb{T}_a\}_{a\in\mathcal{A}}$ and the true observation model $\mathbb{O} = \{\mathbb{O}_a\}_{a\in\mathcal{A}}$. We use $\mathcal{F}_{t-1}$ to denote the filtration defined with respect to the events occurring up to time $t-1$. The second term in the summation defines a martingale. Indeed, by denoting a stochastic process as:

$$X_0 = 0,\ X_t = \sum_{l=0}^{t-1}(\mathbb{E}^{\pi}[r(O_l)|\mathcal{F}_{l-1}] - r(o_l)),$$

we can easily see that X_t represents a martingale. Thus, by applying the Azuma-Hoeffding inequality (Azuma, 1967) we have that with probability at least $1-\delta/4$ we have:

$$\sum_{t=0}^{T-1}(\mathbb{E}^{\pi}[r(O_t)|\mathcal{F}_{t-1}] - r(o_t)) \leqslant \sqrt{2T\ln(4/\delta)}. \tag{35}$$

Since action A_t is adapted to the filtration $\mathcal{F}_{t-1}$, we have:

$$\mathbb{E}^{\pi}[\mu(S_t, A_t)|\mathcal{F}_{t-1}] = g(b_t, A_t),$$

where function $g(\cdot,\cdot)$ is defined in Equation 33, while the belief b_t is computed using the true transition and observation matrices and actions are taken according to policy π. Using analogous notation, we will denote the expected instantaneous reward assuming to have computed the belief using the estimated transition probability $\mathbb{T}_{a,k}$ as:

$$\mathbb{E}^{\pi}_k[\mu(S_t, A_t)|\mathcal{F}_{t-1}] = g(b^k_t, A_t).$$

Given the defined quantities, we can rewrite the first term of (34) as:

$$\sum_{t=0}^{T-1}(\rho^* - \mathbb{E}^{\pi}[r(O_t)|\mathcal{F}_{t-1}]) = \sum_{t=0}^{T-1}(\rho^* - \mathbb{E}^{\pi}[\mu(S_t, A_t)|\mathcal{F}_{t-1}]) = \sum_{t=0}^{T-1}(\rho^* - g(b_t, A_t)). \tag{36}$$

By following the procedure described in the *Action-wise* OAS-UCRL algorithm, at the beginning of each episode k, an optimistic POMDP $\mathcal{Q}_k$ is chosen from the set of possible POMDPs determined by the confidence region $\mathcal{C}_k(\delta_k)$. We recall that the optimistic POMDP $\mathcal{Q}_k$ is defined by the optimistic transition model $\mathbb{T}_k = \{\mathbb{T}_{a,k}\}_{a\in\mathcal{A}}$ provided by the oracle, and the real observation model.
The confidence region $\mathcal{C}_k(\delta_k)$ of the transition model in episode k can be associated with the confidence regions $\mathcal{C}_{a,k}(\delta_{a,k})$ of each action transition model. Each confidence region $\mathcal{C}_{a,k}(\delta_{a,k})$ is centered in the estimated action transition matrix $\widehat{\mathbb{T}}_{a,k}$ and is such that $P(\mathbb{T}_a \in \mathcal{C}_{a,k}(\delta_{a,k})) \geqslant 1-\delta_{a,k}$.

Now we consider two possible events: the *good event* which considers the case where for all episodes k, the true POMDP is contained in the confidence sets $\mathcal{C}_k(\delta_k)$ and the *failure event* which denotes the complementary event. The *good event* implies that all the real action transition models $\mathbb{T}_a$ are contained in their confidence region $\mathcal{C}_{a,k}(\delta_{a,k})$ for all episodes k.
We set the confidence level of the transition model in episode k as $\delta_k \coloneqq \frac{\delta}{k^3}$ and set the confidence level of each action transition model in episode k as $\delta_{a,k} \coloneqq \frac{\delta}{Ak^3}$.
We can thus bound the probability of the *failure event* as:

$$\begin{aligned} P(\mathcal{Q} \notin \mathcal{C}_k(\delta_k), \text{for some k}) &= P(\mathbb{T}_a \notin \mathcal{C}_{a,k}(\delta_{a,k}), \text{for some } a,\, k) \\ &\leqslant \sum_{k=1}^{K-1} \sum_{a \in \mathcal{A}} \delta_{a,k} = \sum_{k=1}^{K} \underbrace{A \frac{\delta}{Ak^3}}_{\delta_k} \leqslant \frac{3}{2}\delta, \end{aligned}$$

From this formulation, it appears that the *good event* holds with probability at least $1 - \frac{3}{2}\delta$. When this is the case, we have that $\rho^* \leqslant \rho^k$ for any k since the optimal average reward is taken from the optimistic POMDP $\mathcal{Q}_k$.

We can now bound the regret under the *good event* during the different K episodes as:

$$\begin{aligned} \sum_{t=0}^{T-1} (\rho^* - g(b_t, A_t)) &\leqslant K + \sum_{k=0}^{K-1} \sum_{t \in E_k} (\rho^* - g(b_t, A_t)) \\ &\leqslant K + (T_0 - 1) + \sum_{k=1}^{K-1} \sum_{t \in E_k} \left(\rho^k - g(b_t, A_t)\right) \\ &= K + \sum_{a \in \mathcal{A}} n_0(a) + \sum_{k=1}^{K-1} \sum_{t \in E_k} \underbrace{\left[\rho^k - g(b_t^k, A_t)\right]}_{\text{First Term}} + \underbrace{\left[g(b_t^k, A_t) - g(b_t, A_t)\right]}_{\text{Second Term}}, \end{aligned} \tag{37}$$

where we have rewritten the summation by highlighting the different episodes K. In particular, for each episode k, we use interval E_k which excludes the last timestamp of that episode, and the term K appearing in the first inequality is obtained by assuming to pay maximum regret for each excluded sample.
In the second inequality instead, we make explicit the length of the first episode T_0 for which we assume to pay maximum regret and from which we subtract 1 (which is the last sample of the episode already counted in the K term). In the last equality, we rewrite the length of the first episode as the sum of counts of the chosen actions. For the moment, we will not consider the terms K and $\sum_{a \in \mathcal{A}} n_0(a)$ but we will focus on the different terms appearing in the summation.

Analysis of the First Term in 37

As a first step, we will consider the first term appearing in the summation in 37. It can be bounded by using the Bellman equation reported in Equation (2) for the optimistic belief MDP. By using the probability distribution U defined on the next belief (see Notation section), we can rewrite the Bellman equation as follows:

$$\begin{aligned} \rho^k + v_k(b_t^k) &= g(b_t^k, A_t) + \int_{b_{t+1} \in \mathcal{B}} v_k(b_{t+1}) U_k(\, db_{t+1} | b_t^k, A_t) \\ &= g(b_t^k, A_t) + \langle U_k(\cdot | b_t^k, A_t), v_k(\cdot) \rangle. \end{aligned}$$

Given that the value function v_k satisfies the Bellman Equation, a shifted version $v_k + c\mathbf{1}$ of the bias function would satisfy it as well. From this consideration, we can assume that $\|v_k\|_\infty \leqslant \text{span}(v_k)/2$. By using the result in Proposition E.1 reported in Zhou et al. (2021), we are able to bound the span of v_k, where the span is defined as $span(v_k) := \max_{b \in \mathcal{B}} v_k(b) - \min_{b \in \mathcal{B}} v_k(b)$. In particular, we use the finite constant D to bound the span. From these considerations, we can write:

$$\|v_k\|_\infty \leqslant \frac{\text{span}(v_k)}{2} \leqslant \frac{D}{2}. \tag{38}$$

By combining the elements reported so far, for the first term in the summation of 37, we can write:

$$\sum_{k=1}^{K-1}\sum_{t\in E_k}(\rho^k - g(b_t^k, A_t)) = \sum_{k=1}^{K-1}\sum_{t\in E_k}\left(-v_k(b_t^k) + \langle U_k(\cdot|b_t^k, A_t), v_k(\cdot)\rangle\right)$$
$$= \sum_{k=1}^{K-1}\sum_{t\in E_k}\left(-v_k(b_t^k) + \langle U(\cdot|b_t^k, A_t), v_k(\cdot)\rangle\right) + \left(\langle U_k(\cdot|b_t^k, A_t) - U(\cdot|b_t^k, A_t), v_k(\cdot)\rangle\right), \quad (39)$$

where the first equality is obtained from the Bellman Equation, while the last equality is obtained by adding and subtracting the term $\langle U(\cdot|b_t^k, A_t), v_k(\cdot)\rangle$ for each time step t and we recall that $U(\cdot|b_t^k, A_t)$ represents the probability distribution over the belief at the next step $t+1$ under the true POMDP instance $\mathcal{Q}$, while $U_k(\cdot|b_t^k, A_t)$ represents this probability distribution under the optimistic instance $\mathcal{Q}_k$.

For the first term of 39, we have:

$$\sum_{k=1}^{K-1}\sum_{t\in E_k}\left(-v_k(b_t^k) + \langle U(\cdot|b_t^k, A_t), v_k(\cdot)\rangle\right) = \sum_{k=1}^{K-1}\sum_{t\in E_k}\left(-v_k(b_t^k) + v_k(b_{t+1}^k)\right) + \left(-v_k(b_{t+1}^k) + \langle U(\cdot|b_t^k, A_t), v_k(\cdot)\rangle\right)$$
$$= \sum_{k=1}^{K-1}\Big(-v_k(b_{s_k}^k) + v_k(b_{e_k+1}^k)\Big) + \sum_{k=1}^{K-1}\sum_{t\in E_k}\mathbb{E}^\pi[v_k(b_{t+1}^k|\mathcal{F}_t)] - v_k(b_{t+1}^k),$$

where the first term appearing in the last equality represents a telescopic summation. For each episode k, the terms appearing in this summation are respectively the value of the bias function of the belief $b_{s_k}^k$ (with s_k denoting the starting time step of episode k) and the value of the bias function of the belief $b_{e_k+1}^k$ (with e_k denoting the last time step of E_k).
The second term appearing in the last equality is instead obtained by showing that:

$$\langle U(\cdot|b_t^k, A_t), v_k(\cdot)\rangle = \int_{b_{t+1}\in\mathcal{B}} v_k(b_{t+1})U(\,db_{t+1}|b_t^k, A_t) = \mathbb{E}^\pi[v_k(b_{t+1}^k|b_t^k)] = \mathbb{E}^\pi[v_k(b_{t+1}^k|\mathcal{F}_t)].$$

By recalling Proposition E.1 to bound the span of the bias function, we can easily see that:

$$\sum_{k=1}^{K-1} -v_k(b_{s_k}^k) + v_k(b_{e_k+1}^k) \leqslant \sum_{k=1}^{K-1} D = (K-1)\,D. \quad (40)$$

By applying analogous considerations as those used for bounding 35, we can state that this sum of differences defines a martingale. Thus, with probability at least $1-\delta/4$, we have that:

$$\sum_{k=1}^{K-1}\sum_{t\in E_k}\mathbb{E}^\pi[v_k(b_{t+1}^k|\mathcal{F}_t)] - v_k(b_{t+1}^k) \leqslant D\sqrt{2T\ln\left(\frac{4}{\delta}\right)}. \quad (41)$$

By combining the previous considerations, we can bound the first term in 39 as:

$$\sum_{k=1}^{K-1}\sum_{t\in E_k}\left(-v_k(b_t^k) + \langle U(\cdot|b_t^k, A_t), v_k(\cdot)\rangle\right) \leqslant (K-1)\,D + D\sqrt{2T\ln\left(\frac{4}{\delta}\right)}. \quad (42)$$

We can now proceed in bounding the second term appearing in 39. Before going on with this step, we need to introduce functions $H(b_t, a_t, o_t)$ and $H_k(b_t, a_t, o_t)$ which return the belief at the next time step b_{t+1} given the current belief b_t, the action taken a_t and the received observation o_t using the real $\mathbb{T}_a$ and the optimistic transition matrix $\mathbb{T}_{a,k}$, respectively.

By analyzing each term appearing in the second summation of 39, we get:

$$\begin{aligned}
\langle U_k(\cdot|b_t^k, A_t) - U(\cdot|b_t^k, A_t), v_k(\cdot)\rangle &\leqslant \left| \int_{\mathcal{B}} v_k(b') U_k(\, db'|b_t^k, A_t) - \int_{\mathcal{B}} v_k(b') U(\, db'|b_t^k, A_t) \right| \\
&= \left| \sum_{o_t \in \mathcal{O}} v_k\left(H_k(b_t^k, A_t, o_t)\right) P(o_t|b_t^k, A_t) - \sum_{o_t \in \mathcal{O}} v_k\left(H(b_t^k, A_t, o_t)\right) P(o_t|b_t^k, A_t) \right| \\
&= \left| \sum_{o_t \in \mathcal{O}} \left[v_k\left(H_k(b_t^k, A_t, o_t)\right) - v_k\left(H(b_t^k, A_t, o_t)\right) \right] P(o_t|b_t^k, A_t) \right| \\
&\leqslant \sum_{o_t \in \mathcal{O}} \left| v_k(H_k(b_t^k, A_t, o_t)) - v_k(H(b_t^k, A_t, o_t)) \right| P(o_t|b_t^k, A_t) \\
&\leqslant \sum_{o_t \in \mathcal{O}} \frac{D}{2} \left| H_k(b_t^k, A_t, o_t) - H(b_t^k, A_t, o_t) \right| P(o_t|b_t^k, A_t) \\
&\leqslant \sum_{o_t \in \mathcal{O}} \frac{D}{2} \Big(L_1 \left\| \mathbb{T}_{A_t} - \mathbb{T}_{A_t,k} \right\|_F \Big) P(o_t|b_t^k, A_t) \qquad \text{(Corollary D.2)} \\
&= \frac{DL_1}{2} \| \mathbb{T}_{A_t} - \mathbb{T}_{A_t,k} \|_F, \qquad (43)
\end{aligned}$$

where in the first equality we have explicitly decoupled the stochasticity induced by the observation from the deterministic update of the belief b' at the next step through the H and H_k functions. The second inequality is simply obtained by using the triangle inequality, while the third inequality is obtained using the bound on the bias span appearing in 38. The last inequality is instead obtained from Corollary D.2 bounding the one-step error of the belief vector updated using different transition matrices. Here, we introduce constant $L_1 = \frac{4(1-\epsilon)}{\epsilon^2}$ derived from the corollary.

By combining the results obtained so far in 42 and 43, we are able to bound the first term appearing in the summation of 37 as:

$$\begin{aligned}
\sum_{k=1}^{K-1} \sum_{t \in E_k} (\rho^k - g(b_t^k, A_t)) &\leqslant (K-1)\, D + D\sqrt{2T \ln\left(\frac{4}{\delta}\right)} + \sum_{k=1}^{K-1} \sum_{t \in E_k} \frac{DL_1}{2} \| \mathbb{T}_{A_t} - \mathbb{T}_{A_t,k} \|_F \\
&= (K-1)\, D + D\sqrt{2T \ln\left(\frac{4}{\delta}\right)} + \frac{DL_1}{2} \sum_{k=1}^{K-1} \sum_{a \in \mathcal{A}} n_k(a) \| \mathbb{T}_a - \mathbb{T}_{a,k} \|_F. \qquad (44)
\end{aligned}$$

Analysis of the Second Term in 37

We can now focus on the second term appearing in the summation of 37. We have that:

$$\begin{aligned}
\sum_{k=1}^{K-1} \sum_{t \in E_k} (g(b_t^k, A_t) - g(b_t, A_t)) &\leqslant \sum_{k=1}^{K-1} \sum_{t \in E_k} \| \boldsymbol{r}^\top \mathbb{O}_{A_t} \|_\infty \| b_t^k - b_t \|_1 \\
&\leqslant \sum_{k=1}^{K-1} \sum_{t \in E_k} \| b_t^k - b_t \|_1, \qquad (45)
\end{aligned}$$

where we use Holder's inequality in the first passage, while the second inequality considers that $\| \boldsymbol{r}^\top \mathbb{O}_{A_t} \|_\infty \leqslant 1$.

The expression we use here to bound line 45 uses a new result which we present in Lemma D.1 that improves over the result employed in Russo et al. (2024a) (see their Proposition H.3).

In particular, Lemma D.1 show that:

$$\sum_{k=1}^{K-1}\sum_{t\in E_k}\|b_t^k - b_t\|_1 \leqslant \sum_{k=1}^{K-1}\left[L + L\sum_{a\in\mathcal{A}} n_k(a)\|\mathbb{T}_a - \mathbb{T}_{a,k}\|_F\right]$$
$$= (K-1)L + L\sum_{k=1}^{K-1}\sum_{a\in\mathcal{A}} n_k(a)\|\mathbb{T}_a - \mathbb{T}_{a,k}\|_F, \tag{46}$$

with constant $L \coloneqq \frac{4(1-\epsilon)^2}{\epsilon^3}$ defined in the lemma.

Merge of Obtained Results to Bound Line 37

By merging the results obtained in 44 and in 46, we bound line 37 as follows:

$$\sum_{t=0}^{T-1}(\rho^* - g(b_t, A_t)) \leqslant K + \sum_{a\in\mathcal{A}} n_0(a) + (K-1)(D+L) + D\sqrt{2T\ln\left(\frac{4}{\delta}\right)} +$$
$$+ \sum_{k=1}^{K-1}\sum_{a\in\mathcal{A}} n_k(a)\left(\frac{DL_1}{2}\|\mathbb{T}_a - \mathbb{T}_{a,k}\|_F + L\|\mathbb{T}_a - \mathbb{T}_{a,k}\|_F\right)$$
$$\leqslant K + \sum_{a\in\mathcal{A}} n_0(a) + (K-1)(D+L) + D\sqrt{2T\ln\left(\frac{4}{\delta}\right)} + \frac{L(2+D)}{2}\sum_{k=1}^{K-1}\sum_{a\in\mathcal{A}} n_k(a)\,\|\mathbb{T}_a - \mathbb{T}_{a,k}\|_F \tag{47}$$

where in the last inequality we used $L_1 \leqslant L$.

Let us now consider the last quantity appearing in line 47 and let us disregard for the moment the multiplicative part $L(2+D)/2$. We proceed with the analysis:

$$\sum_{k=1}^{K-1}\sum_{a\in\mathcal{A}} n_k(a)\,\|\mathbb{T}_a - \mathbb{T}_{a,k}\|_F \leqslant \sum_{k=1}^{K-1}\sum_{a\in\mathcal{A}} \frac{4\,\widetilde{G}\,n_k(a)}{\alpha^2 d_{\min}^{(a,k)}(1-\widetilde{\eta})}\sqrt{\frac{2kSA\,\log\left(2AO^2k/\delta_{a,k}\right)}{N_k(a)}}$$
$$= \sum_{k=1}^{K-1}\sum_{a\in\mathcal{A}} \frac{4\,\widetilde{G}\,n_k(a)}{\alpha^2 d_{\min}^{(a,k)}(1-\widetilde{\eta})}\sqrt{\frac{2kSA\,\log\left(2A^2O^2k^4/\delta\right)}{N_k(a)}}$$
$$\leqslant \frac{4\,\widetilde{G}}{\alpha^2\widetilde{d}_{\min}(1-\widetilde{\eta})}\sqrt{2KSA\,\log\left(\frac{2A^2O^2K^4}{\delta}\right)}\sum_{k=1}^{K-1}\sum_{a\in\mathcal{A}}\frac{n_k(a)}{\sqrt{N_k(a)}}, \tag{48}$$

where the first inequality holds by using Lemma 5.1 and recalling that we are under the *good event*. In the first equality, we make explicit the confidence level $\delta_{a,k} = \frac{\delta}{Ak^3}$ while in the last expression, we use $k \leqslant K$ and we set $\widetilde{d}_{\min} \coloneqq \min_k \min_{a\in\mathcal{A}} d_{\min}^{(a,k)}$, which is always bounded away from 0 because of Assumption 4.1.

Let us now focus on the summation appearing on the right side of the bound in 48. At this point, we can also include the action counts associated with episode 0. It follows that:

$$\sum_{a\in\mathcal{A}} n_0(a) + \sum_{a\in\mathcal{A}}\sum_{k=1}^{K-1}\frac{n_k(a)}{\sqrt{N_k(a)}} = \sum_{a\in\mathcal{A}}\sum_{k=0}^{K-1}\frac{n_k(a)}{\sqrt{\max\{1, N_k(a)\}}}$$
$$\leqslant \sum_{a\in\mathcal{A}}\left(\sqrt{2}+1\right)\sqrt{N_K(a)} \qquad \text{(Lemma E.4)}$$
$$\leqslant \left(\sqrt{2}+1\right)\sqrt{AT}, \qquad \text{(Cauchy-Schwarz inequality)}$$

where in the first equality we bring the summation on the terms $n_0(a)$ in the summation over the episodes by including the max at the denominator. The successive inequality is due to Lemma E.4 taken from Jaksch et al. (2010), while the last expression is simply obtained by the Cauchy-Schwarz inequality and noting that $\sum_{a\in\mathcal{A}} N_K(a) = T - K \leqslant T$.

We are now able to rewrite the bound in 47 as:

$$\begin{aligned}\sum_{t=0}^{T-1}(\rho^* - g(b_t, A_t)) &\leqslant K + (K-1)(D+L) + D\sqrt{2T\ln\left(\frac{4}{\delta}\right)} + \\ &+ (\sqrt{2}+1)\frac{2L(2+D)\,\widetilde{G}}{\alpha^2\widetilde{d}_{\min}(1-\widetilde{\eta})}\sqrt{2KSA^2T\,\log\left(\frac{2A^2O^2K^4}{\delta}\right)} \\ &\leqslant 2K(D+L) + D\sqrt{2T\ln\left(\frac{4}{\delta}\right)} + \frac{6L(2+D)\,\widetilde{G}}{\alpha^2\widetilde{d}_{\min}(1-\widetilde{\eta})}\sqrt{2KSA^2T\,\log\left(\frac{2A^2O^2K^4}{\delta}\right)} \qquad (49)\end{aligned}$$

Final Regret Result

By recalling the definition of the regret in line 34, we are finally able to combine the result appearing in line 49 and the result on the martingale in line 35 using a union bound. Indeed, with probability at least $1-2\delta$, we have:

$$\begin{aligned}\mathcal{R}_T &\leqslant 2K(D+L) + D\sqrt{2T\ln\left(\frac{4}{\delta}\right)} + \sqrt{2T\ln\left(\frac{4}{\delta}\right)} + \frac{6L(2+D)\,\widetilde{G}}{\alpha^2\widetilde{d}_{\min}(1-\widetilde{\eta})}\sqrt{2KSA^2T\,\log\left(\frac{2A^2O^2K^4}{\delta}\right)} \\ &\leqslant 2K(D+L) + 2D\sqrt{2T\ln\left(\frac{4}{\delta}\right)} + \frac{6\,C\,(2+D)\,\widetilde{G}}{\alpha^2\widetilde{d}_{\min}}\sqrt{2KSA^2T\,\log\left(\frac{2A^2O^2K^4}{\delta}\right)}, \qquad (50)\end{aligned}$$

where in the last inequality we used that $D \geqslant 1$ and defined a new constant $C \coloneqq \frac{4(1-\epsilon)^3}{\epsilon^4}$ by using that

$$\frac{L}{1-\widetilde{\eta}} \leqslant \frac{L(1-\epsilon)}{\epsilon} = \frac{4(1-\epsilon)^3}{\epsilon^4} =: C,$$

where the first inequality holds for the properties of the contraction coefficient since $\widetilde{\eta} \leqslant 1 - \frac{\epsilon}{1-\epsilon}$ and the following equality follows from the definition of L in Corollary D.2.
This bound on the regret shows an intricate dependence on the problem parameters and a $\widetilde{\mathcal{O}}(\sqrt{T})$ dependence on time T, while the dependency on the number of episodes K is linear. It can be shown that under the episode termination condition employed in the *Action-wise* OAS Algorithm, the number of episodes can be bounded in the worst case by the following quantity:

$$K \leqslant A\log(T/A),$$

having logarithmic dependence on time T. By using this last result, we are finally able to provide the final expression of the regret, holding with probability at least $1-2\delta$:

$$\mathcal{R}_T \leqslant 2A\log(T/A)(D+L) + 2D\sqrt{2T\ln\left(\frac{4}{\delta}\right)} + \frac{6C(2+D)\,\widetilde{G}}{\alpha^2\widetilde{d}_{\min}}\sqrt{2SA^3T\,\log(T/A)\,\log\left(\frac{2A^6O^2\log^4(T/A)}{\delta}\right)}. \qquad (51)$$

From the formulation above, we can simplify the expression obtaining:

$$\mathcal{R}_T \leqslant \mathcal{O}\left(\frac{CD\widetilde{G}}{\alpha^2\widetilde{d}_{\min}}\sqrt{SA^3T\,\log T\,\log O}\right).$$

This last step completes the proof. □

D CONCENTRATION BOUND OF BELIEF VECTOR UNDER DIFFERENT ACTION MATRICES

We present here Lemma D.1 which will be helpful for the theoretical analysis developed in Theorem 6.1.

Lemma D.1 (Bound on Sum of Belief Errors)**.** *Let $\mathcal{Q}$ be a POMDP instance satisfying Assumption 4.1. Let $\mathbb{T} = \{\mathbb{T}_a\}_{a\in\mathcal{A}}$ be the transition model and let $\widehat{\mathbb{T}} = \{\widehat{\mathbb{T}}_a\}_{a\in\mathcal{A}}$ be its estimate. Let a sequence of actions $(a_t)_{t=0}^T$ be taken while interacting with the environment and let b and $\widehat{b}$ denote the real and estimated belief vector updated using respectively the real and the estimated transition model according to Equation* (1)*. It follows that:*

$$\sum_{t=0}^{T} \|\widehat{b}_t - b_t\|_1 \leqslant L + L \sum_{a\in\mathcal{A}} n(a)\, \|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F$$

where we use constant $L \coloneqq \frac{4(1-\epsilon)^2}{\epsilon^3}$, with ϵ being the minimum action probability appearing in Assumption 4.1, while $n(a)$ represents the number of times each action $a \in \mathcal{A}$ has been chosen during the interaction with the horizon.

Proof. Let $\widehat{b}_t$ and b_t be the estimated and real belief vector at time t updated using Equation 1, each one using respectively the estimated and real transition model. From a belief decomposition reported in De Castro et al. (2017), it is possible to express the belief error as a sum of the errors of the chosen action transition matrices, as follows:

$$\|\widehat{b}_t - b_t\|_1 \leqslant \frac{4\eta^t \|\widehat{b}_0 - b_0\|_2}{\epsilon} + \frac{4(1-\epsilon)}{\epsilon^2} \sum_{l=0}^{t-1} \eta^{t-l-1} \|\mathbb{T}_{a_l} - \widehat{\mathbb{T}}_{a_l}\|_F, \tag{52}$$

$$\leqslant \frac{8\eta^t}{\epsilon} + \frac{4(1-\epsilon)}{\epsilon^2} \sum_{l=0}^{t-1} \eta^{t-l-1} \|\mathbb{T}_{a_l} - \widehat{\mathbb{T}}_{a_l}\|_F, \tag{53}$$

with η being defined as $\eta = 1 - \frac{\epsilon}{1-\epsilon}$, while in the second inequality we simply use that $\|\widehat{b}_0 - b_0\|_2 \leqslant \|\widehat{b}_0 - b_0\|_1 \leqslant 2$.

Basically, this bound states that the error in the belief depends on the sequence of actions taken and a higher contribution is given to the error associated with the most recent actions since the contribution of the error of each action decreases geometrically with time.

Let us consider now the sequence of actions and observations seen during the interaction and let us denote it with $(a_0, o_0, a_1, o_1, \ldots, a_t, o_t)$. First of all, we highlight that the last action-observation pair (a_t, o_t) does not influence the update of b_t but will influence b_{t+1} that does not appear in the summation, hence the last tuple (a_t, o_t) will not influence the final result.
Let us now make explicit the expression in 53 for different values of the belief. For readability, we will use the constant term $C = \frac{2(1-\epsilon)}{\epsilon^2}$.

$$\begin{aligned}
&\|\widehat{b}_0 - b_0\|_1 \leqslant \frac{8}{\epsilon},\\
&\|\widehat{b}_1 - b_1\|_1 \leqslant \frac{8\eta}{\epsilon} + C\|\mathbb{T}_{a_0} - \widehat{\mathbb{T}}_{a_0}\|_F,\\
&\|\widehat{b}_2 - b_2\|_1 \leqslant \frac{8\eta^2}{\epsilon} + \eta C\|\mathbb{T}_{a_0} - \widehat{\mathbb{T}}_{a_0}\|_F + C\|\mathbb{T}_{a_1} - \widehat{\mathbb{T}}_{a_1}\|_F,\\
&\|\widehat{b}_3 - b_3\|_1 \leqslant \frac{8\eta^3}{\epsilon} + \eta^2 C\|\mathbb{T}_{a_0} - \widehat{\mathbb{T}}_{a_0}\|_F + \eta C\|\mathbb{T}_{a_1} - \widehat{\mathbb{T}}_{a_1}\|_F + C\|\mathbb{T}_{a_2} - \widehat{\mathbb{T}}_{a_2}\|_F,\\
&\vdots\\
&\|\widehat{b}_t - b_t\|_1 \leqslant \frac{8\eta^t}{\epsilon} + \eta^{t-1} C\|\mathbb{T}_{a_0} - \widehat{\mathbb{T}}_{a_0}\|_F + \eta^{t-2} C\|\mathbb{T}_{a_1} - \widehat{\mathbb{T}}_{a_1}\|_F + \cdots + C\|\mathbb{T}_{a_{t-1}} - \widehat{\mathbb{T}}_{a_{t-1}}\|_F.
\end{aligned}$$

By reading the expression above along a vertical direction, we can bound the sum of the belief errors across various interaction steps as follows:

$$\begin{aligned}\sum_{t=0}^{T}\|\widehat{b}_t - b_t\|_1 &\leqslant \sum_{t=0}^{T}\frac{4\eta^t\,\|\widehat{b}_0 - b_0\|_2}{\epsilon} + \frac{4(1-\epsilon)}{\epsilon^2}\sum_{t=0}^{T}\sum_{l=0}^{t-1}\eta^{t-l-1}\|\mathbb{T}_{a_l} - \widehat{\mathbb{T}}_{a_l}\|_F \\ &\leqslant \frac{8(1-\epsilon)}{(1-\eta)\epsilon} + \frac{4(1-\epsilon)}{(1-\eta)\epsilon^2}\sum_{a\in\mathcal{A}} n(a)\,\|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F \\ &\leqslant \frac{4(1-\epsilon)}{(1-\eta)\epsilon^2} + \frac{4(1-\epsilon)}{(1-\eta)\epsilon^2}\sum_{a\in\mathcal{A}} n(a)\,\|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F \\ &= \frac{4(1-\epsilon)^2}{\epsilon^3} + \frac{4(1-\epsilon)^2}{\epsilon^3}\sum_{a\in\mathcal{A}} n(a)\,\|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F, \end{aligned} \tag{54}$$

where the first term in the second inequality simply for the bound on geometric series and using that $\|\widehat{b}_0 - b_0\|_2 \leqslant 2$, while the second term holds since it can be noted that the contribution on the error of each action a depends on the number of times it is pulled $n(a)$[10] and the associated error $\|\mathbb{T}_a - \widehat{\mathbb{T}}_a\|_F$ scaled by at most by $1/1-\eta$.
The third inequality holds for any non-trivial problem instance having a number of states $S \geqslant 2$, while the last expression holds by substitution of η.
The statement of the lemma simply follows by defining constant $L \coloneqq \frac{4(1-\epsilon)^2}{\epsilon^3}$. □

In the following, we present a corollary that derives from the considerations reported in the proof of Lemma D.1.

Corollary D.2. *(One-step Belief Bound) Let $\mathcal{Q}$ be a POMDP instance satisfying Assumption 4.1. Let us denote with $\mathbb{T}_a$ and $\widehat{\mathbb{T}}_a$ respectively the real and estimated transition matrix related to action a. Starting from a common belief vector b_0, and choosing action $a \in \mathcal{A}$, the one-step error in the estimated belief vector can be bounded as:*

$$\|\widehat{b}_1 - b_1\|_1 \leqslant L_1\,\|\widehat{\mathbb{T}}_a - \mathbb{T}_a\|_F.$$

where we defined constant $L_1 \coloneqq \frac{4(1-\epsilon)}{\epsilon^2}$, for which it also holds $L_1 = (1-\eta)L$.

Proof. The proof of this corollary easily follows by using the bound in (52) on $t = 1$ and having that $b_0 = \widehat{b}_0$. □

E AUXILIARY RESULTS FOR THE PROOFS OF LEMMA 5.1 AND THEOREM 6.1

This section is devoted to the presentation of different useful results that are used throughout the work.

The first one is taken from Zhou et al. (2021) and provides a bound on the maximum span $span(v)$ of the bias function appearing in the Bellman Equation (2).

Proposition E.1 (Uniform bound on the bias span from Zhou et al. (2021))**.** *Let us assume to have a POMDP instance that can be rewritten as a belief MDP. If Assumption 4.1 holds, then for ρ, v satisfying the Bellman Equation (2), we have the span of the bias function $span(v) := \max_{b\in\mathcal{B}} v(b) - \min_{b\in\mathcal{B}} v(b)$ is bounded by $D(\epsilon)$, where:*

$$D(\epsilon) := \frac{8\Big(\frac{2}{(1-\alpha)^2} + (1+\alpha)\log_\alpha\big(\frac{1-\alpha}{8}\big)\Big)}{1-\alpha}, \qquad \textit{with} \qquad \alpha = \frac{1-2\epsilon}{1-\epsilon} \in (0,1).$$

For all bias functions v associated with a belief MDP generated from a POMDP $\mathcal{Q}$, this proposition ensures that $span(v)$ is bounded by $D = D(\epsilon/2)$.

[10] We highlight here that this count does not consider the last action a_T since it does not influence the bound.

Lemma E.2 (Lemma A.1 in Ramponi et al. (2020)). *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ any pair of vectors, then it holds that:*

$$\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} \right\|_2 \leqslant \frac{2\|\boldsymbol{x} - \boldsymbol{y}\|_2}{\max\{\|\boldsymbol{x}\|_2, \|\boldsymbol{y}\|_2\}}$$

The following result instead shows the relation between vectors which are obtained by the aggregation of higher-dimensional ones.

Lemma E.3 (Aggregation Lemma in Russo et al. (2024a)). *Let $\boldsymbol{M}$ be a matrix of dimension $X \times Y$ and have positive values. Let $\widehat{\boldsymbol{M}}$ be an estimation of $\boldsymbol{M}$. Let now $\boldsymbol{c}$ be a vector of dimension X obtained by summing all the elements of $\boldsymbol{M}$ along the second dimension, such that $\boldsymbol{c}(i) = \sum_{j=1}^{J} \boldsymbol{M}(i,j)$ and let $\widehat{\boldsymbol{c}}$ be a vector obtained with the same procedure from $\widehat{\boldsymbol{M}}$. Then we will have:*

$$\|\widehat{\boldsymbol{c}} - \boldsymbol{c}\|_2 \leqslant \sqrt{Y} \|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_F$$

Lemma E.4 (Lemma 19 in Jaksch et al. (2010)). *For any sequence of numbers $y_0, \ldots, y_{n-1}$ with $0 \leqslant y_k \leqslant Y_k$ and $Y_k \coloneqq \max\{1, \sum_{i=0}^{k-1} y_i\}$:*

$$\sum_{k=0}^{n-1} \frac{y_k}{\sqrt{Y_k}} \leqslant \left(\sqrt{2} + 1\right) \sqrt{Y_n}.$$

Lemma E.5. *Let a policy $\pi \in \mathcal{P}$ interact with a POMDP instance satisfying Assumption 4.1 and let $\boldsymbol{d}_{S^2}^{(a)} \in \Delta(\mathcal{S}^2)$ denote the distribution induced on a pair of consecutive states (S_t, S_{t+1}) conditioned on event $A_t = a$.*
Let us denote with $\boldsymbol{d}_{S^2}^{(a)}(s, \cdot) \in \mathbb{R}^S$ the vector containing the different elements $\boldsymbol{d}_{S^2}^{(a)}(s, s')$ $\forall s' \in \mathcal{S}$ and let us denote the vector of sum as $\boldsymbol{d}_{S^2}^{(a)}(s) \coloneqq \sum_{s' \in \mathcal{S}} \boldsymbol{d}_{S^2}^{(a)}(s, s')$. Then, for any state $s \in \mathcal{S}$, it holds that:

$$\left\| \boldsymbol{d}_{S^2}^{(a)}(s, \cdot) \right\|_2^2 \geqslant \frac{d_{\min}^{(a)}}{\sqrt{S}},$$

with $d_{\min}^{(a)}$ being the minimum value of the distribution $\boldsymbol{d}_{S^2}^{(a)}$.

Proof. The result of the lemma derives from the following considerations:

$$\begin{aligned}
\left\| \boldsymbol{d}_{S^2}^{(a)}(s, \cdot) \right\|_2^2 &= \sum_{s' \in \mathcal{S}} \left[\boldsymbol{d}_{S^2}^{(a)}(s, s') \right]^2 \\
&\geqslant \frac{1}{S} \left(\sum_{s' \in \mathcal{S}} \boldsymbol{d}_{S^2}^{(a)}(s, s') \right)^2 \\
&= \frac{1}{S} \left[\boldsymbol{d}_{S}^{(a)}(s) \right]^2 \\
&\geqslant \frac{1}{S} \left[d_{\min}^{(a)} \right]^2,
\end{aligned}$$

where the first equality simply derives from the definition of $\boldsymbol{d}_{S^2}^{(a)}(s, \cdot)$, while the first inequality derives from the relation $\sqrt{X}\,\|\boldsymbol{x}\|_2 \geqslant \|\boldsymbol{x}\|_1$ holding $\forall \boldsymbol{x} \in \mathbb{R}^X$.
The second equality instead directly derives from the definition of $\boldsymbol{d}_{S}^{(a)}(s)$. For the last inequality, we have defined $d_{\min}^{(a)} := \min_{s' \in \mathcal{S}} \boldsymbol{d}_{S}^{(a)}(s')$ as a lower bound to the values of the distribution $\boldsymbol{d}_{S}^{(a)}$ induced by policy $\pi \in \mathcal{P}$. □

Lemma E.6. *(Link between Transition Model and Induced Distribution) Let a distribution $\boldsymbol{d}_{S^2}^{(a,n,m)} \in \Delta(\mathcal{S}^2)$ be defined on consecutive states, as in Equation (7). Then, the following relation holds:*

$$\mathbb{T}_a(s'|s) = \frac{\boldsymbol{d}_{S^2}^{(a,n,m)}(s, s')}{\sum_{s'' \in \mathcal{S}} \boldsymbol{d}_{S^2}^{(a,n,m)}(s, s'')} \qquad \forall s, s' \in \mathcal{S}.$$

Proof. The result of the lemma easily derives from the following observations.
Starting from the distribution on consecutive states $\boldsymbol{d}_{S^2}^{(a,n,m)}$, we introduce a new distribution $\boldsymbol{d}_{S}^{(a,n,m)} \in \Delta(\mathcal{S})$ defined on a single state and where each of its elements is obtained as follows:

$$\boldsymbol{d}_{S}^{(a,n,m)}(s) = \sum_{s' \in \mathcal{S}} \boldsymbol{d}_{S}^{(a,n,m)}(s, s'). \tag{55}$$

The successive key step is to observe that:

$$\boldsymbol{d}_{S}^{(a,n,m)}(s, s') = \boldsymbol{d}_{S}^{(a,n,m)}(s)\mathbb{T}_a(s'|s) \tag{56}$$

which holds for the Markovianity of the problem since the probability of the next state s' given the current state s and the action a is determined by the transition model.
By using both 55 and 56, we obtain:

$$\frac{\boldsymbol{d}_{S^2}^{(a,n,m)}(s, s')}{\sum_{s'' \in \mathcal{S}} \boldsymbol{d}_{S^2}^{(a,n,m)}(s, s'')} = \frac{\boldsymbol{d}_{S}^{(a,n,m)}(s)\mathbb{T}_a(s'|s)}{\boldsymbol{d}_{S}^{(a,n,m)}(s)} = \mathbb{T}_a(s'|s) \qquad \forall s, s' \in \mathcal{S}.$$

which completes the proof. □

F SIMULATION DETAILS

This section is devoted to providing details about the numerical experiments reported in the main paper. All the reported experiments have been run using 88 Intel(R) Xeon(R) CPU E7-8880 v4 @ 2.20GHz CPUs and 94 GB of RAM.

F.1 Generation of Transition and Observation Models

The instances used in the various experiments have been generated in a random way and, in a successive step, the following modifications are applied:

- concerning each action transition matrix $\mathbb{T}_a$, the generated ones are such that their minimum transition probability is at least $\epsilon = 1/(20S)$.
- for the observation model, for each pair of states and actions, we set a specific observation that will be drawn with higher probability in order to avoid having too much stochasticity in the reward distributions and ensure a diverse observation distribution among states.

F.2 Estimation Error of Transition Matrix

As reported in the main paper, the characteristic of the considered POMDP instances are: for the left plot $S = 5$ states, $A = 4$ actions and $O = 8$ observations, while for the plot on the right we have $S = 10$ states, $A = 4$ actions and $O = 16$ observations.
Instead of using belief-based policies that are optimal in the long horizon but require a planning step, we opted for policies choosing the action that maximizes the instantaneous expected reward based on the current belief state. In order to also select sub-optimal actions (since in this experiment we are not interested in the cumulated reward), we make these policies stochastic, with each action having a minimum probability $\iota = 0.15$ of being selected. Each policy updates its belief based on an internal transition model which is independent and different from the real transition model of the POMDP. After running each policy for 10^4 steps, we change the internal transition model used for the belief update: this will also change the distribution induced by the policy. We apply this methodology to both POMDP instances. We repeat each experiment 10 times and we report in the plot the average result for each transition matrix and a 95% confidence interval.

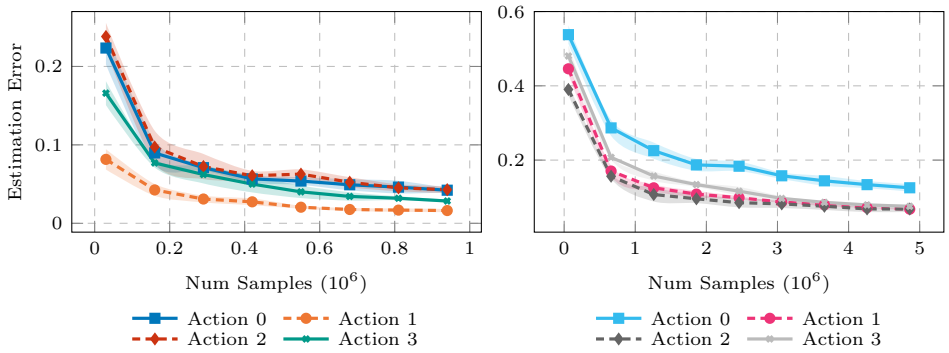

Figure 3: Error in Frobenious Norm of the Different Action Transition Matrices under two POMDP Instances (10 runs, 95 %c.i.).

Table 2: Table representing the minimum singular value of the action observation matrices and the average ($\pm$ std) number of pulls of the actions in the experiments of Figure 3.

	Action 0	**Action** 1	**Action** 2	**Action** 3
Left Figure				
$\sigma_{\min}(\mathbb{O}_a)$	0.214	0.248	0.164	0.670
Number of Pulls	149994.2 ($\pm$162.5)	550202.6 ($\pm$582.2)	150093.6 ($\pm$261.9)	149709.6 ($\pm$508.7)
Right Figure				
$\sigma_{\min}(\mathbb{O}_a)$	0.057	0.155	0.078	0.126
Number of Pulls	1009711.2 ($\pm$706.7)	749978.4 ($\pm$341.1)	2356435.0 ($\pm$1290.5)	883875.4 ($\pm$457.8)

In order to provide a more detailed analysis of the results, we report in Figure 3 the same plot appearing in Figure 1 and provide details about the characteristics of the different actions in Table 2. In particular, the table shows: (i) the minimum singular value $\sigma_{\min}(\mathbb{O}_a)$ associated with each action observation model used in the experiment; (ii) the average number of times each action is chosen along the experiment.

By analyzing the Figure, we can observe that actions with low $\sigma_{\min}(\mathbb{O}_a)$ typically have higher estimation error in the transition model $\mathbb{T}_a$ since they require a larger amount of samples. This aspect can indeed be observed in the dependencies of problem parameters appearing in Lemma 5.1.
Of course, this aspect is mitigated when the number of pulls increases. For example, Action 2 on the right figure presents a low $\sigma_{\min}(\mathbb{O}_a)$ but the high number of pulls makes the estimation error lower.

F.3 Regret Experiments

The experimental results on the regret have been done in the following way. First of all, we consider a POMDP instance with $S = 3$ states, $A = 4$ actions and $O = 4$ observations generated as described in Section F.1. We run each experiment 10 times using the total horizon of approximately $T \approx 4 * 10^5$ timestamps. The methodology employed for this set of experiments is similar to the one adopted in the work of Russo et al. (2024a).

In particular, the planning task is executed by discretizing the belief space. We can thus solve the Bellman Equation by adapting the Extended Value Iteration algorithm (Jaksch et al., 2010) to the discretized state space.

Inspired by similar works such as Azizzadenesheli et al. (2016) and Russo et al. (2024a), the theoretical bounds are replaced by smaller values. This approach is commonly employed when performing experimental comparisons

in these settings and it mostly translates into a regret with bigger multiplicative constants or a result holding with smaller probability.

SEEU algorithm In the experiments, the classical Spectral Decomposition approach is modified to make the comparison between the approaches more fair. Following the procedure highlighted in Russo et al. (2024a), we have that:

- The matrices used by the Spectral Decomposition approach are not updated based on the realizations of the observation received when choosing an action, but we directly provide the matrix with the probabilities defining the real observation distribution associated with the chosen action and the underlying state. This caveat helps the estimation of both the transition and the observation model since it removes the noise given by the realizations of the observations.
- The computation of the optimistic policy for the SEEU algorithm is done by providing the real observation model (together with the estimated transition model) to the Extended Value Iteration algorithm.

The parameters used for the SEEU algorithm are $\tau_1 = 8000$ and $\tau_2 = 20000$, which are used to determine the length of the exploration and the exploitation phase respectively.

PSRL-POMDP In order to implement this algorithm, we opted for the particle filter approach, commonly used in the Bayesian setting. The particle filter strategy does not offer guarantees in terms of consistency but allows updating the model parameters in a tractable manner. We chose this approach since for the moment no consistent estimators for the latent variable model are present in the Bayesian setting. By recalling the parameters of the algorithm described in (Jafarnia Jahromi et al., 2022), we set:

- $SCHED(t_k, T_{k-1}) = t_k + T_{k-1}$ with t_k representing the length of the k-th episode;
- Let us consider that $n_t(a)$ counts the number of times action a has been pulled up to time t. By following the approaches proposed in their work, we set $\widetilde{m}_t(s,a) = n_t(a)$ with $\widetilde{m}_t(s,a)$ being an upper bound to the expected number of times the pair (s,a) has been encountered up to time t.
- We use $N = 100$ particles for each experiment, while updates of the particles are triggered when the *effective sample size* (ESS) associated with their weights goes below 30.

OAS-UCRL Concerning the OAS-UCRL algorithm, we set a minimum action probability $\iota = 0.025$ for all the actions. We chose this value since higher values would have incurred into a higher regret over time. Finally, the initial length of the episode has been set to $T_0 = 2500$.
Differently from the procedure suggested in Russo et al. (2024a) which considers non-overlapping pairs of consecutive elements, we adapted the approach to consider overlapping ones. This modification preserves the theoretical guarantees of the approach and merely translates to adding a multiplicative factor that accounts for the dependency of overlapping samples.

***Action-wise* OAS-UCRL** For our approach, which does not need many parameters to be set, we provided an initial length of the episode of $T_0 = 2500$, analogously as the OAS-UCRL case.

F.4 Ablation Study on Minimum Action Probability of OAS-UCRL

In this set of experiments, we compare the performance of the OAS-UCRL algorithm when different values for the minimum action probability ι are used. The simulations are executed on two different POMDP instances, each one having $S = 3$ states, $A = 3$ actions, and $O = 3$ observations.

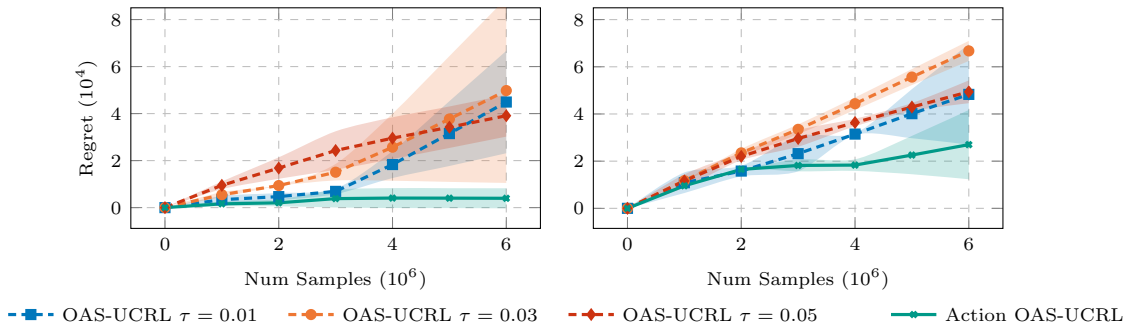

Figure 4: Regret Experiments on two Different POMDP Instances Comparing Action OAS-UCRL and OAS-UCRL with Different Values of ι (10 runs, 95 %c.i.).

The results in terms of regret for each of the two POMDP instances are shown in Figure 4.
For each instance, we run the OAS-UCRL algorithm using as minimum action probability the values $\iota = 0.01$, $\iota = 0.03$, and $\iota = 0.05$.

As highlighted in the figure, there exists a trade-off for the value of ι. In particular, when it has higher values, the amount of exploration increases, which results in having better model estimates but may lead to higher regret since suboptimal actions are chosen more frequently. This aspect can be observed on the right plot.

Contrarily, when ι has lower values, it may result in a lower regret since suboptimal actions are chosen less frequently, but this aspect could also lead to imprecise model estimates, which could in turn increase the suffered regret, as observed from the left plot in Figure 4.

In both plots, we observe the superiority of the Action OAS-UCRL algorithm, whose exploration is driven only by the optimistic approach. In addition, the ability of Action OAS-UCRL to reuse samples across episodes allows for better model estimates, which contributes to experiencing a lower regret.

F.5 Ablation Study on Sample Reuse Strategy of Action OAS-UCRL

The objective of this set of experiments is to show the effect of reusing all samples collected during the different episodes against the case where only samples from the last episodes are used for model estimation (as done in OAS-UCRL (Russo et al., 2024a)).

As shown in Lemma 5.1, the estimation error of each transition matrix scales with the number of times the action has been pulled during the different episodes. Hence, using samples collected from all the episodes leads to lower estimation error with respect to using only samples from the last episode.

Figure 5 presents the results in terms of regret of the standard AOAS-UCRL algorithm and a variant that only uses samples collected from the last episode. Concerning the variant, since AOAS-UCRL may not select all actions during each episode, we use uniform transition matrices for actions that are not chosen during the last episode.

The experiments in Figure 5 are run on three different POMDP instances, each one having $S = 3$ states, $A = 5$ actions and $O = 3$ observations. We observe that reusing all samples generally leads to better performances overall.

In particular, depending on the specific POMDP instance, this advantage can be more or less evident. Indeed, the two instances on the left show more evidently the benefit of reusing all samples, differently from the instance on the right which shows similar performances. It can indeed be the case that even if the estimated model presents higher error, the policies computed using the two strategies are similar and lead to comparable performances.

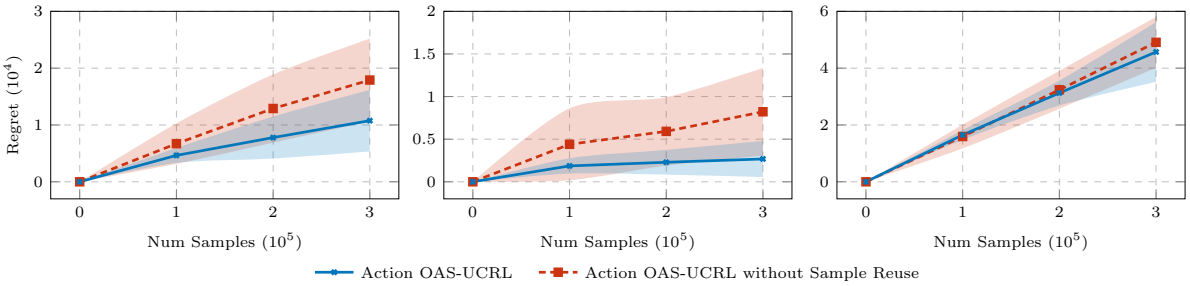

Figure 5: Regret Experiments comparing the Regret Performance of the standard AOAS-UCRL Algorithm Against the case with No Sample Reuse Across Episodes (10 runs, 95 %c.i.).

Another remark, more evident in the plot in the center, is that the confidence intervals for the strategy without sample reuse are larger since the estimates may vary more across the different runs since they rely on less samples.