Received: 10 July 2021 | Accepted: 2 June 2022

DOI: 10.1111/poms.13778

POMS

ORIGINAL ARTICLE

Sublinear regret for learning POMDPs

Yi Xiong[1] | **Ningyuan Chen**[2] | **Xuefeng Gao**[1] | **Xiang Zhou**[1]

[1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China

[2]Rotman School of Management, University of Toronto, Toronto, Canada

Correspondence
Xuefeng Gao, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China.
Email: xfgao@se.cuhk.edu.hk

Handling Editor: Max Z.J. Shen

Funding information
NSERC, Grant/Award Number: RGPIN-2020-04038; Hong Kong RGC GRF, Grant/Award Numbers: 14201520, 14201421

Abstract

We study the model-based undiscounted reinforcement learning for partially observable Markov decision processes (POMDPs). The oracle we consider is the optimal policy of the POMDP with a known environment in terms of the average reward over an infinite horizon. We propose a learning algorithm for this problem, building on spectral method-of-moments estimations for hidden Markov models, the belief error control in POMDPs and upper confidence bound methods for online learning. We establish a regret bound of $O(T^{2/3}\sqrt{\log T})$ for the proposed learning algorithm where T is the learning horizon. This is, to the best of our knowledge, the first algorithm achieving sublinear regret with respect to our oracle for learning general POMDPs.

KEYWORDS

exploration–exploitation, online learning, partially observable MDP, spectral estimator

1 | INTRODUCTION

The partially observable Markov decision process (POMDP) is a framework for dynamic decision making when some evolving state of the system cannot be observed. It extends the Markov decision process (MDP) and can be used to model a wide variety of real-world problems, ranging from healthcare to business. The solution to POMDPs is usually through a reduction to MDPs, whose state is the belief (a probability distribution) of the unobserved state of the POMDP, see, for example, Krishnamurthy (2016) for an overview.

We study the problem of decision making when the environment of the POMDP, such as the transition probability of the hidden state and the probability distribution governing the observation, is unknown to the agent. Thus, the agent has to simultaneously learn the model parameters (we use "environment" and "parameters" interchangeably) and take optimal actions. Such online learning framework has received considerable attention in the last decades (Sutton & Barto, 2018). Despite the practical relevance of POMDPs, the learning of POMDPs is considered much more challenging than finite-state MDPs and few theoretical results are known. This is not surprising: Even with a known environment, the corresponding belief MDP features a continuous state space. When the environment is unknown, we face the additional difficulty of not being able to calculate the belief accurately, whose updating formula is based on the environment. This is in contrast to the learning of standard MDPs, in which the state is always observed exactly.

To tackle this daunting task, we provide an algorithm that achieves sublinear regret, which is a popular measure for the performance of a learning algorithm relative to that of the oracle, that is, the optimal policy in the known environment. This is the first algorithm that achieves sublinear regret, to our knowledge, in the general POMDP setup we consider. We summarize the *three major contributions* of this paper below.

In terms of problem formulation, we benchmark our algorithm against an oracle and measure the performance by calculating the regret. The oracle we consider is the strongest among the recent literature (Azizzadenesheli et al., 2016; Fiez et al., 2018). In particular, the oracle is the optimal policy of the POMDP with a known environment in terms of the average reward over an infinite horizon. Such an oracle has higher average reward than the oracles that use the best fixed action (Fiez et al., 2018) or the optimal memoryless policy (the action only depends on the current observation) (Azizzadenesheli et al., 2016). Still our algorithm is able to attain sublinear regret in the length of the learning horizon. This implies that as the learning horizon increases, the algorithm tends to approximate the strong oracle more accurately.

In terms of the algorithmic design, the learning algorithm we propose (see Algorithm 1) has two key ingredients. First, it builds on the recent advance on the estimation of the parameters of hidden Markov models (HMMs) using

Accepted by Max Z.J. Shen, after two revisions.

spectral method-of-moments methods, which involve the spectral decomposition of certain low-order multivariate moments computed from the data (Anandkumar et al., 2012, 2014; Azizzadenesheli et al., 2016). It benefits from the theoretical finite-sample bound of spectral estimators, while the finite-sample guarantees of other alternatives such as maximum likelihood estimators remain an open problem (Lehéricy, 2019).[1] Second, it builds on the well-known "upper confidence bound" (UCB) method in reinforcement learning (Auer & Ortner, 2006; Jaksch et al., 2010). We divide the horizon into nested exploration and exploitation phases. We use spectral estimators in the exploration phase to estimate the unknown parameters such as the transition matrix of the hidden state, which itself is a function of the action in the period. We apply the UCB method to control the regret in the exploitation phase based on the estimated parameters in the exploration phase and the associated confidence regions. Although the two components have been studied separately before, it is a unique challenge to combine them in our setting. In particular, the belief of the hidden state is subject to the estimation error. We recalibrate the belief at the beginning of each exploitation phase based on the most recent estimate of the parameters. This helps us achieve the sublinear regret.

In terms of regret analysis, we establish a regret bound of $O(T^{2/3}\sqrt{\log(T)})$ for our proposed learning algorithm where T is the learning horizon. Our regret analysis draws inspirations from Jaksch et al. (2010) and Ortner and Ryabko (2012) for learning MDPs and undiscounted reinforcement learning problems, but the analysis differs significantly from theirs since there are two main technical challenges in our problem.

First, the belief in POMDPs, unlike the state in MDPs, is not directly observed and needs to be estimated. This is in stark contrast to learning MDPs (Jaksch et al., 2010; Ortner & Ryabko, 2012) with observed states. As a result, we need to bound the estimation error of the belief, which itself depends on the estimation error of the model parameters. In addition, we also need to bound the error in the belief transition kernel, which depends on the model parameters in a complex way via Bayesian updating. To control these errors, we extend the approach in De Castro et al. (2017) for HMM to POMDP, and relate the error in belief transitions to the estimation error of POMDP parameters and the belief state error.

Second, to establish the regret bound, we need a uniform bound for the span of the bias function (also referred as the relative value function) for the optimistic belief MDPs, which have continuous state spaces. Such a bound is often critical in the regret analysis of undiscounted reinforcement learning of continuous MDP, but it is often shown under restrictive assumptions such as the Hölder continuity that do not hold for the belief state in our setting (Lakshmanan et al., 2015; Ortner & Ryabko, 2012). We develop a novel approach to bound the bias span for the undiscounted POMDP by bounding the Lipschitz modulus of the optimal value function for infinite-horizon discounted problems when the discount factor tends to one. One key step is to bound the Lipschitz module of the belief transition kernels using the Kantorovich metric. Exploiting the connection with the infinite-horizon undiscounted problem via the vanishing discount factor method then yields an explicit bound on the bias span for the optimistic belief MDPs.

1.1 | Related literature

This paper extends the online learning framework popularized by multiarmed bandits to POMDPs. There is a large stream of literature on the topic of bandits, see, for example, Bubeck and Cesa-Bianchi (2012) for a survey. In POMDPs, the rewards across periods are not independent any more. A stream of literature studies nonstationary/switching MAB, including Auer et al. (2019), Auer et al. (2002b), Besbes et al. (2014), Cheung et al. (2022), Garivier and Moulines (2011), and Keskin and Zeevi (2017). The reward can change over periods subject to a number of switches or certain changing budget (the total magnitude of reward changes over the horizon), and the oracle is the best action in each period. It should be noted that the oracle considered is stronger than ours. However, all the designed algorithms in this literature require finite switches or sublinear changing budget (in the order of $o(T)$). This is understandable, as there is no hope to learn such a strong oracle if the actions can be completely different across periods. In our setting, the number of changes (state transitions) is linear in T and the algorithms are expected to fail to achieve sublinear regret even measured against our oracle, which is weaker than the oracle in this stream of literature. There are a few exceptions, including Chen et al. (2021), Zhu and Zheng (2020), and Zhou et al. (2021), which study models with linear changing budget but specific structures. In Chen et al. (2021), the rewards are cyclic, which can be leveraged to learn across cycles despite of the linear change. In Zhu and Zheng (2020), the reward grows over time according to a function. In Zhou et al. (2021), the reward is modulated by an unobserved Markov chain. Another stream of literature investigates the so-called restless Markov bandit problem, in which the state of each action evolves according to independent Markov chains, whose states may not be observable. See, for example, Guha et al. (2010), Ortner et al. (2014), and Slivkins and Upfal (2008). The POMDP model we consider has a more complex structure. Thus, the algorithms proposed in the above studies cannot achieve sublinear regret.

Our work is related to the rich literature on learning MDPs. Jaksch et al. (2010) propose the UCRL2 (Upper Confidence Reinforcement Learning) algorithm to learn finite-state MDPs and prove that the algorithm can achieve the optimal rate of regret measured against the optimal policy in terms of the undiscounted average reward. Follow-up papers have investigated various extensions to Jaksch et al. (2010), including posterior sampling (Agrawal & Jia, 2017), minimax optimal regret (Azar et al., 2017; Zhang & Ji, 2019), and the model-free setting (Jin et al., 2018). Cheung et al. (2019) consider the case where the parameters of the MDP, such as the transition matrix, may change over time. The algorithms are not applicable to our setting, because of the unobserved

state in POMDPs. However, since a POMDP can be transformed to a continuous-state MDP, our setting is related to the literature, especially those papers studying MDPs with a continuous state space. Lakshmanan et al. (2015) and Ortner and Ryabko (2012) extend the algorithm in Jaksch et al. (2010) to a continuous state space. Still, our problem is not equivalent to the learning of continuous-state MDPs. First, in this literature Hölder continuity is typically assumed for the rewards and transition probabilities with respect to the state, in order to aggregate the state and reduce it to the discrete case. However, this assumption does not hold in general for the belief state of POMDPs, whose transition probabilities are not given but arise from the Bayesian updating. Second, even if the continuity holds, the state of the belief MDP in our problem, which is the belief of the hidden state, cannot be observed. It can only be inferred using the estimated parameters. This distinguishes our problem from those studied in this literature. The algorithm and analysis also deviate substantially as a result. There are studies that focus on the applications such as inventory management (Chen et al., 2019, 2020; Nambiar et al., 2021; Zhang et al., 2018, 2020) and handle specific issues such as demand censoring and lost sales.

Our work is related to studies on reinforcement learning for POMDPs, see, for example, Ross et al. (2011) and Spaan (2012) and references therein. Guo et al. (2016) propose a learning algorithm for a class of episodic POMDPs, where the performance metric is the sample complexity, that is, the time required to find an approximately optimal policy. Recently, Jin et al. (2020) give a sample efficient learning algorithm for episodic finite undercomplete POMDPs, where the number of observations is larger than the number of hidden states. Their focus is on the sample complexity, while the method may potentially be used in the regret analysis of our infinite-horizon average reward setting. There is also a growing body of literature that apply deep reinforcement learning methods to POMDPs, see, for example, Hausknecht and Stone (2015) and Igl et al. (2018). Our work differs from these papers in that we study the learning of ergodic POMDPs in an unknown environment and we focus on developing a learning algorithm with sublinear regret guarantees. A concurrent study (Kwon et al., 2021) considers regret minimization for reinforcement learning in a special class of POMDPs called latent MDP. The hidden state is static in their work while it is dynamic in our setting.

Furthermore, our work is related to the literature on the spectral method to estimate HMMs and its application to POMDPs. For instance, Anandkumar et al. (2012, 2014) use the spectral method to estimate the unknown parameters in HMMs, by constructing the so-called multiviews from the observations. The spectral method is not readily applicable to POMDPs, because of the dependence introduced from the actions. Azizzadenesheli et al. (2016) address the issue by restricting to memoryless policies, that is, the action only depends on the observation in the current period instead of the belief state. They extend the spectral estimator to the data generated from an arbitrary distribution other than the stationary distribution of the Markov chain, which is necessary in learning problems when the policy needs to be experimented.

There are two papers whose methodology is closely connected to this paper that warrant more discussion (Azizzadenesheli et al., 2016; De Castro et al., 2017). Azizzadenesheli et al. (2016) use spectral estimators and upper confidence methods to learn POMDPs and establish a regret bound of $O(\sqrt{T}\log(T))$. One main difference between our work and theirs is the choice of the oracle/benchmark. Specifically, their oracle is the optimal memoryless policy, that is, a policy that only depends on the current reward observation instead of using all historical observations to form the belief of the underlying state. For general POMDPs, memoryless policies are suboptimal and the performance gap is linear in T between their oracle and ours. Technically, it allows them to circumvent the introduction of the belief state entirely. In our setting, we need to design a new learning algorithm to achieve sublinear regret with our stronger oracle and analyze the regret. By considering the belief-based policies, several new difficulties arise in our setting. First, the spectral method cannot be applied to samples generated from belief-based policies due to history dependency; second, the belief states cannot be observed and need to be calculated using the estimated parameters, which is not an issue in Azizzadenesheli et al. (2016) because the observation in the current period can be regarded as the state. We tackle these difficulties by using an exploration–exploitation interleaving approach in the algorithm design. In particular, we develop three recipes to analyze the regret. First, we develop a novel approach to upper bound the span of the bias function for the average-reward POMDP model. In general the bias span is a complicated function of the POMDP model parameters. Our paper appears to be the first to make the bound explicit in terms of the smallest element of the transition matrices, and this is one of our main methodological contributions. This result is significant as it simultaneously provides bounds on the bias span of the estimated POMDP and the optimistic POMDP when they are close to the true POMDP model in terms of model parameters. Our bound on the bias span is different from the diameter of the POMDP discussed in Azizzadenesheli et al. (2016). The diameter in Azizzadenesheli et al. (2016) is only for observation-based policies, not for belief state–based policies we consider. Second, to control the regret, we bound the error in the belief state incurred by the errors in the estimation of POMDP parameters. We extend the approach in De Castro et al. (2017), who study filtering and smoothing errors in the context of nonparametric HMMs. Such HMM models do not involve actions or decision making as in the POMDP model we consider. Third, to control the regret, we also bound the error in the (estimated) belief transition law. This is similar to online learning of finite-state MDPs where one often bounds the error in the estimates of transition probabilities in model-based methods. However, our problem is more sophisticated because the belief state is not observed, and hence cannot be

directly estimated. Therefore, we control this error in the transition law of beliefs by relating it to the estimation error of POMDP parameters and the belief state error.

The rest of the paper is organized as follows. In Section 2 we discuss the problem formulation. Section 3 presents our learning algorithm. In Section 4, we state our main results on the regret bounds for the learning algorithm. In Section 5, we present numerical experiments. Finally, we conclude in Section 6. All the proofs of the results in the paper are deferred to the Supporting Information.

2 | PROBLEM FORMULATION

We first introduce the notation for the POMDP. A POMDP model with horizon T consists of the tuple

$$\{\mathcal{M}, \mathcal{I}, \mathcal{O}, \mathcal{P}, \Omega, R\}, \tag{1}$$

where

- $\mathcal{M} := \{1, 2, \dots, M\}$ denotes the state space of the hidden state. We use $M_t \in \mathcal{M}$ to denote the state at time $t = 1, 2, \dots, T$.
- $\mathcal{I} := \{1, 2, \dots, I\}$ denotes the action space with $I_t \in \mathcal{I}$ representing the action chosen by the agent at time t.
- $\mathcal{O} := \{o_1, o_2, \dots, o_O\}$ is a finite set of possible observations and O_t denotes the observation at time t.
- $\mathcal{P} := \{P^{(1)}, \dots, P^{(I)}\}$ describes a family of transition probability matrices, where $P^{(i)} \in \mathbb{R}^{M\times M}$ is the transition probability matrix for states in $\mathcal{M}$ after the agent takes action $i \in \mathcal{I}$. That is, $P^{(i)}(m, m') = \mathbb{P}(M_{t+1} = m' | M_t = m, I_t = i)$ for $m, m' \in \mathcal{M}$.
- The observation density function $\Omega(o|m, i)$ is a distribution over observations $o \in \mathcal{O}$ that occur in state $m \in \mathcal{M}$ after the agent takes action i in the last period, that is, $\Omega(o|m, i) = \mathbb{P}(O_t = o | M_t = m, I_{t-1} = i)$.
- The reward function $R(m, i)$ specifies the immediate reward for each state–action pair (m, i), and we assume the reward function $R : \mathcal{M} \times \mathcal{I} \to [0, r_{\max}]$ for some constant $r_{\max} > 0$.

The following sequence of events occur in order in each period. In period t, the underlying state transits to M_t. Then the agent observes $O_t \in \mathcal{O}$, whose distribution depends on M_t and I_{t-1}. The agent then chooses an action $I_t \in \mathcal{I}$ and receives reward R_t determined by reward function, which depends on the state M_t and the action I_t. Then the time proceeds to $t + 1$ and the state transits to M_{t+1}, whose transition probability depends on the action I_t.

In the POMDP model, the agent does not observe the state M_t, but only the noisy observation O_t, after which an action is chosen. Moreover, since the agent does not know the state M_t at time t, it does not know the reward $R_t = R(M_t, I_t)$. Hence, it is typical to assume that the action does not depend on the reward in the literature (Cao & Guo, 2007; Krishnamurthy, 2016). Therefore, the action taken in period t, I_t, depends on the history up to time t, denoted by

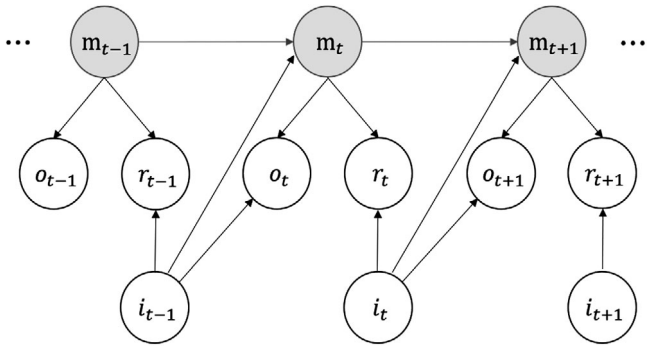

FIGURE 1 A graph showing the dependence structure of the POMDP

$$\begin{aligned} \mathcal{H}_0 &:= \{I_0\}, \\ \mathcal{H}_t &:= \{I_0, O_1, \dots, I_{t-1}, O_t\}, \quad t \geq 1. \end{aligned} \tag{2}$$

The agent attempts to optimize the expected cumulative reward over a finite horizon T. The information structure is illustrated by the graph in Figure 1.

2.1 | Reformulate POMDP as belief MDP

If the model environment in (1) is known to the agent, then it is well known (see, for example, Krishnamurthy, 2016) that to maximize the expected reward, the agent can reformulate the POMDP as an MDP with a continuous state space. The state of the MDP reflects the belief, or the distribution, over the hidden states, and thus it is referred to as the belief MDP. More precisely, define an M-dimensional vector $b_t = (b_t(1), \dots, b_t(M)) \in \mathcal{B} := \{b \in \mathbb{R}_+^M : \sum_{m=1}^M b(m) = 1\}$ as the belief of the underlying state in period t:

$$\begin{aligned} b_0(m) &:= \mathbb{P}(M_0 = m), \\ b_t(m) &:= \mathbb{P}(M_t = m | \mathcal{H}_t), \quad t \geq 1. \end{aligned} \tag{3}$$

Because of the Markovian structure of the belief, we can show (see, e.g., Krishnamurthy, 2016; Puterman, 2014) that the belief in period $t + 1$ can be updated based on the current belief $b_t = b$, the chosen action $I_t = i$, and the observation $O_{t+1} = o$. In particular, the updating function $H_{\mathcal{P},\Omega} : \mathcal{B} \times \mathcal{I} \times \mathcal{O} \to \mathcal{B}$ determines

$$b_{t+1} = H_{\mathcal{P},\Omega}(b_t, I_t, O_{t+1}). \tag{4}$$

We may omit the dependence on $\mathcal{P}$ and Ω if it does not cause confusion. By Bayes' theorem, we have

$$b_{t+1}(m) = \frac{\Omega(o|m, i) \sum_{m' \in \mathcal{M}} P^{(i)}(m', m) b_t(m')}{\mathbb{P}(o|b, i)}, \tag{5}$$

where $\mathbb{P}(o|b,i) = \sum_{m''\in\mathcal{M}} \Omega(o|m'',i) \sum_{m'\in\mathcal{M}} P^{(i)}(m', m'')b_t(m')$ is the distribution of the observation under belief b and action i.

We next introduce some notations to facilitate the discussion and analysis. Define the expected reward conditional on the belief and action

$$\bar{R}(b,i) := \sum_{m\in\mathcal{M}} R(m,i)b(m). \tag{6}$$

We can also define the transition kernel of the belief conditional on the action:

$$\begin{aligned}\bar{T}(b_{t+1}|b_t,i_t) &:= \mathbb{P}(b_{t+1}|b_t,i_t)\\ &= \sum_{o_{t+1}\in\mathcal{O}} \mathbb{1}_{\{H(b_t,i_t,o_{t+1})=b_{t+1}\}}\mathbb{P}(o_{t+1}|b_t,i_t).\end{aligned} \tag{7}$$

A policy $\mu = (\mu_t)_{t\geq 0}$ for the belief MDP is a mapping from the belief states to actions, that is, the action chosen by the agent at time t is $I_t = \mu_t(b_t)$. Following the literature (see, e.g., Agrawal & Jia, 2017), we define the gain of a policy and the optimal gain.

Definition 1. The gain of a policy μ, given the initial belief state b, is defined as the long-run average reward for the belief MDP over an infinite horizon, given by:

$$\rho_b^{\mu} := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} R(M_t,\mu_t)|b_0 = b\right], \tag{8}$$

where the expectation is taken with respect to the interaction sequence when policy μ interacts with the belief MDP. The optimal gain ρ^* is defined by

$$\rho^* := \sup_b \sup_\mu \rho_b^{\mu}. \tag{9}$$

2.2 | Assumptions

Next we provide the technical assumptions for the analysis.

Assumption 1. The entries of all transition matrices are bounded away from zero $\epsilon := \min_{i\in\mathcal{I}} \min_{m,n\in\mathcal{M}} P^{(i)}(m,n) > 0$.

Assumption 2. $\xi := \min_{i\in\mathcal{I}} \min_{o\in\mathcal{O}} \sum_{m\in\mathcal{M}} \Omega(o|m,i) > 0$.

Assumptions 1 and 2 can be strong in general, but they are required by the state-of-art method to bound the belief error caused by the parameter miscalibration (see De Castro et al., 2017, for the HMM setting), which is essential in learning POMDPs. Moreover, the two assumptions provide sufficient conditions to guarantee the existence of the solution to the Bellman optimality equation of the belief MDP and the boundedness of the bias span; See Propositions 1 and 3. Note that Assumption 1 itself implies that for any fixed i, the Markov chain with transition matrix $P^{(i)}$ is geometrically ergodic with a unique stationary distribution denoted by $\omega^{(i)}$, and the geometric rate is upper bounded by $1-\epsilon$. See, for example, Theorems 2.7.2 and 2.7.4 in Krishnamurthy (2016). This geometric ergodicity, which can hold under weaker assumptions, is needed for spectral estimations of the POMDP model as in Azizzadenesheli et al. (2016).

Assumption 3. For each $i \in \mathcal{I}$, the transition matrix $P^{(i)}$ is invertible.

Assumption 4. For all $i \in \mathcal{I}$, $\Omega(\cdot|1,i), \ldots, \Omega(\cdot|m,i)$ are linearly independent.

Assumption 3 and 4 are required for the finite-sample guarantee of spectral estimators (Anandkumar et al., 2012, 2014). See Section 3.1 for more details. Since our learning algorithm uses the spectral estimator to estimate HMMs, our approach inherits the assumptions.

Before we proceed, we first state a result on the characterization of the optimal gain ρ^* given in Definition 1 and the existence of stationary optimal policies for the belief MDP (8) under the average reward criterion. Note that in general (without the assumptions), there is no guarantee that a stationary optimal policy would exist for problem (8) (see, e.g., Yu & Bertsekas, 2004).

Proposition 1. *Suppose Assumptions 1 and 2 hold. There exists a bounded function $v : \mathcal{B} \to \mathbb{R}$ and a constant $\rho^* \in \mathbb{R}$ such that the Bellman optimality equation holds for problem (9):*

$$\rho^* + v(b) = \max_{i\in\mathcal{I}} \left[\bar{R}(b,i) + \int_{\mathcal{B}} v(b')\bar{T}(db'|b,i)\right], \quad \forall b \in \mathcal{B}. \tag{10}$$

Moreover, there exists a stationary deterministic optimal policy μ^* for problem (9), which prescribes an action that maximizes the right side of (10). The constant ρ^* is the optimal gain defined in (9).

Although Proposition 1 is necessary for the subsequent analysis, it mainly guarantees the regularity condition and the proof is independent of the algorithmic design. To understand its intuition at a high level, note that the function v is referred to as the bias function, or the relative value function of the belief state for the undiscounted problem (8) (Chapter 8 of Puterman, 2014). Conditions for the existence are known in the literature; see, for example, Hsu et al. (2006) and Ross (1968). To establish the existence, the bias functions $\{v^{\beta}(b) : \beta \in (0,1)\}$ for the infinite-horizon discounted POMDPs with discount factor β are studied. As $\beta \to 1$, the problems converge to the undiscounted one. However, the key condition to ensure the convergence is that

the bias functions $\{v^\beta\}$ are *uniformly* bounded in β. For this purpose, we use the following ideas. First, to bound v^β, it suffices to bound the Lipschitz modulus of the optimal value function for the infinite-horizon discounted problem since the belief state is a probability vector and is bounded. Then it reduces to bound the Lipschitz modulus of the optimal value function of the finite-horizon discounted problem by the results in Hinderer (2005). For finite-horizon problems, we use backward inductions to obtain bounds (uniform in β) on the Lipschitz modulus of optimal value functions recursively. A challenge in the last step is to show that the transition kernel of the belief state is a contraction when the number of steps is large, that is, it has Lipschitz modulus strictly smaller than one. This partly follows from the geometric ergodicity or uniform forgetting property of the belief (see Lemmas EC.1 and EC.3 in the Supporting Information). Our proof technique yields an explicit upper bound on the bias span in terms of the smallest element of the transition matrices in Assumption 1. Such a bound on the bias span is known to be critical in the regret analysis of learning continuous MDPs, see, for example, Lakshmanan et al. (2015).

We also remark that solving the optimality equation (10) and finding the optimal policy for POMDP with average reward criteria in a known environment are computationally challenging due to the continuous belief states. Various methods have been proposed to compute an approximately optimal policy for belief MDPs or more general continuous-state MDPs with average reward criterion. See, for example, Ormoneit and Glynn (2002), Saldi et al. (2017), Sharma et al. (2020), Yu and Bertsekas (2004), Yu and Bertsekas (2008), and the references therein. In this work, we do not focus on this planning problem and assume the access to an optimization oracle that solves the Bellman equation (10) and returns ρ^* and the optimal stationary policy μ^*.

2.3 | Learning POMDP

We consider learning algorithms to learn the POMDP model when some model parameters are unknown. In particular, the agent knows the state space $\mathcal{M}$, the action space $\mathcal{I}$, the observation space $\mathcal{O}$, and the reward function $R(m, i)$, but has no knowledge about the underlying hidden state M_t, the transition matrices $P^{(i)}$ for all actions, and the observation density function $\Omega(o|m, i)$. The goal is to design a learning policy to decide which action to take in each period to maximize the expected cumulative reward over T periods even if T is unknown in advance. Note that the setting is slightly different from multiarmed bandits, in which the reward distribution of each arm is unknown. In POMDP, it is typical to assume $R(m, i)$ to be a deterministic function and the random noise mainly comes from the observation. Moreover, the realized reward is usually not observed or used to determine the action, as mentioned previously. Therefore, it is reasonable to set up the environment to learn the parameters related to the observations. Our approach can be used to learn the reward function as well, if the historical reward can be observed.

For a learning policy π, the action taken in period t, which we denote by π_t, is adapted to the history $\mathcal{H}_t = \{\pi_0, O_1^\pi, \ldots, \pi_{t-1}, O_t^\pi\}$, where O_t^π denotes the observation received under the learning policy π in period t. Note that π_t maps the initial belief b and the history $\mathcal{H}_t$ to an action in period t. Similar to Definition 1, we may define the reward in period t for the policy π when the initial belief b as

$$R_t^\pi(b) := R(M_t, \pi_t). \tag{11}$$

Note that both M_t and π_t depend on the initial belief b, which we omit in the notation.

To measure the performance of a learning policy, we follow the literature (see, e.g., Agrawal & Jia, 2017; Jaksch et al., 2010; Ortner et al., 2014) and set the optimal gain as the benchmark. In particular, we define the total regret of π in T periods as

$$\mathcal{R}_T^\pi : \max_b \left\{ (T+1)\rho^* - \sum_{t=0}^{T} R_t^\pi(b) \right\}. \tag{12}$$

The objective is to design efficient learning algorithms whose regret grows sublinearly in T with theoretical guarantees. In the sequel, the dependency of $\mathcal{R}_T^\pi$ on π may be dropped if it is clear from the context.

3 | THE SPECTRAL EXPLORATION AND EXPLOITATION WITH UPPER CONFIDENCE BOUND (SEEU) LEARNING ALGORITHM

This section describes our learning algorithm for the POMDP, which is referred to as the SEEU algorithm. We first provide a high-level overview of the algorithm and then elaborate on the details.

To device a learning policy for the POMDP with unknown $\mathcal{P}$ (transition probabilities) and Ω (observation distributions), one needs a procedure to estimate those quantities from the history, that is, the past actions and observations. Anandkumar et al. (2012, 2014) propose the spectral estimator for the unknown parameters in HMMs, with finite-sample theoretical properties. It serves as a major component in the SEEU algorithm.

However, the spectral estimator is not directly applicable to ours, because there is no decision making involved in HMMs. In a POMDP, the action may depend on past observations and such dependency violates the assumptions of the spectral estimator. To address the issue, we divide the horizon T into nested "exploration" and "exploitation" phases. In the exploration phase, we choose each action successively for a fixed length of periods. This transforms the system into an HMM so that we can apply the spectral method to estimate $\mathcal{P}$ and

ALGORITHM 1 The SEEU Algorithm

Input: Precision δ, exploration hyperparameter τ_1, exploitation hyperparameter τ_2.

1: Initialize: time $T_1 = 0$, initial belief b_0.
2: **for** $k = 1, 2, 3, \ldots$ **do**
3: **for** $t = T_k, T_k + 1, \ldots, T_k + \tau_1 I$ **do**
4: Select each action τ_1 times successively.
5: Observe next observation o_{t+1}.
6: **end for**
7: Use the realized actions and observations in all previous exploration phases $\hat{I}_k : \{i_{T_1:T_1+\tau_1 I}, \ldots, i_{T_k:T_k+\tau_1 I}\}$ and $\hat{\mathcal{O}}_k : \{o_{T_1+1:T_1+\tau_1 I+1}, \ldots, o_{T_k+1:T_k+\tau_1 I+1}\}$ as input to Algorithm 2 to compute

$$(\hat{\mathcal{P}}_k, \hat{\Omega}_k) = \textbf{SpectralEstimation}(\hat{I}_k, \hat{\mathcal{O}}_k).$$

8: Compute the confidence region $C_k(\delta_k)$ centered at $(\hat{\mathcal{P}}_k, \hat{\Omega}_k)$ from (20) using the confidence level $1 - \delta_k = 1 - \delta/k^3$ such that $\mathbb{P}\{(\mathcal{P}, \Omega) \in C_k(\delta_k)\} \geq 1 - \delta_k$.
9: Find the optimistic POMDP in the confidence region (ρ^* given in (9) and (10)):

$$(\mathcal{P}_k, \Omega_k) = \underset{(\mathcal{P},\Omega)\in C(\delta_k)}{\arg\max}\ \rho^*(\mathcal{P}, \Omega).$$

10: **for** $t = 0, 1, \ldots, T_k + \tau_1 I$ **do**
11: Update belief b_t^k to $b_{t+1}^k = H_{\mathcal{P}_k,\Omega_k}(b_t^k, i_t, o_{t+1})$ under the new parameters $(\mathcal{P}_k, \Omega_k)$.
12: **end for**
13: **for** $t = T_k + \tau_1 I + 1, \ldots, T_k + \tau_1 I + \tau_2\sqrt{k}$ **do**
14: Execute the optimal policy $\pi^{(k)}$ by solving the Bellman equation (10) with parameters $(\mathcal{P}_k, \Omega_k)$: $i_t = \pi_t^{(k)}(b_t^k)$.
15: Observe next observation o_{t+1}.
16: Update the belief at $t + 1$ following $b_{t+1}^k = H_{\mathcal{P}_k,\Omega_k}(b_t^k, i_t, o_{t+1})$.
17: **end for**
18: $T_{k+1} \leftarrow t + 1$
19: **end for**

Ω from the observed actions and observations in that phase. In the exploitation phase, based on the confidence region of the estimators obtained from the exploration phase, we use a UCB-type policy to implement the optimistic policy (the optimal policy for the best-case estimators in the confidence region) for the POMDP.

The SEEU algorithm is presented in Algorithm 1. The algorithm proceeds with episodes with increasing length, similar to the UCRL2 algorithm in Jaksch et al. (2010) for learning MDPs. Each episode is divided into exploration and exploitation phases. The exploration phase lasts $\tau_1 I$ periods (Step 3), where τ_1 is a tunable hyperparameter and I is the total number of actions in the action space. In this phase, the algorithm chooses each action successively for τ_1 periods. In Step 7 it applies the spectral estimator (Algorithm 2 to be introduced in Section 3.1) to (re-)estimate $\mathcal{P}$ and Ω. Moreover, it constructs a confidence region based on Proposition 2 with a confidence level $1 - \delta_k$, where $\delta_k : \delta/k^3$ is a vanishing sequence with $\delta > 0$ in episode k (Step 8). The key information to extract from the exploration phase is

- the optimistic POMDP inside the confidence region (Step 9) and
- the updated belief vector according to the new estimators (Step 11).

Then the algorithm enters the exploitation phase (Step 13), whose length is $\tau_2\sqrt{k}$ in episode k and τ_2 is another tunable hyperparameter. In the exploitation phase, an action is chosen according to the optimal policy associated with the optimistic estimators for $\mathcal{P}$ and Ω inside the confidence region. This is the principle of "optimisim in the face of uncertainty" for UCB-type algorithms.

Before getting into the details, we comment on the major difficulties of designing and analyzing such an algorithm. To apply the spectral estimator, in the exploration phase the actions are chosen deterministically to "mimic" an HMM, as mentioned above. This is necessary as the spectral estimator requires fast convergence to a stationary distribution, guaranteed by Assumption 1. Moreover, at the first sight, the recalculation of the belief in Step 11 may deviate significantly from the actual belief using the exact parameters. The belief relies on the whole history, and a small error in the estimation may accumulate over t periods and lead to an erroneous calculation. We show in Proposition 4 that the belief error can actually be well controlled. This is important for the algorithm to achieve the sublinear regret.

We remark that the horizon is divided into nested phases, instead of a single exploration phase followed by a exploitation phase, because we do not require the knowledge T in advance. If T is known to the agent, then it is indeed true that one can use a single episode (exploration followed by exploitation) to attain the same rate of regret. This is often not the case in practice.

3.1 | Exploration: Spectral method

We next zoom in to the exploration phase of a particular episode, in order to show the details of the spectral estimator in Steps 7 and 8 (Anandkumar et al., 2012, 2014; Azizzadenesheli et al., 2016). Suppose the exploration phase lasts from period 0 to N, with a fixed action $i_t \equiv i$ and realized observations $\{o_1, o_2, \ldots, o_{N+1}\}$ sampled according to the observation density Ω. When the action is fixed, the underlying state M_t converges to the steady state geometrically fast due to Assumption 1. For the ease of exposition, we assume that the system has reached the steady state at $t = 0$. In Remark 1, we discuss how to control the error as the system starts from an arbitrary state distribution.

For $t \in \{2, \ldots, N\}$, we consider three "views" (o_{t-1}, o_t, o_{t+1}). (Here a view is simply a feature of the collected data, a term commonly used in data fusion (Zhao

ALGORITHM 2 The subroutine to estimate POMDP parameters.

Input: Observed actions $\{i_0, \dots, i_N\}$ and observations $\{o_1, \dots, o_{N+1}\}$

Output: POMDP parameters $\hat{\mathcal{P}}, \hat{\Omega}$

1: **for** $i = 1, \dots, I$ **do**
2: Construct $v_{1,t}^{(i)} = o_{t-1}$, $v_{2,t}^{(i)} = o_t$ and $v_{3,t}^{(i)} = o_{t+1}$.
3: Compute $\hat{W}_{p,q}^{(i)}, p, q \in \{1,2,3\}$ according to (17),
4: Compute $\hat{v}_{1,t}^{(i)}$ and $\hat{v}_{2,t}^{(i)}$ according to (18).
5: Compute $\hat{M}_2^{(i)}$ and $\hat{M}_3^{(i)}$ according to (19).
6: Apply tensor decomposition (Anandkumar et al., 2014) to compute

$$\hat{A}_3^{(i)} = \textbf{TensorDecomposition}(\hat{M}_2, \hat{M}_3).$$

7: Compute $\hat{\theta}_{2,m}^{(i)} = \hat{W}_{2,1}^{(i)}(\hat{W}_{3,1}^{(i)})^{\dagger}\hat{\theta}_{3,m}^{(i)}$ for each $m \in \mathcal{M}$.
8: Return $\hat{\Omega}(o|m,i) = \hat{A}_2^{(i)}(o,m)$.
9: Return $\hat{P}^{(i)} = ((\hat{A}_2^{(i)})^{\dagger}\hat{A}_3^{(i)})^{\top}$.
10: **end for**

et al., 2017). We stick to the term as in the original description of the spectral estimator.) We can see from Figure 1 that given m_t and $i_t \equiv i$, all the three views are independent. Since the system has reached the steady state, the distribution of (o_{t-1}, o_t, o_{t+1}) is also stationary. The key of the spectral estimator is to express the distribution of (o_{t-1}, o_t, o_{t+1}) as a function of the parameters to learn. Then the relevant moments are matched to the samples, which is similar to the spirit of methods of moments.

We represent the views in the vector form for convenience. Formally, we encode o_{t-1} into a unit vector $v_{1,t}^{(i)} \in \{0,1\}^O$, satisfying $\mathbb{1}_{\{v_{1,t}^{(i)} = \boldsymbol{e}_o\}} = \mathbb{1}_{\{O_{t-1}=o\}}$. Similarly, o_t and o_{t+1} can also be expressed as unit vectors $v_{2,t}^{(i)} \in \{0,1\}^O$ and $v_{3,t}^{(i)} \in \{0,1\}^O$. Define three matrices $A_1^{(i)}, A_2^{(i)}, A_3^{(i)} \in \mathbb{R}^{O \times M}$ for action i such that:

$$\begin{aligned} A_1^{(i)}(o,m) &= \mathbb{P}\big(v_{1,t}^{(i)} = \boldsymbol{e}_o | m_t = m, i_t = i\big), \\ A_2^{(i)}(o,m) &= \mathbb{P}\big(v_{2,t}^{(i)} = \boldsymbol{e}_o | m_t = m, i_t = i\big), \\ A_3^{(i)}(o,m) &= \mathbb{P}\big(v_{3,t}^{(i)} = \boldsymbol{e}_o | m_t = m, i_t = i\big). \end{aligned} \tag{13}$$

By stationarity, the distribution of the matrices is independent of t. We use $\theta_{1,m}^{(i)}, \theta_{2,m}^{(i)}$, and $\theta_{3,m}^{(i)}$ to denote the m-th column of $A_1^{(i)}$, $A_2^{(i)}$, and $A_3^{(i)}$, respectively. Let $W_{p,q}^{(i)} = \mathbb{E}[v_{p,t}^{(i)} \otimes v_{q,t}^{(i)}]$ be the correlation matrix between $v_{p,t}^{(i)}$ and $v_{q,t}^{(i)}$, for $p, q \in \{1,2,3\}$.[2]

The spectral estimator uses the following modified views, which are linear transformations of $v_{1,t}^{(i)}$ and $v_{2,t}^{(i)}$:

$$\widetilde{v}_{1,t}^{(i)} := W_{3,2}^{(i)}\big(W_{1,2}^{(i)}\big)^{\dagger} v_{1,t}^{(i)}, \quad \widetilde{v}_{2,t}^{(i)} := W_{3,1}^{(i)}\big(W_{2,1}^{(i)}\big)^{\dagger} v_{2,t}^{(i)}, \tag{14}$$

where $\dagger$ represents the pseudoinverse of a matrix. It turns out that the second and third moments of the modified views,

$$M_2^{(i)} := \mathbb{E}\Big[\widetilde{v}_{1,t}^{(i)} \otimes \widetilde{v}_{2,t}^{(i)}\Big], \quad M_3^{(i)} := \mathbb{E}\Big[\widetilde{v}_{1,t}^{(i)} \otimes \widetilde{v}_{2,t}^{(i)} \otimes v_{3,t}^{(i)}\Big], \tag{15}$$

can be compactly represented by the model parameters. More precisely, by Theorem 3.6 in Anandkumar et al. (2014), we have the following spectral decomposition:

$$\begin{aligned} M_2^{(i)} &= \sum_{m \in \mathcal{M}} \omega^{(i)}(m) \theta_{3,m}^{(i)} \otimes \theta_{3,m}^{(i)}, \\ M_3^{(i)} &= \sum_{m \in \mathcal{M}} \omega^{(i)}(m) \theta_{3,m}^{(i)} \otimes \theta_{3,m}^{(i)} \otimes \theta_{3,m}^{(i)}, \end{aligned} \tag{16}$$

where we recall that $\omega^{(i)}(m)$ is the state stationary distribution under the policy $i_t \equiv i$ for all t.

With the relationship (16), we can describe the procedures of the spectral estimator. Suppose a sample path $\{o_t\}_{t=1}^{N+1}$ is observed under the policy $i_t \equiv i$. It can be translated to $N-1$ samples of $(v_{1,t}^{(i)}, v_{2,t}^{(i)}, v_{3,t}^{(i)})$, $t \in \{2, \dots, N\}$. They can be used to construct the sample average of $W_{p,q}^{(i)}$ for $p, q \in \{1,2,3\}$:

$$\hat{W}_{p,q}^{(i)} = \frac{1}{N-1} \sum_{t=2}^{N} v_{p,t}^{(i)} \otimes v_{q,t}^{(i)}. \tag{17}$$

By (14) and (15), we can construct the following estimators:

$$\hat{v}_{1,t}^{(i)} = \hat{W}_{3,2}^{(i)}\big(\hat{W}_{1,2}^{(i)}\big)^{\dagger} v_{1,t}^{(i)}, \quad \hat{v}_{2,t}^{(i)} = \hat{W}_{3,1}^{(i)}\big(\hat{W}_{2,1}^{(i)}\big)^{\dagger} v_{2,t}^{(i)}, \tag{18}$$

$$\hat{M}_2^{(i)} = \frac{1}{N-1} \sum_{t=2}^{N} \hat{v}_{1,t}^{(i)} \otimes \hat{v}_{2,t}^{(i)},$$

$$\hat{M}_3^{(i)} = \frac{1}{N-1} \sum_{t=2}^{N} \hat{v}_{1,t}^{(i)} \otimes \hat{v}_{2,t}^{(i)} \otimes v_{3,t}^{(i)}. \tag{19}$$

Plugging $\hat{M}_2^{(i)}$ and $\hat{M}_3^{(i)}$ into the left-hand sides of (16), we can apply the tensor decomposition method (Anandkumar et al., 2014) to solve $\theta_{3,m}^{(i)}$ from (16), which is denoted as $\hat{\theta}_{3,m}^{(i)}$. It can also be shown that $\theta_{1,m}^{(i)} = W_{1,2}^{(i)}(W_{3,2}^{(i)})^{\dagger}\theta_{3,m}^{(i)}$ and $\theta_{2,m}^{(i)} = W_{2,1}^{(i)}(W_{3,1}^{(i)})^{\dagger}\theta_{3,m}^{(i)}$, which naturally lead to estimators $\hat{\theta}_{1,m}^{(i)}$ and $\hat{\theta}_{2,m}^{(i)}$. As a result, the unknown parameters $P^{(i)}$ and Ω can be estimated according to the following lemma.

Lemma 1. *The unknown transition matrix $P^{(i)}$ and the observation density function Ω satisfy $\Omega(o|m,i) = A_2^{(i)}(o,m)$ and $P^{(i)} = ((A_2^{(i)})^{\dagger}A_3^{(i)})^{\top}$.*

We remark that Assumption 3 and Assumption 4 imply that all three matrices $A_1^{(i)}, A_2^{(i)}, A_3^{(i)}$ are all of full column rank (Azizzadenesheli et al., 2016), and hence the pseudoinverse

$(A_2^{(i)})^\dagger$ in Lemma 1 is well defined. The subroutine to estimate POMDP estimators is summarized in Algorithm 2.

Remark 1. The stationary distribution for a fixed action $i_t \equiv i$ is crucial for the spectral estimator $\hat{\Omega}(o|m,i)$ and $\hat{P}^{(i)}$, which allows (15) and (16) to be independent of t. In our case, the spectral estimator is applied to a sequence of samples in the exploration phase, which does not start in a steady state. This is a similar situation as Azizzadenesheli et al. (2016). Fortunately, Assumption 1 allows fast mixing so that the distribution converges to the stationary distribution at a sufficiently fast rate. We can still use Algorithm 2, which is originally designed for stationary HMMs. The theoretical result in Proposition 2 already takes into account the error attributed to mixing.

Note that Algorithm 2 is directly adapted from the spectral estimator studied in the literature, for example, Anandkumar et al. (2012) and Azizzadenesheli et al. (2016). The following result, adapted from Azizzadenesheli et al. (2016), provides the confidence regions of the estimators in Algorithm 2.

Proposition 2 (**Finite-sample guarantee of spectral estimators**). *Suppose Assumptions 1, 3, and 4 hold. For any* $\delta \in (0,1)$. *If for any action* $i \in \mathcal{I}$, *the number of samples* $N^{(i)}$ *satisfies* $N^{(i)} \geq N_0^{(i)}$ *for some* $N_0^{(i)}$, *then with probability* $1-\delta$, *the estimated* $\hat{P}^{(i)}$ *and* $\hat{\Omega}$ *by Algorithm 2 satisfy*

$$\left\|\Omega(\cdot|m,i) - \hat{\Omega}(\cdot|m,i)\right\|_1 \leq C_1 \sqrt{\frac{\log\left(\frac{6(O^2+O)}{\delta}\right)}{N^{(i)}}},$$

$$\left\|P^{(i)}(m,:) - \hat{P}^{(i)}(m,:)\right\|_2 \leq C_2 \sqrt{\frac{\log\left(\frac{6(O^2+O)}{\delta}\right)}{N^{(i)}}}. \quad (20)$$

for $i \in \mathcal{I}$ *and* $m \in \mathcal{M}$. *Here,* C_1 *and* C_2 *are constants independent of any* $N^{(i)}$.

The explicit expressions of constants $N_0^{(i)}, C_1, C_2$ are given in Section EC.4 in the Supporting Information.

Remark 2. Note that Ω and $P^{(i)}$ are identifiable up to a proper permutation of the hidden state labels, because the exact index of the states cannot be recovered. In Azizzadenesheli et al. (2016), because the reward can be observed, the states associated with the observations and rewards can be correctly identified up to a permutation. In this paper, since we follow the standard setup in the POMDP literature and assume unobserved rewards, we have to make sure the inferred states match the indices of the reward function. In practice, this issue can be resolved if one can rely on external identifiability conditions to sort and label the states (Stephens, 2000): For example, in the application of financial economics, state 1 may represent a bullish market and state 2 maps to a bearish market. Once the observation density function Ω is accurately estimated (suppose the observations are macroeconomic indicators), when the sample size N is sufficiently large, the agent can naturally construct a mapping to the states. This is similar to the assumptions made in Azizzadenesheli et al. (2016). In the numerical experiment (Section 5), the algorithm always correctly labels the states. We do not explicitly mention the permutation in the statement of Proposition 2 for simplicity, consistent with the literature such as Azizzadenesheli et al. (2016).

3.2 | Exploitation: UCB-type method

After the confidence region and the implied coverage probabilities are derived in Proposition 2 at the end of the exploration phase, we adopt the principle of "optimism in the face of uncertainty" and the associated UCB algorithm (Auer et al., 2002a) in the subsequent exploitation phase. It is based on the intuition that the UCB creates a collection of "plausible" environments and selects the one with the largest optimal gain. The UCB algorithm has been successfully applied to reinforcement learning (UCRL) (Auer & Ortner, 2006; Jaksch et al., 2010) to control the regret. We apply this idea to Step 9 of Algorithm 1. Selecting the optimal action based on the optimistic yet plausible environment helps to balance the exploration and exploitation.

3.3 | Discussions on the SEEU Algorithm

We discuss a few points related to the implementation of Algorithm 1.

3.3.1 | Computational cost of Algorithm 1

For given parameters $(\mathcal{P}, \Omega)$, we need to compute the optimal average reward $\rho^*(\mathcal{P}, \Omega)$ that depends on the parameters (Step 9 in Algorithm 1). Various computational and approximation methods have been proposed in the literature to tackle this planning problem for belief MDPs, which we have already discussed in the introduction. These methods can be applied to our algorithm. In addition, we need to find out the optimistic POMDP in the confidence region $\mathcal{C}_k(\delta_k)$ with the best average reward (Step 9 in Algorithm 1). For low-dimensional models, one can discretize $\mathcal{C}_k(\delta_k)$ into grids and calculate the corresponding optimal average reward ρ^* at each grid point so as to find (approximately) the optimistic model $(\mathcal{P}_k, \Omega_k)$. However, in general it is not clear whether there is an efficient computational method to find the optimistic plausible POMDP model in the confidence region when the unknown parameters are high dimensional. This issue is also present in other recent studies on learning continuous-state MDPs with the UCB approach, see, for example, Lakshmanan et al. (2015) for a discussion. In our regret analysis below, we do not take into account the

approximation errors arising from the computational aspects discussed here. We point out that the main contribution of this paper is not *computational*. Rather, it is to develop a theoretical framework for learning model-based POMDPs. To implement the algorithm efficiently remains an intriguing future direction.

3.3.2 | Dependence on the unknown parameters

Algorithm 1 requires some information about the unknown Markov chain to compute the confidence bounds. In particular, when computing the confidence region in Step 8 of Algorithm 1, the agent needs the information of the constants C_1 and C_2 in Proposition 2. Although C_1 and C_2 do not depend on the parameters to learn directly, such as the transition matrices, they do depend on a few "primitives" that are hard to know, for example, the mixing rate of the underlying Markov chain when the action is fixed. See Section EC.4 in the Supporting Information for more details. Still, we would argue that it is relatively easy to acquire such information and this setup is innocuous: We only need upper bounds for C_1 and C_2 for the theoretical guarantee and not exact values. Therefore, a rough and conservative estimate would be sufficient. Such dependence on some unknown parameters is common in learning problems: The parameter of the sub-Gaussian noise in multiarmed bandits is usually assumed to be known; the confidence bound in Azizzadenesheli et al. (2016) is constructed based on similar information of the underlying Markov chain; Lakshmanan et al. (2015) and Ortner et al. (2014) require the knowledge of the Hölder constant for rewards and transition probabilities. A common remedy is to dedicate the beginning of the horizon to estimate the unknown parameters, which typically does not increase the rate of the regret. Alternatively, C_1 and C_2 can be replaced by parameters that are tuned by hand. See Remark 3 of Azizzadenesheli et al. (2016) for a discussion on this issue.

3.3.3 | The explore-then-commit (ETC) algorithm

As an alternative to the SEEU algorithm, the ETC algorithm can also be applied to POMDPs. The structure of the ETC algorithm is similar to the SEEU algorithm. The main difference is that in each exploitation phase, the ETC algorithm uses the point estimator of the POMDP parameters directly while the SEEU algorithm uses the optimistic estimator in the confidence region. The detailed steps of the algorithm are summarized in Algorithm EC.1 in the Supporting Information. The ETC algorithm is conceptually simpler than the SEEU algorithm, because it does not need to calibrate the size of the confidence region and simply uses the point estimators. However, the regret analysis requires a different set of techniques, in particular, the sensitivity of POMDPs to its parameters. We show that it achieves the same rate of regret as the SEEU algorithm in Section EC.7 in the Supporting Information.

4 | REGRET ANALYSIS

In this section, we state our main results, the upper bound for the regret of Algorithm 1 in high probability and expectation. We first show a uniform bound on the span of v_k, where v_k is the bias function satisfying the Bellman equation (10) for the optimistic belief MDP in episode k of Algorithm 1. Such a bound is known to be critical in the regret analysis of learning continuous MDPs.

Proposition 3 (**Uniform bound for the span of the bias function**). *Suppose Assumption 1 and Assumption 2 hold, and* (ρ_k^*, v_k) *satisfies the Bellman equation* (10). *Fix the hyperparameter* τ_1 *in Algorithm 1 to be sufficiently large. Then, there exists a constant* D *such that* $span(v_k) := \max_{b\in\mathcal{B}} v_k(b) - \min_{b\in\mathcal{B}} v_k(b) \le D$ *for all* k, *where*

$$D := \frac{8r_{\max}\left(\frac{2}{(1-\bar{\alpha})^2} + (1+\bar{\alpha})\log_{\bar{\alpha}}\frac{1-\bar{\alpha}}{8}\right)}{1-\bar{\alpha}}, \tag{21}$$

and $\bar{\alpha} = \frac{1-\epsilon}{1-\epsilon/2} \in (0,1)$ *with* $\epsilon = \min_{i\in\mathcal{I}} \min_{m,n\in\mathcal{M}} P^{(i)}(m,n) > 0$.

Next, recall that in Algorithm 1, we recompute the belief after re-estimating the parameters in each exploration phase. If a small error in the parameter estimation may propagate over time and cause the belief to deviate from the true value, then the algorithm cannot perform well in the exploitation phase. In the next result (Proposition 4), we show guarantees that the error does not accumulate and as a result, the regret incurred in the exploitation phase is proportional to the estimation error.

Proposition 4 (**Controlling the belief error**). *Suppose Assumption 1 and Assumption 2 hold. Given the estimators* $((\hat{P}^{(i)})_i, \hat{\Omega})$ *of the true model parameters* $((P^{(i)})_i, \Omega)$, *for an arbitrary reward–action sequence* $\{i_{0:t-1}, o_{1:t}\}_{t\ge1}$, *let* b_t *and* $\hat{b}_t$ *be the corresponding beliefs in period* t *computed from the same initial belief* b_0 *under* $((P^{(i)})_i, \Omega)$ *and* $((\hat{P}^{(i)})_i, \hat{\Omega})$, *respectively. Then there exist constants* L_1, L_2 *such that*

$$\|b_t - \hat{b}_t\|_1 \le L_1 \max_{m\in\mathcal{M}, i\in\mathcal{I}} \left\|\Omega(\cdot|m,i) - \hat{\Omega}(\cdot|m,i)\right\|_1 + L_2 \max_{m\in\mathcal{M}, i\in\mathcal{I}} \left\|P^{(i)}(m,:) - \hat{P}^{(i)}(m,:)\right\|_2, \tag{22}$$

where $L_1 = \frac{4(1-\epsilon)^2}{\epsilon^2\xi}$ *and* $L_2 = \frac{4(1-\epsilon)^2}{\epsilon^3}$.

With Propositions 3 and 4, we are now ready to state our main result about the high-probability regret bound for our algorithm.

Theorem 1. *Fix the hyperparameter τ_1 in Algorithm 1 to be sufficiently large. Suppose Assumptions 1 to 4 hold. There exist constants T_0 such that for $T > T_0$, with probability at least $1 - \frac{7}{2}\delta$, the regret of Algorithm 1 satisfies*

$$\mathcal{R}_T \le CT^{2/3}\sqrt{\log\left(\frac{9(O+1)}{\delta}T\right)} + T_0\rho^*, \tag{23}$$

where $\rho^* \le r_{\max}$, and

$$\begin{aligned} C = 3\sqrt{2}\Bigg[&\left(\frac{D}{2} + \frac{D}{2}L_1 + r_{\max}L_1\right)C_1 \\ &+ \left(\frac{D\sqrt{M}}{2} + \frac{D}{2}L_2 + r_{\max}L_2\right)C_2\Bigg]\tau_1^{-1/2}\tau_2^{1/3} \\ &+ 3(\tau_1 I\rho^* + D)\tau_2^{-2/3} + D\sqrt{2\log\left(\frac{1}{\delta}\right)} \\ &+ \sqrt{2r_{\max}\log\left(\frac{1}{\delta}\right)}, \end{aligned} \tag{24}$$

with C_1, C_2 given in Proposition 2, D given in Proposition 3, and L_1, L_2 given in Proposition 4.

We briefly discuss the dependency of C on the model primitives. The dependence is square-root in M (similar to the learning of MDPs in Jaksch et al., 2010) and linear in I. In contrast, the regret of MAB typically scales in $\sqrt{I}$. This is because only one arm emerges optimal. In our setting, all actions may be optimal depending on the belief of the underlying state, and thus their parameters need to be learned equally accurately. The dependency on C_1 and C_2 is directly inherited from the confidence bounds in Proposition 2 (see also Azizzadenesheli et al., 2016). In addition, C depends on L_1 and L_2, which arise from controlling the propagated error when updating the belief of the hidden state (Proposition 4; see also De Castro et al., 2017). Finally, C depends on the bound D of the bias span (similar to the diameter of MDP in UCRL2 in Jaksch et al., 2010) in Proposition 3. The constant C may not be tight, and its dependence on some parameters may be just an artifact of our proof, but it is the best bound we can obtain.

Since $\mathcal{R}_T \le (T+1)\rho^*$ by the definition (12), we can choose $\delta = \frac{9(O+1)}{T+1}$ in Theorem 1, and obtain

$$\begin{aligned} \mathbb{E}[\mathcal{R}_T] &\le \left(CT^{2/3}\sqrt{\log\left(\frac{9(O+1)}{\delta}T\right)} + T_0\rho^*\right)\left(1 - \frac{7}{2}\delta\right) \\ &\quad + (T+1)\rho^* \cdot \frac{7}{2}\delta \\ &\le CT^{2/3}\sqrt{2\log T} + T_0\rho^* + \frac{63(O+1)}{2}\rho^*. \end{aligned} \tag{25}$$

We summarize this bound for the expected regret in the following result.

Theorem 2. *Under the same setting as in Theorem 1, the regret of Algorithm 1 satisfies*

$$\mathbb{E}[\mathcal{R}_T] \le CT^{2/3}\sqrt{2\log T} + (T_0 + 32(O+1))\rho^*, \tag{26}$$

where constants C and T_0 are given in Theorem 1.

Remark 3 (**Lower bound of the regret**). The typical optimal regret bound for online learning is $O(T^{1/2})$. The gap is probably caused by the split of exploration/exploitation phases in our algorithm, which resembles the $O(T^{2/3})$ regret for ETC algorithms in classic multiarmed bandit problems (see Chapter 6 in Lattimore & Szepesvári, 2020). We cannot integrate the two phases because of the following barrier: The spectral estimator cannot use samples generated from the belief-based policy due to the history dependency. This is also why Azizzadenesheli et al. (2016) focus on memoryless policies in POMDPs to apply the spectral estimator. In some simpler settings such as linear bandit, the exploration–exploitation interleaving approach can lead to $O(\sqrt{T})$ regret by adaptively varying the length of the phases, see, for example, Rusmevichientong and Tsitsiklis (2010). The reason for this difference in regret lies in the structure of the problem. In Rusmevichientong and Tsitsiklis (2010), the linear parameterization and smoothness assumption on the set of actions give rise to an instantaneous regret that is proportional to $\|\text{estimation error}\|_2^2$ in the exploitation phase, while for us it is proportional to $\|\text{estimation error}\|_2$.

5 | NUMERICAL EXPERIMENTS

In this section, we present proof-of-concept experiments to demonstrate the performance of the SEEU algorithm. Note that for large-scale POMDPs, it is computationally expensive even to solve the oracle, that is, finding the optimal policy to maximize the long-run average reward in a known environment. Therefore, we focus on small-scale experiments, following some recent literature on reinforcement learning for POMDPs (Azizzadenesheli et al., 2016; Igl et al., 2018).

As a representative example, we consider a POMDP model (1) with two hidden states, two actions, and two possible observations where the model parameters are given as follows.

- The transition matrices $P^{(i)}$ for action $i = 1, 2$, are $P^{(1)} = \begin{pmatrix} 0.2 & 0.8 \\ 0.9 & 0.1 \end{pmatrix}$ and $P^{(2)} = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}$;
- the observation densities are $\Omega^{(1)} = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$ and $\Omega^{(2)} = \begin{pmatrix} 0.2 & 0.8 \\ 0.9 & 0.1 \end{pmatrix}$ where $\Omega^{(i)}(m, o) = \Omega(o|m, i)$;

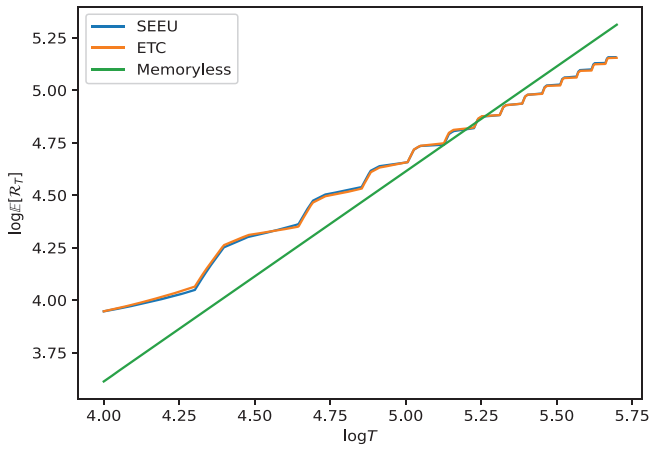

FIGURE 2 Regret comparison of different algorithms [Color figure can be viewed at wileyonlinelibrary.com]

- the reward function is $R = \begin{pmatrix} 1 & 4 \\ 3 & 2 \end{pmatrix}$, where $R(m, i)$ denotes the reward for the pair (m, i).

We compare the regret of our algorithm with two other benchmarks. First, we implement the ETC-type algorithm mentioned in Section 3.3. Second, we also implement the optimal memoryless policy for POMDPs with known parameters discussed in Azizzadenesheli et al. (2016). A memoryless policy maps the observation in the current period to an action. In Figure 2, we plot the average regret versus T in the log-log scale. Due to computational cost of the search for the optimistic parameters in the confidence region for the SEEU algorithm, we run 100 replications of three algorithms. The relative standard error (the standard error divided by the estimated mean) of all three methods are less than 10%. For SEEU and ETC algorithms, we choose the hyperparameters $\tau_1 = 5000$ and $\tau_2 = 10,000$. The choices of these parameters do not affect the order of the regret as shown in our theoretical results. We can observe from Figure 2 that the memoryless policy suffers linear regret. This is expected because the optimal policy for such a POMDP is a belief-based policy which relies on the historical observations and memoryless policies are generally suboptimal. On the other hand, both the SEEU and the ETC algorithm have sublinear growth of regret, where the slopes of the corresponding curves are both close to $2/3$. The similar trends between SEEU and ETC are due to fact that they both share the same nested structure of exploration and exploitation phases.

6 | CONCLUSION AND FUTURE WORK

This paper studies learning of POMDPs in an unknown environment under the average-reward criterion. We develop a learning algorithm that integrates spectral estimators for HMMs and upper confidence methods from online learning. We also establish a regret bound of order of $O(T^{2/3}\sqrt{\log T})$ for the learning algorithm.

There are two research directions based on the work that is worth exploring. First, it is not clear if other algorithms can attain regret of order $\sqrt{T}$ or the regret lower bound is $T^{2/3}$. A related open problem is whether spectral methods can be applied to samples generated from adaptive policies, so that exploration and exploitation can be integrated to improve the theoretical regret bound. We hope to address both problems in future studies.

ACKNOWLEDGMENTS

We thank the department editor, the senior editor, and two referees for many constructive comments, which have led to an improved version of the paper. The research of Ningyuan Chen is supported by NSERC Discovery Grants Program RGPIN-2020-04038. The research of Xuefeng Gao is supported by Hong Kong RGC GRF (No. 14201520 and No. 14201421).

ENDNOTES

[1] There are recent advances on the expectation-maximization (EM) algorithm that are applied to likelihood-based methods for HMMs (Balakrishnan et al., 2017; Yang et al., 2017) with finite-sample analysis. However, the conditions on the Q function and the resulting basin of attraction are hard to translate to our setting explicitly.

[2] For any vectors $v \in \mathbb{R}^{n_1}, u \in \mathbb{R}^{n_2}, w \in \mathbb{R}^{n_3}$, the tensor products are defined as follows: $v \otimes u \in \mathbb{R}^{n_1 \times n_2}$ with $[v \otimes u]_{i,j} = v_i u_j$, and $v \otimes u \otimes w \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with $[v \otimes u \otimes w]_{i,j,k} = v_i u_j w_k$.

REFERENCES

Agrawal, S., & Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, *30*, 1184–1194.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., & Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, *15*, 2773–2832.

Anandkumar, A., Hsu, D., & Kakade, S. M. (2012). A method of moments for mixture models and hidden Markov models. In Mannor, Shie, Srebro, Nathan & Williamson, Robert C. (Eds.), *Conference on learning theory* (Vol. 23, pp. 33.1–33.34), PMLR.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2), 235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, *32*(1), 48–77.

Auer, P., Gajane, P., & Ortner, R. (2019). Adaptively tracking the best bandit arm with an unknown number of distribution changes. In Beygelzimer, Alina & Hsu, Daniel (Eds.), *Conference on learning theory* (Vol. 99, pp. 138–158), PMLR.

Auer, P., & Ortner, R. (2006). Logarithmic online regret bounds for undiscounted reinforcement learning. In B. Schölkopf, J. Platt & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems*, (Vol. 19, pp. 49–56), MIT Press.

Azar, M. G., Osband, I., & Munos, R. (2017). Minimax regret bounds for reinforcement learning. In Precup, Doina & Teh, Yee Whye (Eds.), *International conference on machine learning* (Vol. 70, pp. 263–272), PMLR.

Azizzadenesheli, K., Lazaric, A., & Anandkumar, A. (2016). Reinforcement learning of POMDPs using spectral methods. In Feldman, Vitaly, Rakhlin,

Alexander & Shamir, Ohad (Eds.), *Conference on learning theory* (Vol. 49, pp. 193–256), PMLR.

Balakrishnan, S., Wainwright, M., & Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, *45*(1), 77–120.

Besbes, O., Gur, Y., & Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, *27*, 199–207.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, *5*(1), 1–122.

Cao, X., & Guo, X. (2007). Partially observable Markov decision processes with reward information: Basic ideas and models. *IEEE Transactions on Automatic Control*, *52*(4), 677–681.

Chen, B., Chao, X., & Ahn, H. (2019). Coordinating pricing and inventory replenishment with nonparametric demand learning. *Operations Research*, *67*(4), 1035–1052.

Chen, N., Wang, C., & Wang, L. (2021). *Learning and optimization with seasonal patterns*. ArXiv preprint arXiv:2005.08088.

Chen, W., Shi, C., & Duenyas, I. (2020). Optimal learning algorithms for stochastic inventory systems with random capacities. *Production and Operations Management*, *29*(7), 1624–1649.

Cheung, W., Simchi-Levi, D., & Zhu, R. (2019). *Non-stationary reinforcement learning: The blessing of (more) optimism*. Available at SSRN 3397818.

Cheung, W., Simchi-Levi, D., & Zhu, R. (2022). Hedging the drift: Learning to optimize under nonstationarity. *Management Science*, *68*(3), 1696–1713.

De Castro, Y., Gassiat, E., & Le Corff, S. (2017). Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Transactions on Information Theory*, *63*(8), 4758–4777.

Fiez, T., Sekar, S., & Ratliff, L. (2018). *Multi-armed bandits for correlated Markovian environments with smoothed reward feedback*. ArXiv preprint arXiv:1803.04008.

Garivier, A., & Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In Kivinen, Jyrki, Szepesvári, Csaba, Ukkonen, Esko & Zeugmann, Thomas (Eds.), *International conference on algorithmic learning theory* (pp. 174–188), Springer Berlin Heidelberg.

Guha, S., Munagala, K., & Shi, P. (2010). Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM)*, *58*(1), 1–50.

Guo, Z., Doroudi, S., & Brunskill, E. (2016). A PAC RL algorithm for episodic POMDPs. In Gretton, Arthur & Robert, Christian C. (Eds.), *International conference on artificial intelligence and statistics* (pp. 510–518), PMLR.

Hausknecht, M., & Stone, P. (2015). Deep recurrent q-learning for partially observable MDPS. In *Association for the advancement of artificial intelligence fall symposium series* (pp. 29–37).

Hinderer, K. (2005). Lipschitz continuity of value functions in Markovian decision processes. *Mathematical Methods of Operations Research*, *62*(1), 3–22.

Hsu, S., Chuang, D., & Arapostathis, A. (2006). On the existence of stationary optimal policies for partially observed MDPS under the long-run average cost criterion. *Systems & Control Letters*, *55*(2), 165–173.

Igl, M., Zintgraf, L., Le, T., Wood, F., & Whiteson, S. (2018). Deep variational reinforcement learning for POMDPs. In Dy, Jennifer & Krause, Andreas (Eds.), *International conference on machine learning* (pp. 2117–2126), PMLR.

Jaksch, T., Ortner, R., & Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *The Journal of Machine Learning Research*, *11*, 1563–1600.

Jin, C., Allen-Zhu, Z., Bubeck, S., & Jordan, M. (2018). Is Q-learning provably efficient? *Advances in Neural Information Processing Systems*, *31*, 4868–4878.

Jin, C., Kakade, S., Krishnamurthy, A., & Liu (2020). Sample-efficient reinforcement learning of undercomplete POMDPs. *Advances in Neural Information Processing Systems*, *33*, 18530–18539.

Keskin, N., & Zeevi, A. (2017). Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research*, *42*(2), 277–307.

Krishnamurthy, V. (2016). *Partially observed Markov decision processes*. Cambridge University Press.

Kwon, J., Efroni, Y., Caramanis, C., & Mannor, S. (2021). Rl for latent MDPS: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, *34*, 24523–24534.

Lakshmanan, K., Ortner, R., & Ryabko, D. (2015). Improved regret bounds for undiscounted continuous reinforcement learning. In Bach, Francis & Blei, David (Eds.), *International conference on machine learning* (pp. 524–532), PMLR.

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lehéricy, L. (2019). Consistent order estimation for nonparametric hidden Markov models. *Bernoulli*, *25*(1), 464–498.

Nambiar, M., Simchi-Levi, D., & Wang, H. (2021). Dynamic inventory allocation with demand learning for seasonal goods. *Production and Operations Management*, *30*(3), 750–765.

Ormoneit, D., & Glynn, P. (2002). Kernel-based reinforcement learning in average-cost problems. *IEEE Transactions on Automatic Control*, *47*(10), 1624–1636.

Ortner, R., & Ryabko, D. (2012). Online regret bounds for undiscounted continuous reinforcement learning. In F. Pereira, C.J. Burges, L. Bottou & K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, (Vol. 25, 1772–1780), Curran Associates, Inc.

Ortner, R., Ryabko, D., Auer, P., & Munos, R. (2014). Regret bounds for restless Markov bandits. *Theoretical Computer Science*, *558*, 62–76.

Puterman, M. (2014). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons.

Ross, S. (1968). Arbitrary state Markovian decision processes. *The Annals of Mathematical Statistics*, *39*(6), 2118–2122.

Ross, S., Pineau, J., Chaib-draa, B., & Kreitmann, P. (2011). A Bayesian approach for learning and planning in partially observable Markov decision processes. *The Journal of Machine Learning Research*, *12*, 1729–1770.

Rusmevichientong, P., & Tsitsiklis, J. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, *35*(2), 395–411.

Saldi, N., Yüksel, S., & Linder, T. (2017). On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. *Mathematics of Operations Research*, *42*(4), 945–978.

Sharma, H., Jafarnia-Jahromi, M., & Jain, R. (2020). Approximate relative value learning for average-reward continuous state MDPS. In Adams, Ryan P. & Gogate, Vibhav (Eds.), *Conference on uncertainty in artificial intelligence* (pp. 956–964), PMLR.

Slivkins, A., & Upfal, E. (2008). Adapting to a changing environment: The Brownian restless bandits. In Rocco, Servedio & Tong, Zhang (Eds.), *Conference on learning theory* (pp. 343–354), PMLR.

Spaan, M. (2012). Partially observable Markov decision processes. *Reinforcement Learning*, 387–414.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Sutton, R., & Barto, A. (2018). *Reinforcement learning: An introduction*. MIT Press.

Yang, F., Balakrishnan, S., & Wainwright, M. (2017). Statistical and computational guarantees for the Baum-Welch algorithm. *The Journal of Machine Learning Research*, *18*(1), 4528–4580.

Yu, H., & Bertsekas, D. (2004). Discretized approximations for POMDP with average cost. In Christopher Meek (Ed.), *Conference on uncertainty in artificial intelligence* (pp. 619–627), AUAI Press.

Yu, H., & Bertsekas, D. (2008). On near optimality of the set of finite-state controllers for average cost POMDP. *Mathematics of Operations Research*, *33*(1), 1–11.

Zhang, H., Chao, X., & Shi, C. (2018). Perishable inventory systems: Convexity results for base-stock policies and learning algorithms under censored demand. *Operations Research*, *66*(5), 1276–1286.

Zhang, H., Chao, X., & Shi, C. (2020). Closing the gap: A learning algorithm for lost-sales inventory systems with lead times. *Management Science*, *66*(5), 1962–1980.

Zhang, Z., & Ji, X. (2019). Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, *32*, 2827–2836.

Zhao, J., Xie, X., Xu, X., & Sun, S. (2017). Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, *38*, 43–54.

Zhou, X., Xiong, Y., Chen, N., & Gao, X. (2021). Regime switching bandits. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang & J. Wortman Vaughan (Eds.), *Advances in Neural Information Processing Systems*, *34*, (pp. 4542–4554).

Zhu, F., & Zheng, Z. (2020). When demands evolve larger and noisier: Learning and earning in a growing environment. In III, Hal Daumé & Singh, Aarti (Eds.), *International conference on machine learning* (Vol. 119, pp. 11629–11638), PMLR.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Xiong, Y., Chen, N., Gao, X., & Zhou, X. (2022). Sublinear regret for learning POMDPs. *Production and Operations Management*, *31,* 3491–3504. https://doi.org/10.1111/poms.13778