Ergodic Unobservable MDPs: Decidability of Approximation

Krishnendu Chatterjee, David Lurie, Raimundo Saona, Bruno Ziliotto

Ergodic Unobservable MDPs: Decidabiliy of Approximation

Krishnendu Chatterjee[1] David Lurie[2] Raimundo Saona[1] Bruno Ziliotto[3]

May 22, 2024

Abstract

Unobservable Markov decision processes (UMDPs) serve as a prominent mathematical framework for modeling sequential decision-making problems. A key aspect in computational analysis is the consideration of decidability, which concerns the existence of algorithms. In general, the computation of the exact and approximated values is undecidable for UMDPs with the long-run average objective. Building on matrix product theory and ergodic properties, we introduce a novel subclass of UMDPs, termed ergodic UMDPs. Our main result demonstrates that approximating the value within this subclass is decidable. However, we show that the exact problem remains undecidable. Finally, we discuss the primary challenges of extending these results to partially observable Markov decision processes.

Keywords. Markov, finite state, decidability, ergodicity, approximation algorithms.

Contents

[1]Institute of Science and Technology Austria, Austria.
[2]Paris Dauphine University and NyxAir, France.
[3]CEREMADE, CNRS, Paris Dauphine University, France.

1 Context

In this study, we consider unobservable Markov decision processes (UMDPs), and their partially observable variants (POMDPs), with the long-run average objective. We introduce a new theoretical class of UMDPs by leveraging the "ergodicity" property of products of matrices. The emphasis of this work is to provide a mathematical proof that analyzes the decidability of this newly defined class. Consequently, we intentionally avoid examining the complexity or efficiency of the proposed algorithm, focusing instead on the theoretical foundations over the practical applications.

UMDPs and POMDPs serve as a powerful model for sequential decision-making problems. UMDPs are equivalent to probabilistic finite automata [7, 27, 29]. In parallel, POMDPs extend the classical model with perfect observations of Markov decision processes (MDPs) [5], which assume perfect information of the system. These models effectively capture the dynamic in state transitions and the unobservability, or partial observability, of the environment. Here, we present UMDPs and POMDPs with finite state space, finite action space, finite signal space, and bounded stage reward functions.

Researchers have increasingly explored the use of UMDPs and POMDPs to solve various real-world problems [8, 25]. Notably, scenarios involving a long duration stand out. In such contexts, the long-run average objective holds particular relevance as it emphasizes long-term optimization. For instance, this objective finds practical applications in numerous fields, such as communication networks and queueing systems [2]. In [29], Rabin also explored probabilistic automata in the context of control theory.

Addressing UMDPs and POMDPs with the long-run average objective presents significant challenges. Computationally, finite-horizon UMDPs are classified as NP-complete, and finite-horizon POMDPs as PSPACE-complete [23]. Moreover, infinite-horizon UMDPs are known as undecidable [21]. The analytical difficulties of the long-run average objective is mainly due to the absence of the contraction property, as seen in the discounted-sum objective. Consequently, this work aims to develop innovative algorithms to overcome these difficulties.

We propose a new approach which relies on the property of "ergodicity" in products of matrices. This concept [17, 20, 32, 33] describes how a system progressively forgets its past. These techniques have found wide applications in the literature, including demographic studies and economic analysis [33]. However, only MDPs have been explored in the context of this hypothesis: see [1, 18, 35] for MDPs with finite state space and [4, 24] for MDPs with denumerable state space.

Main Contributions. In this paper, we present the following main results:

- *New theoretical class.* We identify a novel theoretical class of UMDPs, termed *ergodic UMDPs*, by leveraging the concept of ergodicity in products of stochastic matrices.
- *Sufficient conditions.* We provide sufficient conditions to establish that the class of ergodic UMDPs is not empty.
- *Positive Result.* We demonstrate that the decision version of approximating the value is decidable for the class of ergodic UMDPs with the long-run average objective.
- *Negative Result.* We prove that the decision version of determining the exact value is undecidable for the class of ergodic UMDPs with the long-run average objective.

Significance. To the best knowledge of the authors, this is the first result that exploits ergodic properties and matrix theory to establish the decidability of approximating the value for a subclass of UMDPs. Even in this scenario, the exact problem remains undecidable, underscoring the subtleties of our approach. Furthermore, our research also draws from a wide range of matrix theory literature to establish computational results in UMDPs, a classical problem in control theory.

Methodology. When considering UMDPs, a common approach involves constructing an equivalent completely observable MDP, termed *belief MDP*, where the state space is infinite. In this model, the state corresponds to the belief on the original set of states. In parallel, it is well-known in the literature that MDPs with finite state spaces are not only decidable, but can also be solved efficiently. By leveraging the ergodicity property, we construct a finite-state MDP, termed *abstract MDP*, and prove that it effectively approximates the belief MDP.

Structure of the paper. The structure of the paper is outlined as follows. We start by reviewing UMDPs in Section 2. In Section 3, we present the class of ergodic UMDPs. The decidability of approximating the value is examined in Section 4, while undecidablity aspects are tackled in Section 5. Finally, Section 6 discusses potential challenges with extending our approach to POMDPs.

2 Model Description

Notation. *Sets are represented by calligraphic letters such as $\mathcal{A}, \mathcal{H}, \mathcal{K}$, and $\mathcal{S}$. Elements within these sets are denoted by lowercase letters, such as a, h, k, and s. Random elements are denoted by uppercase letters, such as A, H, K, and S. For a set $\mathcal{C}$, let $\Delta(\mathcal{C})$ be the set of probability measure distribution over $\mathcal{C}$, and let δ_c be the Dirac measure at some element $c \in \mathcal{C}$. For $a, b \in \mathbb{R}$, the set $[a, b] \cap \mathbb{Z}$ is represented as $[a..b]$.*

2.1 Model

In this section, we introduce UMDPs, the primary model discussed throughout this paper. Formally, we define it as follows.

Definition 2.1 (UMDP)**.** *A UMDP, denoted as Γ, is defined by a 5-tuple $\Gamma = (\mathcal{K}, \mathcal{A}, p, g, b_1)$, where:*

- *$\mathcal{K}$ is the finite set of states;*
- *$\mathcal{A}$ is the finite set of actions;*
- *$p\colon \mathcal{K} \times \mathcal{A} \to \Delta(\mathcal{K})$, represented as $p(k'|k, a)$, is the probabilistic transition function that gives the probability distribution over the successor states given a state $k \in \mathcal{K}$ and an action $a \in \mathcal{A}$. We represent by $P(a)$ the transition matrix for each action $a \in \mathcal{A}$;*
- *$g\colon \mathcal{K} \times \mathcal{A} \to [0, 1]$ is the stage reward function;*
- *$b_1 \in \Delta(\mathcal{K})$ is the initial belief, which represents the initial probability distribution over the state space.*

Starting from $b_1 \in \Delta(\mathcal{K})$, the UMDP evolves as follows:

- An initial state k_1 is selected according to b_1. The decision-maker knows b_1 but does not know k_1, the realization of K_1.
- At each stage $m \geq 1$, the decision-maker chooses some action $a_m \in \mathcal{A}$. This action results in a stage reward $G_m := g(k_m, a_m)$. Subsequently, the next state K_{m+1} is determined according to the transition probability function $p(k_m, a_m)$. The decision-maker receives no information about the environment. Therefore, he cannot observe the state K_{m+1} nor the reward G_m.

In contrast with a UMDP, where the decision-maker has no information about the current state, a MDP [28] offers full information over the state space. At all stages, the decision-maker is fully aware of his current state. A MDP can be defined as a 5-tuple $(\mathcal{K}, \mathcal{A}, p, g, b_1)$, where $\mathcal{K}$ is the finite set of states, $\mathcal{A}$ is the finite set actions,

$p\colon \mathcal{K} \times \mathcal{A} \to \Delta(\mathcal{K})$ the transition function, $g\colon \mathcal{K} \times \mathcal{A} \to [0,1]$ the stage reward function, and $b_1 \in \Delta(\mathcal{K})$ the initial belief.

At stage m, the decision-maker recalls all previous actions, which is referred to as the *history before stage m*. Let $\mathcal{H}_m \coloneqq (\mathcal{A})^{m-1}$ define the set of histories before stage m, with $(\mathcal{A})^0 \coloneqq \{\emptyset\}$. A strategy is a mapping $\sigma\colon \bigcup_{m\geq 1} \mathcal{H}_m \to \mathcal{A}$. The set of strategies is denoted by Σ. Given $b_1 \in \Delta(\mathcal{K})$ and $\sigma \in \Sigma$, we define $\mathbb{P}_\sigma^{b_1}$ as the law induced by the strategy σ and the initial belief b_1 on the set of plays of the game $\Omega = (\mathcal{K} \times \mathcal{A})^{\mathbb{N}}$. Similarly, $\mathbb{E}_\sigma^{b_1}$ represents the expectation with respect to this law.

Remark 2.1. *There exists strategies, known as behavioural strategies, that consider a probability distribution over the action space $\mathcal{A}$. These strategies are defined as $\sigma\colon \bigcup_{m\geq 1} \mathcal{H}_m \to \Delta(\mathcal{A})$. When $\Delta(\mathcal{A})$ is a dirac δ_a over some action a, such strategies are referred to as pure strategies. Kuhn's Theorem establishes a connection between pure and behavioural strategies, leading to the conclusion that using behavioural strategies does not alter the results presented in this paper (see [14, 34]).*

Let $\gamma(\sigma)$ represent the long-run average reward given by some strategy $\sigma \in \Sigma$ as

$$\gamma(\sigma) \coloneqq \liminf_{N\to\infty} \mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N} G_m\right), \tag{1}$$

and v denote the optimal long-run average reward, called value, defined as

$$v \coloneqq \sup_{\sigma\in\Sigma} \gamma(\sigma), \tag{2}$$

where the supremum is taken over all strategies $\sigma \in \Sigma$.

Remark 2.2. *In the literature, the long-run average objective is also commonly defined as*

$$v \coloneqq \sup_{\sigma\in\Sigma} \mathbb{E}_\sigma^{b_1}\left(\liminf_{N\to\infty} \frac{1}{N}\sum_{m=1}^{N} G_m\right).$$

These values coincide, as proven in [34].

Next, we introduce the definition of ε-optimal strategies as follows.

Definition 2.2 (ε-Optimal Strategy)**.** *Let $b_1 \in \Delta(\mathcal{K})$ and $\varepsilon > 0$. A strategy $\sigma \in \Sigma$ is ε-optimal in Γ if*

$$\gamma(\sigma) \geq v - \varepsilon. \tag{3}$$

2.2 Computational Formalism

In computational theory, a *decision problem* is a binary question that determines whether a specific property holds for a given input. The decidability of these problems is characterized by Turing machines: they process an input and, upon reaching a halting state, accept or reject it. Algorithms are Turing machines that halt for all possible inputs within a finite number of stages. If it takes the correct decision for all input, the algorithm is said to solve a decision problem. The class of decision problems solvable by an algorithm is referred to as being *decidable*. Conversely, if no algorithm exists to solve a decision problem, it belongs to the *undecidable* class. In [21], Madani et al. established the undecidability of determining the value for UMDPs with the long-run average objective. This decision problem can be defined as follows.

Definition 2.3 (Decision Version of Determining the Value)**.** *Let $\Gamma = (\mathcal{K}, \mathcal{A}, p, g, b_1)$ be a UMDP. Given $x \in [0,1]$, the problem consists in deciding which one is the case: to accept means that $v > x$ holds, whereas to reject means that $v \leq x$ holds.*

In many real-world scenarios, finding exact solutions can be very challenging. To address this issue, approximation algorithms offer practical and efficient solutions by finding ε-optimal solutions with guaranteed performance bounds. The study of the approximability of decision problems involves analyzing the trade-off between the quality of the solution and the computational resources. The decision problem of approximating the value can be defined as follows.

Definition 2.4 (Decision Version of Approximating the Value)**.** *Let $\Gamma = (\mathcal{K}, \mathcal{A}, p, g, b_1)$ be a UMDP. Given $x \in [0,1]$, $\varepsilon > 0$ such that $v > x + \varepsilon$ or $v < x - \varepsilon$, the problem consists in deciding which one is the case: to accept means that $v > x + \varepsilon$ holds, whereas to reject means that $v < x - \varepsilon$ holds.*

In the context of UMDPs, Madani et al. [21] have proved that even the decision version of the approximation problem remains undecidable for the long-run average objective. As a consequence, it becomes essential to propose conditions that define decidable subclasses of UMDPs.

2.3 Reduction of UMDPs to Belief MDPs

Recall that the state dynamic in UMDPs is unobservable: the decision-maker is "blind" over the states of the system. In contrast, the classical MDP model assumes that the decision-maker observes the state of the system at the beginning of each stage.

We introduce the definition of history before stage m as follows.

Definition 2.5 (m-Stage History)**.** *Given a strategy $\sigma \in \Sigma$ and an initial belief $b_1 \in \Delta(\mathcal{K})$, denote the (random) history at stage m by*

$$H_m \coloneqq (A_1, A_2, ..., A_{m-1}).$$

The random variable H_m takes values in $\mathcal{H}_m$.

In UMDPs, the decision-maker remembers the history of its past actions when deciding on a new action. Unfortunately, the representation of past histories is expensive. Indeed, the set of possible histories up to stage m grows exponentially with m, while for infinite-horizon POMDPs, the set of histories becomes infinite. An alternative approach was introduced by Åström in [3]. He proved that it is possible to summarize all information from past actions into a probability distribution (the belief) over the state space $\mathcal{K}$: the belief is a sufficient statistic for a given history $H_m = h_m$. As a result, a strategy can take the form of a mapping from "states" to actions if the states of the model are defined differently.

We consider the definition of the m-stage belief in UMDPs as follows.

Definition 2.6 (m-Stage Belief)**.** *Given an initial belief $b_1 \in \Delta(\mathcal{K})$, a strategy $\sigma \in \Sigma$, and a history $h_m \in \mathcal{H}_m$, denote the belief at stage m by $b^{b_1}_{m,\sigma}$, which is given by, for all $k \in \mathcal{K}$,*

$$b^{b_1}_{m,\sigma}(k) \coloneqq \mathbb{P}^{b_1}_{\sigma}(K_m = k | H_m = h_m).$$

For clarity, we will simplify the notation of the belief as b_m, omitting its dependence on b_1 and σ. Given a fixed σ and b_1, one can use Bayes' rule to compute b_m.

When dealing with UMDPs, a common approach is to construct a completely observable MDP $\mathcal{G}$, referred to as the *belief MDP*. The observability stems from the fact that the beliefs are functions of the set of histories and the initial belief. This method offers several advantages, such as relying on the more developed theory of completely observable MDPs, including optimality equations or structures of optimal strategies. Unfortunately, the state space of this belief MDP is infinite (cf. [15]).

Consider a UMDP $\Gamma = (\mathcal{K}, \mathcal{A}, p, g, b_1)$. The belief MDP, denoted as $\mathcal{G}$, is defined by a 5-tuple $\mathcal{G} = (\Delta(\mathcal{K}), \mathcal{A}, \bar{p}, \bar{g}, b_1)$, where:

- $\Delta(\mathcal{K})$ is the countable set of belief states;
- $\mathcal{A}$ is the finite set of actions;
- $\bar{p}\colon \Delta(\mathcal{K}) \times \mathcal{A} \to \Delta(\mathcal{K})$, denoted $\bar{p}(b'|b,a)$, is the deterministic transition function that gives the probability distribution over the successor belief states given a belief state $b \in \Delta(\mathcal{K})$ and an action $a \in \mathcal{A}$;
- $\bar{g}\colon \Delta(\mathcal{K}) \times \mathcal{A} \to [0,1]$ is the stage reward function such that for all $b \in \Delta(\mathcal{K})$ and $a \in \mathcal{A}$, $\bar{g}(b,a) \coloneqq \sum_{k \in \mathcal{K}} b(k) g(k,a)$;
- $b_1 \in \Delta(\mathcal{K})$ is the initial belief state.

At each stage, the belief changes each time an action is taken. For each stage $m \geq 1$ and state $k' \in \mathcal{K}$, the *belief update* is defined as:

$$\begin{aligned} b_{m+1}(k') :=& \tau(b_m, a_m) \\ =& \sum_{k \in \mathcal{K}} p(k'|k, a_m) b_m(k) \end{aligned}$$

and the *deterministic transition function* is defined as:

$$\bar{p}(b_{m+1}|b_m, a_m) := \begin{cases} 1 & \text{if } b_{m+1} = \tau(b_m, a_m) \\ 0 & \text{otherwise} \end{cases}$$

For each stage $m \geq 1$, let $\bar{G}_m := \bar{g}(B_m, A_m)$ denote the stage reward where $B_m \in \Delta(\mathcal{K})$ and $A_m \in \mathcal{A}$.

For a given strategy $\sigma \in \Sigma$, the long-run average objective of the belief MDP is defined as follows

$$\gamma'(\sigma) := \liminf_{N \to \infty} \mathbb{E}_{\sigma}^{b_1} \left(\frac{1}{N} \sum_{m=1}^{N} \bar{G}_m \right),$$

and the value v' is defined as

$$v' := \sup_{\sigma} \gamma(\sigma),$$

where the supremum is taken over all strategies $\sigma \in \Sigma$.

As evidenced by prior works [6, 13, 30, 37]: given a strategy $\sigma \in \Sigma$ and initial belief $b_1 \in \Delta(\mathcal{K})$, the long-run average objective of the belief MDP is equal to the long-run average objective of the corresponding UMDP. Similarly, given an initial belief $b_1 \in \Delta(\mathcal{K})$, the value of the belief MDP and the value of the corresponding UMDP are equivalent. Finally, given its equivalence to the original UMDP, an optimal solution to the belief MDP coincides with an optimal solution for the original UMDP.

3 Class Description

This section presents a novel class of UMDPs that leverages the property of ergodicity of products of stochastic matrices. Ergodicity [17, 32, 33] focuses on the progression of forward products of stochastic matrices towards equivalence of the rows. This property essentially describes a tendency to overlook its distant past [17]. Building upon this concept, we introduce a new class of UMDPs, referred to as *ergodic UMDPs*. Additionally, we will present sufficient conditions for specific subclasses of ergodic UMDPs.

To begin, given $n \geq 1$ and action sequence $a^n = (a_1, ..., a_n)$, define the forward products of transition matrices as

$$T^n(a^n) := P(a_1)P(a_2)\cdots P(a_n) = \prod_{i=1}^{n} P(a_i).$$

We say that $P > 0$ if $p_{k,k'} > 0$ for each k, k'. Similarly, we write $P \geq 0$ if $p_{k,k'} \geq 0$ for each k, k'. When it is clear from the context, we denote $T^n(a^n) = T^n = (t^n_{k,k'})_{k,k' \in \mathcal{K}}$. For a given matrix P, we represent its k-th column as $(P)_{k \in \mathcal{K}}$. Finally, the transpose of a vector p will be denoted as $p^\top$. A square matrix P is stochastic if $p_{k,k'} \geq 0$ for each $k, k' \in \mathcal{K}$, and the terms of each row sum to one i.e., $\sum_{k' \in \mathcal{K}} p_{k,k'} = 1$.

We define the ergodicity of products of stochastic matrices as follows [33, Definition 4.4, p. 136].

Definition 3.1 (Ergodicity)**.** *A sequence of stochastic matrices $\{P_i\}_{i \geq 1}$ on $\mathcal{K}$ is ergodic if we have, for all $k, \bar{k}, k' \in \mathcal{K}$, that*

$$t^n_{k,k'} - t^n_{\bar{k},k'} \longrightarrow 0, \tag{4}$$

as n goes to infinity.

Remark 3.1. *In the literature, Definition 3.1 is commonly known as the weak ergodicity property of forward products of matrices [32].*

The "strong" ergodicity of product of stochastic matrices [33, Definition 4.5, p. 136] necessitates entrywise convergence, that is, each term $t^n_{k,k'}$ must converge to a limit as $n \to \infty$. In contrast, Definition 3.1 emphasizes a convergence based on the differences between rows. Specifically, it indicates that the values of $t^n_{k,k'}$ for each $k, k' \in \mathcal{K}$ may not necessarily converge to a limit as $n \to \infty$.

Coefficient of ergodicity has been introduced as a tool to characterize the convergence speed of forward products of matrices. For a deeper dive into this topic, we refer the reader to the following papers [19, 22, 33]. A stochastic matrix P is called *stable* if every rows are identical.

We present a formal definition of coefficient of ergodicity for stochastic matrices [33, Definition 4.6, p. 136].

Definition 3.2 (Coefficient of Ergodicity)**.** *A scalar function $\tau(\cdot)$, continuous on the set of stochastic matrices and satisfying $0 \leq \tau(P) \leq 1$, is called a coefficient of ergodicity. It is proper if*

$$\tau(P) = 0 \text{ if and only if } P = \mathbf{1}v^\top,$$

where $\mathbf{1} = (1, \ldots, 1)$ and v is some probability vector ($v \geq 0$, $v^\top \mathbf{1} = 1$) that is, whenever the matrix P is stable.

Given a stochastic matrix P, an example of proper coefficient of ergodicity (see [33]), denoted as $\tau_1(\cdot)$, is defined by

$$\tau_1(P) \coloneqq \frac{1}{2} \max_{k,\bar{k}} \sum_{k'=1}^{|\mathcal{K}|} |p_{k,k'} - p_{\bar{k},k'}|.$$

Using Seneta [33, Lemma 4.3, p. 139], the coefficient of ergodicity τ_1 is submultiplicative, i.e., for all stochastic matrices P and Q, we have that $\tau_1(PQ) \leq \tau_1(P)\tau_1(Q)$. The coefficient of ergodicity τ_1 plays a crucial role in characterizing ergodicity [33, p. 140]. More specifically, ergodicity of forward products of stochastic matrices is equivalent to

$$\tau_1(T^n) \longrightarrow 0,$$

as n goes to infinity. Using this definition, we identify a novel theoretical class of UMDPs, referred to as *ergodic UMDPs*.

Definition 3.3 (Ergodic UMDP)**.** *A UMDP Γ is ergodic if, for all $\varepsilon \in (0,1)$, there exists an integer n_0 such that, for all action sequences a^n with $n \geq n_0$, the following inequality holds:*

$$\tau_1(T^n(a^n)) \leq \varepsilon. \tag{5}$$

The existence of an integer n_0, such that inequality (5) holds for each sequence of actions a^{n_0}, is crucial in analyzing the decidability of the class of ergodic UMDPs.

Remark 3.2. *Definition 3.3 of ergodic UMDP is equivalent to the concept of ergodic probabilistic finite automata introduced by Paz in [25]. Here, it displays an integer n_0 that is uniform. Definition 3.1 characterizes the ergodicity property using pointwise convergence. Similarly, we could call a UMDP Γ pointwise ergodic if, for each given action sequence, forward products of transition matrices satisfies the ergodicity condition. By [12, Theorem 6.1], this is equivalent to Definition 3.3.*

Paz provided a characterization of ergodic UMDPs [25, Theorem 3.1, p. 80].

Theorem 3.4. *A UMDP is ergodic if and only if there is n_0 such that for each action sequence a^{n_0} we have $\tau_1(T^{n_0}(a^{n_0})) < 1$.*

Having characterized ergodic UMDPs, we now provide interesting subclasses. In [33], Seneta provided conditions for ergodicity of forward products of stochastic matrices. Therefore, our objective is to highlight connections between them and subclasses of ergodic UMDPs which are of particular interest. We start by introducing the following class of matrices.

Definition 3.5 ([11, 25, 33, 36])**.**

- *A matrix P is stochastic indecomposable and aperiodic (SIA), if $\lim_{n\to\infty} P^n = Q$ exists, where Q is a stable stochastic matrix. The class of SIA matrices is denoted G_1.*
- *A stochastic matrix P is a G_2-matrix if $P \in G_1$ and, for all $Q \in G_1$, we have that $QP \in G_1$. Denote G_2 the class of G_2-matrices.*

- *For all $\mathcal{Q} \subseteq \mathcal{K}$ and stochastic matrix P, define the consequence function F as $F_P(\mathcal{Q}) = \{k' | \exists k \in \mathcal{Q} \text{ s.t. } p_{k,k'} > 0\}$. A stochastic matrix P is a Sarymsakov matrix if for all two disjoint nonempty sets $\mathcal{Q}, \mathcal{Q}' \subseteq \mathcal{K}$, $F_P(\mathcal{Q}) \cap F_P(\mathcal{Q}') \neq \emptyset$ or $|F_P(\mathcal{Q}) \cup F_P(\mathcal{Q}')| > |\mathcal{Q} \cup \mathcal{Q}'|$. Denote S the class of Sarymsakov matrices.*
- *A stochastic matrix P is scrambling if given two rows k and $\bar{k}$, there is at least one column k' such that $p_{k,k'} > 0$ and $p_{\bar{k},k'} > 0$, or equivalently $P \in G_3$ if and only $\tau_1(P) < 1$. Denote G_3 the class of scrambling matrices.*
- *A stochastic matrix P is Markov if at least one column of P has all entries strictly positive. Denote M the class of Markov matrices.*

In the following definition, we introduce a condition commonly referred to as the Wolfowitz Condition, denoted by (WC) in the literature and discussed in [25].

Definition 3.6 (SIA UMDP)**.** *A UMDP is SIA if condition (WC) holds i.e., for all $n \geq 1$ and action sequence a^n, the matrix $T^n(a^n)$ belongs to the class of matrices G_1.*

A SIA UMDP is ergodic. Indeed, under condition (WC), there is an integer n_0 such that for all action sequences a^{n_0} the inequality $\tau_1(T^{n_0}(a^{n_0})) < 1$ holds [26]. Therefore, the ergodicity holds by Theorem 3.4.

From [33], we have $M \subsetneq G_3 \subsetneq S \subsetneq G_2 \subsetneq G_1$. The class of Sarymsakov matrices [33] represents the largest known subclass within G_1 that possesses the property of being closed under multiplication. Therefore, we can construct subclasses of ergodic UMDPs under the assumption that, for each action $a \in \mathcal{A}$, the transition matrix $P(a)$ belongs to M, G_3, S, or G_2.

Example 3.1 (Convex Combination)**.** *The strict convex combination of a UMDP and an ergodic UMDP with the same sets and stage rewards results in an ergodic UMDP. More formally, for each $\alpha \in (0,1)$, consider a UMDP $\Gamma = (\mathcal{K}, \mathcal{A}, p, g, b_1)$ and an ergodic UMDP $\Gamma' = (\mathcal{K}, \mathcal{A}, p', g, b_1)$. For example, one may have that the transitions in Γ' satisfy $p'(k'|k,a) > 0$ for all $k, k' \in \mathcal{K}$ and $a \in \mathcal{A}$. We define the combined UMDP, $\Gamma'' = (\mathcal{K}, \mathcal{A}, p'', g, b_1)$, by setting, for each action a, that $p''(k'|k,a) = \alpha p(k'|k,a) + (1-\alpha)p'(k'|k,a))$. This approach ensures that Γ'' satisfies the ergodicity property, as it keeps the transition matrix structure of Γ''.*

4 Decidability Analysis

In this section, we consider the decision version of approximating the value for the class of ergodic UMDPs. The main contribution of this section is the following theorem.

Theorem 4.1. *The decision version of approximating the value for the class of ergodic UMDPs is decidable.*

To tackle this problem, we present a new approximation scheme for computing the ε-approximate value for the class of ergodic UMDPs.

Interpretation. The goal is to construct a finite-state MDP whose value differs by at most ε from the value of the UMDP. This approach can be considered as an aggregation scheme, where similar beliefs of the ergodic UMDPs are approximated by an "abstract" belief in the finite-state MDP. A related interpretation has been discussed in the robust MDP literature, such as in [16].

We remind that the belief at stage $(n+1)$ after actions $a_1, \ldots, a_n$ can be expressed in "matrix" form as

$$b_{n+1}^\top = b_1^\top T^n(a^n) = b_1^\top P(a_1) \cdots P(a_n) \tag{6}$$

with $a^n = (a_1, ..., a_n)$.

4.1 Determining the Ergodicity Property

A legitimate question would be whether the ergodicity property holds for a given UMDP. We show that it is decidable and that it can be done within exponential space. This claim is supported by the following result from Paz [25, Theorem 4.7, p. 90].

Theorem 4.2. *A UMDP Γ is ergodic if and only if, there exists an integer $n_0 \leq \left(3^{|\mathcal{K}|} - 2^{|\mathcal{K}|+1} + 1\right)/2$ such that, for every action sequence a^n with $n \geq n_0$,*

$$\tau_1(T^n(a^n)) < 1. \tag{7}$$

Building on Theorem 4.2, we deduce the following proposition.

Proposition 4.3. *Let Γ be a UMDP. Determining whether the ergodicity property holds for Γ is decidable. Moreover, it can be achieved within exponential space.*

Proof. Although Theorem 4.2 states a condition for all sequences of actions a^n with $n \geq n_0$, it is sufficient that the condition holds for sequences of actions of length n_0 only. Therefore, Theorem 4.2 immediately implies an algorithm to decide whether a UMDP satisfies the ergodicity property. Indeed, it is sufficient to verify if there is $n_0 \leq \left(3^{|\mathcal{K}|} - 2^{|\mathcal{K}|+1} + 1\right)/2$ such that for all action sequences of length n_0 we have that $\tau_1(T^{n_0}(a^{n_0})) < 1$, which can be done by enumeration.

Moreover, checking whether a UMDP satisfies the ergodicity property requires exponential space complexity. Indeed, one needs to verify through enumeration whether $\tau_1(T^n(a^n)) < 1$ is satisfied for every action sequences of size $n \leq (3^{|\mathcal{K}|} - 2^{|\mathcal{K}|+1} + 1)/2$. □

4.2 Construction of the Abstract MDP

We proceed to construct a finite-state MDP, termed *abstract MDP*. Consider an ergodic UMDP $\Gamma = (\mathcal{K}, \mathcal{A}, p, g, b_1)$ and $\varepsilon \in (0,1)$.

First, the following statement allows us to relate Theorem 4.2 with every $\varepsilon \in (0,1)$.

Proposition 4.4. *Given an ergodic UMDP Γ and $\varepsilon \in (0,1)$, consider $n_0 \leq \left(3^{|\mathcal{K}|} - 2^{|\mathcal{K}|+1} + 1\right)/2$ such that (7) is satisfied for every action sequence a^{n_0}. Then, there exists a finite integer $n \leq \bar{n}(n_0, \varepsilon) n_0/2$ such that for every action sequence $a^{n_1} = (a_1, ..., a_{n_1})$ with $n_1 \geq n$, we have*

$$\tau_1(T^{n_1}(a^{n_1})) \leq \varepsilon\,, \tag{8}$$

where $\bar{n}(n_0, \varepsilon) \coloneqq \lceil \ln(\varepsilon)/\ln(\sup_{a^{n_0}} \tau_1(T^{n_0}(a^{n_0}))) \rceil$.

Proof of Proposition 4.4. By Theorem 4.2, a UMDP is ergodic if and only if there exists $n_0 \leq (3^{|\mathcal{K}|} - 2^{|\mathcal{K}|+1} + 1)/2$ such that for every action sequence a^{n_0} we have

$$\tau_1(T^{n_0}(a^{n_0})) < 1.$$

Now, consider the set of products of stochastic matrices $T^{n_0}(a^{n_0})$ where a^{n_0} is every action sequences of length n_0. Denote $\bar{a}^{n_0}$ the sequence of actions of length n_0 that maximizes $\tau_1(T^{n_0}(a^{n_0}))$, i.e., $\bar{a}^{n_0} \coloneqq \operatorname{argmax}_{a^{n_0}} \tau_1(T^{n_0}(a^{n_0}))$ and $\bar{\tau}(n_0) \coloneqq \tau_1(T^{n_0}(\bar{a}^{n_0}))$. By Theorem 4.2, we have that $\bar{\tau}(n_0) < 1$. For every $\varepsilon \in (0,1)$, taking $\bar{n}(n_0, \varepsilon) = \lceil \ln(\varepsilon)/\ln(\bar{\tau}(n_0)) \rceil$, we have

$$[\bar{\tau}(n_0)]^{\bar{n}(n_0,\varepsilon)} \leq \varepsilon.$$

Also, for all action sequences of length $\bar{n}(n_0, \varepsilon) n_0$, it holds that

$$\tau_1\left(T^{\bar{n}(n_0,\varepsilon)n_0}\left(a^{\bar{n}(n_0,\varepsilon)n_0}\right)\right) \leq [\bar{\tau}(n_0)]^{\bar{n}(n_0,\varepsilon)}$$

by submultiplicativity of the coefficient of ergodicity τ_1. Therefore, the result follows. □

Given $\varepsilon \in (0,1)$, define $\mathcal{T}(\varepsilon)$ as the finite set of forward products of transition matrices satisfying (8) of length n where n is given by Proposition 4.4. In particular, for all $T^n(a^n) \in \mathcal{T}(\varepsilon)$, we have that $\tau_1(T^n(a^n)) \leq \varepsilon$.

We construct an "abstract" set of stable matrices, denoted by $\widetilde{\mathcal{T}}(\varepsilon)$ which approximate $\mathcal{T}(\varepsilon)$. Each matrix is approximated by another with equal rows corresponding to the average over each row. Formally, given every action sequence a^n, consider the matrix $T^n(a^n) \in \mathcal{T}(\varepsilon)$. We define $\widetilde{T}^n(a^n) \in \widetilde{\mathcal{T}}(\varepsilon)$ by

$$\tilde{t}^n_{k,k'}(a^n) = \frac{1}{|\mathcal{K}|} \sum_{\bar{k}=1}^{|\mathcal{K}|} t^n_{\bar{k},k'}(a^n).$$

Note that a matrix $\widetilde{T}^n \in \widetilde{\mathcal{T}}(\varepsilon)$ is stable, i.e., all rows are equal. In the context of UMDPs, we can interpret a stable matrix as a belief state. Indeed, given each initial belief $b \in \Delta(\mathcal{K})$ and matrix $\widetilde{T}^n \in \widetilde{\mathcal{T}}(\varepsilon)$, the belief update is $b'(k') = (b\widetilde{T}^n)_{k'} = \sum_{k\in\mathcal{K}} b(k)\tilde{t}^n_{k,k'}$ for each $k' \in \mathcal{K}$. By definition of stable matrices, for each column k', the transition $\tilde{t}^n_{k,k'}$ is constant in the row k. Thus, the belief update becomes a convex combination of terms with equal values. Consequently, the belief update is independent of the initial belief, and each stable matrix can be interpreted as a mapping to a single belief. As a result, the stable property of stochastic matrices becomes fundamental to the finiteness of the state space in the abstract MDP.

Now, for a given initial belief $b_1 \in \Delta(\mathcal{K})$, define the set of abstract beliefs as follows:

$$\mathcal{B}^* \coloneqq \{b^* \in \Delta(\mathcal{K}) \,|\, \exists \widetilde{T}^n \in \widetilde{\mathcal{T}}(\varepsilon) \text{ such that } b^* = b_1^\top \widetilde{T}^n\} \cup \{b_1\}.$$

For $i \in [1..n-1]$, we will write $\mathcal{B}^* \times \mathcal{A}^i \coloneqq \{(b^*, a_1, ..., a_i) | b^* \in \mathcal{B}^* \text{ and } a_j \in \mathcal{A} \text{ for } j \in [1..i]\}$, and when $i = 0$, $\mathcal{B}^* \times \mathcal{A}^0 \coloneqq \{(b^*) | b^* \in \mathcal{B}^*\}$ for convenience.

Consider an ergodic UMDP $\Gamma = (\mathcal{K}, \mathcal{A}, p, g, b_1)$. The abstract MDP, denoted $\mathcal{G}^*(\varepsilon)$, is defined by a 5-tuple $\mathcal{G}^*(\varepsilon) = (\mathcal{X}, \mathcal{A}, \bar{p}^*, \bar{g}^*, b_1)$, where:

- $\mathcal{X}$ is the finite set of states, defined by
$$\mathcal{X} \coloneqq \bigcup_{i=0}^{n-1} (\mathcal{B}^* \times \mathcal{A}^i);$$
- $\mathcal{A}$ is the finite set of actions;
- $\bar{p}^* \colon \mathcal{X} \times \mathcal{A} \to \mathcal{X}$ is the deterministic transition function that gives the successor state according to current state and action;
- $\bar{g}^* \colon \mathcal{X} \times \mathcal{A} \to [0, 1]$ is the stage reward function;
- $b_1 \in \Delta(\mathcal{K})$ is the initial state.

Define $proj \colon \mathcal{X} \to \Delta(\mathcal{K})$ the function that assigns a belief state to each state of the abstract MDP. Given $x \in \mathcal{X}$, the function $proj$ is defined as follows:

$$proj(x)(k) \coloneqq \begin{cases} b^*(k) & \text{if } x = (b^*) \\ \sum_{\bar{k}\in\mathcal{K}} b^*(\bar{k}) t^i_{\bar{k},k}(a_1, ..., a_i) & \text{if } x = (b^*, a_1, ..., a_i) \end{cases}$$

where $T^i(a^i) = T^i(a_1, ..., a_i) = \prod_{j=1}^{i} P(a_j)$ for $i \in [1..n-1]$.

The *abstract reward function* is defined as:

$$\bar{g}^*(x, a) \coloneqq \sum_{k\in\mathcal{K}} proj(x)(k) \cdot g(k, a).$$

Given $x \in \mathcal{X}$, where x is of the form (b^*, a^i) for $i \in [0..n-1]$, and an action $a \in \mathcal{A}$, define the *abstract update* as

$$\tau^*(x,a) \coloneqq \begin{cases} (b^*, a_1, \cdots, a_i, a) & \text{if } i \in [0, n-2] \\ (proj(x)^\top \widetilde{T}^n(a^n)) & \text{if } i = n-1 \end{cases}$$

where $a^n = (a^{n-1}, a)$. In this context, for a given current state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$, the *abstract update* will compute the deterministic successor state $x' \in \mathcal{X}$. Define the *abstract transition function* as

$$\bar{p}^*(x'|x,a) \coloneqq \begin{cases} 1 & \text{if } x' = \tau^*(x,a) \\ 0 & \text{otherwise} \end{cases}$$

where $x, x' \in \mathcal{X}$ and $a \in \mathcal{A}$.

For each stage $m \geq 1$, let $\bar{G}^*_m \coloneqq \bar{g}^*(x_m, a_m)$ denote the stage reward where $x_m \in \mathcal{X}$ and $a_m \in \mathcal{A}$.

For a given strategy $\sigma \in \Sigma$, the long-run average objective of the abstract MDP is defined as follows

$$\gamma^*(\sigma) \coloneqq \liminf_{N \to \infty} \mathbb{E}^{b_1}_\sigma \left(\frac{1}{N} \sum_{m=1}^{N} \bar{G}^*_m \right),$$

and v^* is defined as

$$v^* \coloneqq \sup_\sigma \gamma^{b_1,*}_\infty(\sigma),$$

where the supremum is taken over all strategies $\sigma \in \Sigma$.

4.3 Decidability of Ergodic UMDPs

We now prove the decidability of approximating the value in the class of ergodic UMDPs in three steps.

Given a stochastic matrix P and a probability vector b, we consider the following norms: $\|P\|_1 \coloneqq \max_{k' \in \mathcal{K}} \sum_{k \in \mathcal{K}} |p_{k,k'}|$, $\|P\|_\infty \coloneqq \max_{k \in \mathcal{K}} \sum_{k' \in \mathcal{K}} |p_{k,k'}|$, and $\|b\|_1 \coloneqq \sum_{k \in \mathcal{K}} |b(k)|$.

Step 1. We first analyze the belief dynamics within an ergodic UMDP and its corresponding abstract MDP, proving that they remain closely aligned.

Lemma 4.5. *Let Γ be an ergodic UMDP and $\varepsilon \in (0,1)$. For all strategies $\sigma \in \Sigma$ and $m \geq 1$, the states of the abstract MDP $\mathcal{G}^*(\varepsilon)$ satisfies*

$$\left\| b^{b_1}_{m,\sigma} - proj(x^{b_1}_{m,\sigma}) \right\|_1 \leq 4\varepsilon,$$

where $x^{b_1}_{m,\sigma}$ denote the state in the abstract MDP at stage m induced by strategy σ and starting from initial belief b_1.

Proof of Lemma 4.5. Fix an ergodic UMDP Γ and $\varepsilon \in (0,1)$. Let n given by Proposition 4.4. Consider the abstract MDP $\mathcal{G}^*(\varepsilon)$. Recall that the MDP $\mathcal{G}^*(\varepsilon)$ is constructed as follows:

1. By Proposition 4.4, for $\varepsilon \in (0,1)$, there exists a number n with the associated set of matrices $\mathcal{T}(\varepsilon) = \{T^n(a^n)\}$ satisfying that, for all action sequences a^n,

$$\tau_1(T^n(a^n)) \leq \varepsilon\,,$$

   i.e., each matrix in $\mathcal{T}(\varepsilon)$ has similar rows;

2. Associated with $\mathcal{T}(\varepsilon)$, we construct the abstract set of stable matrices, denoted $\widetilde{\mathcal{T}}(\varepsilon)$;

3. Each matrix in $\widetilde{\mathcal{T}}(\varepsilon)$ can be regarded as a belief;

4. In the abstract MDP:

   - Using function $proj$, each state $x \in \mathcal{X}$ is related to a specific belief in $\Delta(\mathcal{K})$;
   - Actions correspond to actions in the original ergodic UMDP.

Fix a strategy $\sigma \in \Sigma$. We prove that

$$\left\| b^{b_1}_{m,\sigma} - proj(x^{b_1}_{m,\sigma}) \right\|_1 \leq 4\varepsilon, \qquad \forall m \geq 1.$$

We consider blocks of size n. Let $i \geq 0$, and a^n an action sequence. We recall the following relations:

- $b^{b_1}_{(i+1)n+1,\sigma} = b^{b_1\top}_{in+1,\sigma} T^n(a^n)$;
- $proj(x^{b_1}_{(i+1)n+1,\sigma}) = proj(x^{b_1}_{in+1,\sigma})^\top \widetilde{T}^n(a^n)$.

Define $\tilde{b}^{b_1}_{(i+1)n+1,\sigma} \coloneqq proj(x^{b_1}_{in+1,\sigma})^\top T^n(a^n)$.

We prove the claim by induction on $i \in \mathbb{N}$. We start by observing that the base case holds, i.e., when $i = 0$. Indeed, by construction of the abstract MDP, we have for all $m \in [0..n-1]$ that

$$\left\| b^{b_1}_{m+1,\sigma} - proj(x^{b_1}_{m+1,\sigma}) \right\|_1 = 0$$

and,

$$
\begin{aligned}
&\left\| b^{b_1}_{n+1,\sigma} - proj(x^{b_1}_{n+1,\sigma}) \right\|_1 \\
&\quad \leq \sum_{k=1}^{|\mathcal{K}|} \sum_{j=1}^{|\mathcal{K}|} b_1(j) \left| t^n_{j,k}(a^n) - \tilde{t}^n_{j,k}(a^n) \right| \\
&\quad = \sum_{j=1}^{|\mathcal{K}|} b_1(j) \sum_{k=1}^{|\mathcal{K}|} \left| t^n_{j,k}(a^n) - \tilde{t}^n_{j,k}(a^n) \right| \\
&\quad \leq \sum_{j=1}^{|\mathcal{K}|} b_1(j)\, 2\tau_1(T^n) && \text{(Def. } \tau_1 \text{ and } \widetilde{T}^n\text{)} \\
&\quad \leq 2\varepsilon. && \text{(Proposition 4.4)}
\end{aligned}
$$

Now, assume that the claim holds for the first i blocks. We prove it holds for the next block, i.e., block $i+1$. For each action sequence a^n,

$$
\begin{aligned}
&\left\| b^{b_1}_{(i+1)n+1,\sigma} - proj(x^{b_1}_{(i+1)n+1,\sigma}) \right\|_1 \\
&\quad = \left\| b^{b_1}_{(i+1)n+1,\sigma} - \tilde{b}^{b_1}_{(i+1)n+1,\sigma} + \tilde{b}^{b_1}_{(i+1)n+1,\sigma} - proj(x^{b_1}_{(i+1)n+1,\sigma}) \right\|_1 \\
&\quad \leq \left\| b^{b_1}_{(i+1)n+1,\sigma} - \tilde{b}^{b_1}_{(i+1)n+1,\sigma} \right\|_1 \\
&\qquad + \left\| \tilde{b}^{b_1}_{(i+1)n+1,\sigma} - proj(x^{b_1}_{(i+1)n+1,\sigma}) \right\|_1 \\
&\quad = \left\| b^{b_1\top}_{in+1,\sigma} T^n(a^n) - proj(x^{b_1}_{in+1,\sigma})^\top T^n(a^n) \right\|_1 \\
&\qquad + \left\| proj(x^{b_1}_{in+1,\sigma})^\top T^n(a^n) - proj(x^{b_1}_{in+1,\sigma})^\top \widetilde{T}^n(a^n) \right\|_1 \\
&\quad \leq 2\varepsilon + \left\| proj(x^{b_1}_{in+1,\sigma})^\top T^n(a^n) - proj(x^{b_1}_{in+1,\sigma})^\top \widetilde{T}^n(a^n) \right\|_1 && \text{(Proposition 4.4)} \\
&\quad \leq 2\varepsilon + \sum_{k=1}^{|\mathcal{K}|} \sum_{j=1}^{|\mathcal{K}|} proj(x^{b_1}_{in+1,\sigma})(j) \left| t^n_{j,k}(a^n) - \tilde{t}^n_{j,k}(a^n) \right| \\
&\quad = 2\varepsilon + \sum_{j=1}^{|\mathcal{K}|} proj(x^{b_1}_{in+1,\sigma})(j) \sum_{k=1}^{|\mathcal{K}|} \left| t^n_{j,k}(a^n) - \tilde{t}^n_{j,k}(a^n) \right| \\
&\quad \leq 2\varepsilon + \sum_{j=1}^{|\mathcal{K}|} proj(x^{b_1}_{in+1,\sigma})(j)\, 2\tau_1(T^n) && \text{(Def. } \tau_1 \text{ and } \widetilde{T}^n\text{)} \\
&\quad \leq 2\varepsilon + 2\varepsilon = 4\varepsilon. && \text{(Proposition 4.4)}
\end{aligned}
$$

We now compute the difference between each belief inside a block. For each $m \in [1..n-1]$, denote $Y^m \coloneqq \prod_{k=1}^{m} P(a_k)$, then

$$\begin{aligned}
&\left\| b^{b_1}_{in+1+m,\sigma} - proj(x^{b_1}_{in+1+m,\sigma}) \right\|_1 \\
&= \left\| (b^{b_1}_{in+1,\sigma} - proj(x^{b_1}_{in+1,\sigma}))^\top Y^m \right\|_1 \\
&\leq \sum_{k \in \mathcal{K}} \left| \left((b^{b_1}_{in+1,\sigma} - proj(x^{b_1}_{in+1,\sigma}))^\top Y^m \right)_k \right| \\
&\leq \left\| b^{b_1}_{in+1,\sigma} - proj(x^{b_1}_{in+1,\sigma}) \right\|_1 \|Y^m\|_\infty \\
&\leq 4\varepsilon.
\end{aligned}$$

Therefore, the result follows. □

Step 2. Knowing that beliefs remain close in the belief and abstract MDPs, we prove that the difference between the long-run average rewards are also close.

Lemma 4.6. *Let Γ be an ergodic UMDP and $\varepsilon \in (0,1)$. For all strategy $\sigma \in \Sigma$, the reward of the abstract MDP $\mathcal{G}^*(\varepsilon)$ satisfies*

$$\left| \liminf_{N\to\infty} \mathbb{E}^{b_1}_\sigma \left(\frac{1}{N} \sum_{m=1}^{N} \bar{G}_m \right) - \liminf_{N\to\infty} \mathbb{E}^{b_1}_\sigma \left(\frac{1}{N} \sum_{m=1}^{N} \bar{G}^*_m \right) \right| \leq 4\varepsilon.$$

Proof of Lemma 4.6. Let us consider $\tilde{n}$ blocks of size n, where n is given by Proposition 4.4. The difference in reward inside a block $i \in [0..\tilde{n}-1]$ is as follows

$$\begin{aligned}
\left| \mathbb{E}^{b_1}_\sigma \left(\frac{1}{n} \sum_{m=in+1}^{(i+1)n} [\bar{G}_m - \bar{G}^*_m] \right) \right| &= \left| \frac{1}{n} \sum_{m=in+1}^{(i+1)n} \bar{g}_m(b^{b_1}_{m,\sigma}, a_m) - \bar{g}^*_m(x^{b_1}_{m,\sigma}, a_m) \right| \\
&= \left| \frac{1}{n} \sum_{m=in+1}^{(i+1)n} \sum_{k \in \mathcal{K}} g(k, a_m)[b^{b_1}_{m,\sigma}(k) - proj(x^{b_1}_{m,\sigma})(k)] \right| \\
&\leq \frac{1}{n} \sum_{m=in+1}^{(i+1)n} \left\| b^{b_1}_{m,\sigma} - proj(x^{b_1}_{m,\sigma}) \right\|_1 \\
&\leq 4\varepsilon,
\end{aligned}$$

where the last inequality follows from Lemma 4.5.

By summing over $\tilde{n}$ blocks and considering that $N = \tilde{n}\, n$, we have that

$$\left| \mathbb{E}^{b_1}_\sigma \left(\frac{1}{N} \sum_{m=1}^{N} [\bar{G}_m - \bar{G}^*_m] \right) \right| = \left| \frac{1}{\tilde{n}} \sum_{j=0}^{\tilde{n}-1} \mathbb{E}^{b_1}_\sigma \left(\frac{1}{n} \sum_{m=jn+1}^{(j+1)n} [\bar{G}_m - \bar{G}^*_m] \right) \right| \leq 4\varepsilon.$$

Taking the limit as $\tilde{n}$ grows,

$$\begin{aligned}
\liminf_{N\to\infty} \mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}\bar{G}_m\right) &= \left|\liminf_{N\to\infty} \mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}[\bar{G}_m - \bar{G}_m^* + \bar{G}_m^*]\right)\right| \\
&\leq \liminf_{N\to\infty} \left|\mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}[\bar{G}_m - \bar{G}_m^* + \bar{G}_m^*]\right)\right| \\
&\leq \liminf_{N\to\infty} \left[\mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}|\bar{G}_m - \bar{G}_m^*|\right) + \mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}\bar{G}_m^*\right)\right] \\
&\leq \liminf_{N\to\infty} \left[\mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}\bar{G}_m^*\right)\right] + 4\varepsilon.
\end{aligned}$$

Similarly,

$$\liminf_{N\to\infty} \mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}\bar{G}_m^*\right) \leq \liminf_{N\to\infty} \left[\mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}\bar{G}_m\right)\right] + 4\varepsilon.$$

Therefore, we can conclude that

$$\left|\liminf_{N\to\infty} \mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}\bar{G}_m\right) - \liminf_{N\to\infty} \mathbb{E}_\sigma^{b_1}\left(\frac{1}{N}\sum_{m=1}^{N}\bar{G}_m^*\right)\right| \leq 4\varepsilon.$$

□

Step 3. Finally, we can prove Theorem 4.1.

Proof of Theorem 4.1. Given an ergodic UMDP and $\varepsilon \in (0,1)$, Proposition 4.4 guarantees the existence of a finite integer $n \leq \bar{n}(n_0,\varepsilon)n_0/2$, with $n_0 \leq (3^{|\mathcal{K}|} - 2^{|\mathcal{K}|+1} + 1)/2$, such that we can construct the finite set of matrices $\mathcal{T}(\varepsilon)$. Relying on this set, we derive its finite set of abstract matrices $\widetilde{\mathcal{T}}(\varepsilon)$. Finally, we can construct a finite-state MDP, termed abstract MDP, in a finite amount of time.

Moreover, Lemma 4.6 demonstrates that the abstract MDP provides an approximate value for the ergodic UMDPs. Given that the decision version of approximating the value is decidable for finite-state MDPs [28], the decidability for the class of ergodic UMDPs follows. □

Finally, Algorithm 1 summarizes the approximation scheme for the class of ergodic UMDPs.

Algorithm 1 Algorithm for Ergodic UMDPs

1: **Input:** UMDP $\Gamma = (\mathcal{K}, \mathcal{A}, p, g, b_1)$, $\varepsilon \in (0,1)$
2: Check ergodicity property of Γ
3: Construct the set of matrices $\mathcal{T}(\varepsilon)$
4: Derive the abstract set of matrices $\widetilde{\mathcal{T}}(\varepsilon)$ from $\mathcal{T}(\varepsilon)$
5: Construct the abstract MDP $\mathcal{G}^*(\varepsilon)$
6: Compute the optimal value v^* of the abstract MDP $\mathcal{G}^*(\varepsilon)$
7: **Output:** Optimal value v^*

4.4 Some Properties about ergodic UMDPs

In this section, we present some properties about ergodic UMDPs. Given a strategy $\sigma \in \Sigma$, define the discounted-sum objective for blind MDPs as follows: for all $\theta \in (0,1)$,

$$\gamma(\sigma, \theta) = \mathbb{E}_\sigma^{b_1}\left((1-\theta)\sum_{m\geq 1}\theta^{m-1}G_m\right),$$

and the discounted value as

$$v(\theta) = \sup_{\sigma\in\Sigma}\gamma(\sigma,\theta).$$

Similarly, for a given strategy $\sigma \in \Sigma$, the discounted-sum objective for the abstract MDP is defined as follows: for all $\theta \in (0,1)$

$$\gamma^*(\sigma, \theta) = \mathbb{E}_\sigma^{b_1}\left((1-\theta)\sum_{m\geq 1}\theta^{m-1}\bar{G}_m^*\right),$$

and the discounted value as

$$v^*(\theta) = \sup_{\sigma\in\Sigma}\gamma(\sigma,\theta).$$

Proposition 4.7. *Consider an ergodic UMDP Γ. The following properties hold:*

- *For all $\varepsilon \in \left(0, \dfrac{1}{2|\mathcal{K}|(|\mathcal{K}|-1)}\right]$, there exists $\theta_0 \in \left[1-\dfrac{1}{4|\mathcal{K}|^3}, 1\right)$ such that the discounted value of the abstract MDP $\mathcal{G}^*(\varepsilon)$ satisfies*

$$|v^*(\theta_0) - v| \leq 5\varepsilon.$$

- *The long-run average value is constant with respect to the initial belief.*

Proof of Proposition 4.7. We consider the first claim. Lemma 4.6 shows that, for every $\varepsilon \in \left(0, \dfrac{1}{2|\mathcal{K}|(|\mathcal{K}|-1)}\right]$, we have

$$|v^* - v| \leq 4\varepsilon$$

and, using Zwick and Paterson [38, Theorem 5.2, p. 352-353], there exists $\theta_0 \in [1 - 1/(4|\mathcal{K}|^3), 1)$ such that

$$|v^*(\theta_0) - v^*| \leq \varepsilon.$$

Therefore, we deduce the result as follows

$$\begin{aligned} |v^*(\theta_0) - v| &\leq |v^*(\theta_0) - v^*| + |v^* - v| \\ &\leq 5\varepsilon. \end{aligned}$$

Now, we tackle the second claim as follows. Consider an ergodic UMDP Γ and some initial belief $b_A \in \Delta(\mathcal{K})$ and $b_B \in \Delta(\mathcal{K})$. By property of ergodicity of stochastic matrices, for each action sequence $a^n = (a_1, ..., a_n)$ and $k, k' \in \mathcal{K}$, we have that

$$\sum_{k \in \mathcal{K}} \left(b_A(k) t^n_{k,k'}(a^n) - b_B(k) t^n_{k,k'}(a^n) \right) \longrightarrow 0,$$

as n goes to infinity. For all strategy $\sigma \in \Sigma$, similar computations as in the proof of Lemma 4.6 lead to

$$\left| \gamma^{b_A}_\infty(\sigma) - \gamma^{b_B}_\infty(\sigma) \right| \leq \varepsilon.$$

We conclude the proof by considering the limit as $\varepsilon \to 0$. □

5 Undecidability Analysis

In this section, we introduce an undecidability result for the decision version of determining the value for the class of ergodic UMDPs. This indicates that the approximation problem can be interpreted as "tight": computing the exact value is undecidable and therefore reducing the class of ergodic UMDPs to (perfectly observable) MDPs is not possible.

To approach the exact problem, we define probabilistic finite automata (PFA) [21]. While PFA are models that accepts or rejects strings, UMDPs are used for solving stochastic sequential optimization problems where the decision-maker receives no information of the system. These models are tightly connected. Indeed, the alphabet in PFA corresponds to actions in UMDPs. Further, the notion of acceptance in PFA corresponds to a reachability objective in UMDPs. Due to this connection, undecidability results in PFAs also holds for UMDPs. We define a state $k \in \mathcal{K}$ as absorbing if $p(k, a)(k) = 1$ for all $a \in \mathcal{A}$.

Definition 5.1 (PFA [9])**.** *A PFA, denoted as $\mathcal{M}$, is defined as a 5-tuple $\mathcal{M} = (\mathcal{K}, \mathcal{B}, \mathcal{A}, p, k_1)$, where:*

- *$\mathcal{K}$ is the finite set of states;*
- *$\mathcal{B} \subseteq \mathcal{K}$ is the set of nonabsorbing accepting states;*

- $\mathcal{A}$ *is the finite set of symbols;*
- $p\colon \mathcal{K}\times\mathcal{A}\to\Delta(\mathcal{K})$ *is the probability distribution over the successor state given current state* $k\in\mathcal{K}$ *and symbol* $a\in\mathcal{A}$*;*
- k_1 *is the initial state.*

Denote $\mathbb{P}_w^{k_1}(K_{N+1}\in\mathcal{B})$ the probability of acceptance of a word $w\in\mathcal{A}^*$ with $|w|=N$. Recall the following well-known undecidability result from [21, Theorem 3.2].

Theorem 5.2. *Given a PFA, deciding whether there exists a word with acceptance probability strictly greater than* $1/2$ *is undecidable.*

Theorem 5.2 implies the undecidability for ergodic UMDPs as proved by the following theorem.

Theorem 5.3. *The decision version of determining the value for the class of ergodic UMDPs is undecidable, and the undecidability holds even for the subclass of Markov UMDPs.*

Proof of Theorem 5.3. Consider a PFA $\mathcal{M}=(\mathcal{K},\mathcal{B},\mathcal{A},p,k_1)$ where k_1 is the starting state and $\mathcal{B}\subseteq\mathcal{K}$ the set of accepting states. Given $\mathcal{M}$ and $\lambda\in(0,1)$, we construct a UMDP $\Gamma=(\mathcal{K}',\mathcal{A}',p',g,\delta_{k_1})$ with the long-run average objective as follows:

- $\mathcal{K}'=\mathcal{K}\cup\{\hat{k}\}$ is the finite set of states;
- $\mathcal{A}'=\mathcal{A}\cup\{restart\}$ is the finite set of actions;
- $p'\colon\mathcal{K}'\times\mathcal{A}'\to\Delta(\mathcal{K}')$ is the probabilistic transition function, defined as follows:
  - For every action $a\in\mathcal{A}$ and $k\in\mathcal{K}$, $p'(k,a)(\hat{k})=\lambda$, $p'(k,a)(k')=(1-\lambda)p(k,a)(k')$ for all $k'\in\mathcal{K}$, $p'(\hat{k},a)(\hat{k})=1$;
  - Given action $a=restart$, $p'(k,restart)(k_1)=1$ for all $k\in\mathcal{K}'$.
- $g\colon\mathcal{K}'\times\mathcal{A}'\to[0,1]$ is the stage reward function, defined as follows:
  - Given some action $a\in\mathcal{A}$, we have $g(k',a)=1/2$ for all states $k'\in\mathcal{K}'$;
  - For action $restart$, we have $g(\hat{k},restart)=1/2$. Also, for all $k\in\mathcal{B}$, we have $g(k,restart)=+1$, and, for all $k\in\mathcal{K}\setminus\mathcal{B}$, we have $g(k,restart)=0$.

In the UMDP Γ, the set of accepting states $\mathcal{B}\subseteq\mathcal{K}'$ is transient. Note that, for all action $a\in\mathcal{A}'$, the transition matrix $P(a)$ is Markov. Therefore, the UMDP Γ is Markov which implies that Γ is ergodic.

Let us show that the acceptance probability of the PFA $\mathcal{M}$ is strictly greater than $1/2$ if and only if the value of the UMDP is strictly greater than $1/2$.

Consider a PFA $\mathcal{M}$ where there exists a word w of length $|w| = N$ that has an acceptance probability strictly greater than 1/2. We provide a strategy in the UMDP $\mathcal{G}$ that guarantees a payoff strictly greater than 1/2.

The strategy consists in repeatedly playing the actions in w followed by *restart*. To compute the payoff it guarantees, focus on a single block. The expected average reward is

$$\begin{aligned}
&\mathbb{E}^{\delta_{k_1}}\left(\frac{1}{N+1}\sum_{m=1}^{N+1} G_m\right)\\
&\quad= \frac{1}{N+1}\left[\frac{N}{2} + \mathbb{P}_w^{k_1}(K_{N+1}=\hat{k})\frac{1}{2} + \left(1-\mathbb{P}_w^{k_1}(K_{N+1}=\hat{k})\right)p_w^{k_1}(K_{N+1}\in\mathcal{B})\right]\\
&\quad> \frac{1}{N+1}\left[\frac{N}{2} + \mathbb{P}_w^{k_1}(K_{N+1}=\hat{k})\frac{1}{2} + \left(1-\mathbb{P}_w^{k_1}(K_{N+1}=\hat{k})\right)\frac{1}{2}\right]\\
&\quad= 1/2.
\end{aligned}$$

At stage $N+2$, the state is again k_1. By repeating the argument on each block of size $N+1$, we obtain that this strategy guarantees strictly more than 1/2.

Now consider that the UMDP $\mathcal{G}$ has a value strictly greater than 1/2. We show that the PFA $\mathcal{M}$ has a word with an acceptance probability strictly greater than 1/2. By [10], there exists an eventually periodic strategy that guarantees strictly more than 1/2. We claim that such a strategy should play *restart* an infinite number of times. Indeed, otherwise the game would remain in $\hat{k}$ from some stage, and the strategy would achieve payoff 1/2, which is a contradiction. Since the strategy is eventually cyclic and after playing once the action *restart* the state moves to k_1 with probability 1, we may consider that the strategy repeats a cycle of the form $(a_1, a_2, \ldots, a_n, restart)$. Moreover, if there exists $i \in [1..n]$ such that $a_i = restart$, the payoff the strategy guarantees is a weighted average between the payoff that the two strategies given by the cycles $(a_1, \ldots, a_{i-1}, restart)$ and $(a_{i+1}, \ldots, a_n, restart)$ guarantee. Therefore, repeating this argument, we may consider a strategy that repeats a cycle of the form $(a_1, a_2, \ldots, a_n, restart)$ where $a_i \in \mathcal{A}$ for all $i \in [1..n]$. We prove that the word $w = a_1 \ldots a_n$ is accepted by the PFA $\mathcal{M}$ with probability strictly larger than 1/2.

Indeed, denote the strategy σ and note that

$$\begin{aligned}
\frac{1}{2} &< \mathbb{E}_\sigma^{\delta_{k_1}}\left(\frac{1}{N+1}\sum_{m=1}^{N+1} G_m\right)\\
&= (1-\lambda)^n \mathbb{P}_w^{k_1}(K_{n+1}\in\mathcal{B}) + (1-(1-\lambda)^n)\frac{1}{2}.
\end{aligned}$$

Therefore, $\mathbb{P}_w^{k_1}(K_{n+1}\in\mathcal{B}) > 1/2$, i.e., w is accepted by $\mathcal{M}$ with probability strictly larger than 1/2. By Theorem 5.2, it follows that the decision version of determining the value for the class of Markov UMDPs is undecidable. Moreover, a result showing undecidability in a subclass indicates undecidability of the broader class. We conclude that determining the value for the class of ergodic UMDPs is undecidable. □

6 Extending Results to POMDPs

6.1 Class Description

In section 4, we established the decidability of approximating the value for ergodic unobservable MDPs. POMDPs extend UMDPs by allowing the decision-maker to receive a signal after taking an action. This generalization introduces one legitimate question whether a similar approximation scheme would be possible for a generalization of ergodic UMDPs to POMDPs. The definition of a POMDP is the following.

Definition 6.1 (POMDP)**.** *A POMDP, denoted as* Γ*, is defined by a* 7*-tuple* $\Gamma = (\mathcal{K}, \mathcal{A}, \mathcal{S}, p, q, g, b_1)$*, where:*

- $\mathcal{K}$ *is the finite set of states;*
- $\mathcal{A}$ *is the finite set of actions;*
- $\mathcal{S}$ *is the finite set of observations;*
- $p\colon \mathcal{K} \times \mathcal{A} \to \Delta(\mathcal{K})$*, represented as* $p(k'|k,a)$*, is the probabilistic transition function that gives the probability distribution over the successor states given a state* $k \in \mathcal{K}$ *and an action* $a \in \mathcal{A}$*. We represent by* $P(a)$ *the transition matrix for each action* $a \in \mathcal{A}$*;*
- $q\colon \mathcal{K} \times \mathcal{A} \to \Delta(\mathcal{S})$*, expressed as* $q(s|k,a)$*, is the probabilistic observation function that gives the probability distribution over the observations given a state* $k \in \mathcal{K}$ *and an action* $a \in \mathcal{A}$*. The observation matrix corresponding to some action* $a \in \mathcal{A}$ *is denoted as* $Q(a)$*;*
- $g\colon \mathcal{K} \times \mathcal{A} \to [0,1]$ *is the stage reward function;*
- $b_1 \in \Delta(\mathcal{K})$ *is the initial belief.*

For every action $a \in \mathcal{A}$ and signal $s \in \mathcal{S}$, define the matrix $R(a,s) \in M_{|\mathcal{K}|\times|\mathcal{K}|}$ such that $r_{k,k'}(a,s) \coloneqq q(s|k',a) \times p(k'|k,a)$ for all k, k' and let $\mathcal{R}$ denote the set of sub-stochastic matrices such that $\mathcal{R} \coloneqq \{R(a,s)|a \in \mathcal{A}, s \in \mathcal{S}\}$. We say that a sub-stochastic matrix R is scrambling if for every two rows there exists a common successor, i.e., a column with a positive entry in both rows. Then, we can generalize the class of scrambling UMDPs to POMDPs as follows.

Definition 6.2 (Scrambling POMDP)**.** *A POMDP is scrambling if every matrix in the set of matrices* $\mathcal{R}$ *is scrambling.*

The next Corollary establishes that determining the exact value for the class of scrambling POMDPs remains undecidable.

Corollary 6.3. *The decision version of determining the value for the class of scrambling POMDPs is undecidable.*

Proof of Corollary 6.3. The class of Markov UMDPs is a subclass of scrambling POMDPs. By Theorem 5.3, determining the value for the class of Markov UMDPs is undecidable. Since hardness results follow for larger classes, determining the value for the class of scrambling POMDPs is undecidable as well. □

6.2 Discussion

We discuss the main obstacles in extending the method developed for ergodic UMDPs to POMDPs. Recall that the main approach for both UMDPs and POMDPs involves considering their corresponding belief MDPs.

Stochastic Transitions. A significant difference exists in the transition properties of the belief MDPs for UMDPs and POMDPs. In UMDPs, the belief MDP features deterministic transitions. In contrast, the belief MDP in POMDPs exhibits stochastic transitions, which are expressed as a mapping from $\Delta(\mathcal{K}) \times \mathcal{A}$ to $\Delta(\Delta(\mathcal{K}))$. The stochastic nature of these transitions arises from signals which are drawn randomly from the signal function q. Consequently, these signals affect the transitions of the belief MDP in POMDPs, as illustrated in Figure 6.2.

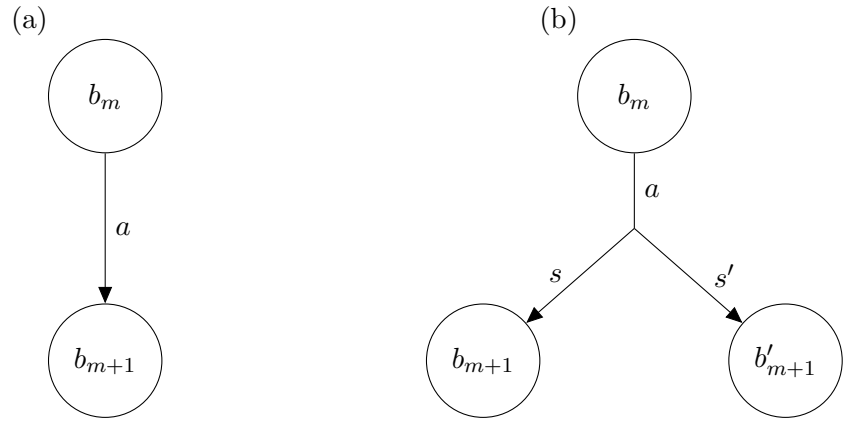

Figure 1: (a) Belief MDP of a UMDP which displays deterministic transitions; (b) Belief MDP of a POMDP which presents stochastic transitions from signals.

Perfect Coupling. The proof of decidability for the class ergodic UMDPs relies heavily on a "perfect coupling" between the ergodic UMDP and the abstract MDP given by the

deterministic transitions present in both models. Specifically, for all given strategy, the induced Markov chain can be expressed as a deterministic graph. As a result, this property facilitates a direct comparison between the two MDPs.

In the case of POMDPs, the stochastic nature of the transitions in the belief MDP could lead to diverging paths between the POMDP and the abstract MDP, even when the transitions are nearly identical. Such divergence potentially implies an error propagation between the two models. This problem has already been discussed in the literature. For instance, Rosenberg et al. [31] encountered a similar problem when trying to extend their result on ε-optimal stationary strategies for UMDPs to POMDPs.

Consequently, the lack of perfect or even strong coupling in POMDPs severely restricts the direct comparison between the two models. Overcoming these fundamental differences requires new innovative approaches.

Undecidability for POMDPs. Finally, one could consider exploring the undecidability of the approximation problem in scrambling POMDPs. Undecidability results in the literature typically stem from the unobservable case. Indeed, this method is commonly adopted as proving undecidability in this scenario guarantees that the result will hold for the more general class. Therefore, proving undecidability for a general class of POMDPs, such as scrambling POMDPs, could be very challenging to obtain since the decidability holds for the unobservable MDPs counterpart.

Acknowledgements

This material is based upon work supported by the ANRT under the French CIFRE Ph.D program, in collaboration between NyxAir (France) and Paris-Dauphine University (Contract: CIFRE N° 2022/0513), by the French Agence Nationale de la Recherche (ANR) under reference ANR-21-CE40-0020 (CONVERGENCE project), and partially supported by the ERC CoG 863818 (ForM-SMArt) grant. Part of this work was done at NyxAir (France) by David Lurie. Part of this work was done during a 1-year visit of Bruno Ziliotto to the Center for Mathematical Modeling (CMM) at University of Chile in 2023, under the IRL program of CNRS.

References

[1] Jeffrey M Alden and Robert L Smith. Rolling horizon procedures in nonhomogeneous markov decision processes. Operations Research, 40(3-supplement-2):S183–S194, 1992.

[2] Aristotle Arapostathis, Vivek S Borkar, Emmanuel Fernández-Gaucherand, Mrinal K Ghosh, and Steven I Marcus. Discrete-time controlled markov processes with average cost criterion: a survey. SIAM Journal on Control and Optimization, 31(2):282–344, 1993.

[3] Karl Johan Åström. Optimal control of markov processes with incomplete state information. Journal of mathematical analysis and applications, 10(1):174–205, 1965.

[4] James C Bean, Robert L Smith, and Jean B Lasserre. Denumerable state nonhomogeneous markov decision processes. Journal of Mathematical Analysis and Applications, 153(1):64–77, 1990.

[5] Richard Bellman. A markovian decision process. Journal of mathematics and mechanics, pages 679–684, 1957.

[6] DP Bertsekas and SE Shreve. Optimal control: The discrete time case, 1978.

[7] RG Bukharaev. Probabilistic automata. Journal of Soviet Mathematics, 13:359–386, 1980.

[8] Anthony R Cassandra. A survey of pomdp applications. In Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes, volume 1724, 1998.

[9] Krishnendu Chatterjee and Thomas A Henzinger. Probabilistic automata on infinite words: Decidability and undecidability results. In International Symposium on Automated Technology for Verification and Analysis, pages 1–16. Springer, 2010.

[10] Krishnendu Chatterjee, Raimundo Saona, and Bruno Ziliotto. Finite-memory strategies in pomdps with long-run average objectives. Mathematics of Operations Research, 2021.

[11] Pierre-Yves Chevalier, Vladimir V Gusev, Raphaël M Jungers, and Julien M Hendrickx. Sets of stochastic matrices with converging products: bounds and complexity. arXiv preprint arXiv:1712.02614, 2017.

[12] Ingrid Daubechies and Jeffrey C Lagarias. Sets of matrices all infinite products of which converge. Linear algebra and its applications, 161:227–263, 1992.

[13] Eugene Borisovich Dynkin and Alexander Adolph Yushkevich. Controlled markov processes, volume 235. Springer, 1979.

[14] Eugene A Feinberg. On measurability and representation of strategic measures in markov decision processes. Lecture Notes-Monograph Series, pages 29–43, 1996.

[15] Emmanuel Fernández-Gaucherand, Aristotle Arapostathis, and Steven I Marcus. On partially observable markov decision processes with an average cost criterion. In Proceedings of the 28th IEEE Conference on Decision and Control,, pages 1267–1272. IEEE, 1989.

[16] Robert Givan, Sonia Leach, and Thomas Dean. Bounded-parameter markov decision processes. Artificial Intelligence, 122(1-2):71–109, 2000.

[17] John Hajnal and Maurice S Bartlett. Weak ergodicity in non-homogeneous markov chains. In Mathematical Proceedings of the Cambridge Philosophical Society, volume 54, pages 233–246. Cambridge University Press, 1958.

[18] Wallace J Hopp, James C Bean, and Robert L Smith. A new optimality criterion for nonhomogeneous markov decision processes. Operations Research, 35(6):875–883, 1987.

[19] Marius Iosifescu. On two recent papers on ergodicity in nonhomogeneous markov chains. The Annals of Mathematical Statistics, 43(5):1732–1736, 1972.

[20] Andrei Kolmogoroff. Über die analytischen methoden in der wahrscheinlichkeitsrechnung. Mathematische Annalen, 104:415–458, 1931.

[21] Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. Artificial Intelligence, 147(1-2):5–34, 2003.

[22] JL Mott. Xxiv.—conditions for the ergodicity of non-homogeneous finite markov chains. Proceedings of the Royal Society of Edinburgh Section A: Mathematics, 64(4):369–380, 1957.

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

[24] YS Park, James C Bean, and Robert L Smith. Optimal average value convergence in nonhomogeneous markov decision processes. Journal of mathematical analysis and applications, 179(2):525–536, 1993.

[25] A Paz. Introduction to Probabilistic Automata. Computer Science and Applied Mathematics. Academic Press, Cambridge, MA, 1971.

[26] Azaria Paz. Definite and quasidefinite sets of stochastic matrices. Proceedings of the American Mathematical Society, 16(4):634–641, 1965.

[27] Azaria Paz. Introduction to probabilistic automata. Academic Press, 2014.

[28] ML Puterman et al. Discrete stochastic dynamic programming. Markov Decision Processes, 1994.

[29] Michael O Rabin. Probabilistic automata. Information and control, 6(3):230–245, 1963.

[30] Detlef Rhenius. Incomplete information in markovian decision models. The Annals of Statistics, pages 1327–1334, 1974.

[31] Dinah Rosenberg, Eilon Solan, and Nicolas Vieille. Blackwell optimality in markov decision processes with partial observation. Annals of statistics, pages 1178–1193, 2002.

[32] Eugene Seneta. Coefficients of ergodicity: structure and applications. Advances in applied probability, 11(3):576–590, 1979.

[33] Eugene Seneta. Non-negative matrices and Markov chains. Springer Science & Business Media, 2006.

[34] Xavier Venel and Bruno Ziliotto. Strong uniform value in gambling houses and partially observable markov decision processes. SIAM Journal on Control and Optimization, 54(4):1983–2008, 2016.

[35] Allise O Wachs, Irwin E Schochetman, and Robert L Smith. Average optimality in nonhomogeneous infinite horizon markov decision processes. Mathematics of Operations Research, 36(1):147–164, 2011.

[36] Jacob Wolfowitz. Products of indecomposable, aperiodic, stochastic matrices. Proceedings of the American Mathematical Society, 14(5):733–737, 1963.

[37] AA Yushkevich. Reduction of a controlled markov model with incomplete data to a problem with complete information in the case of borel state and control space. Theory of Probability & Its Applications, 21(1):153–158, 1976.

[38] Uri Zwick and Mike Paterson. The complexity of mean payoff games on graphs. Theoretical Computer Science, 158(1-2):343–359, 1996.