# The steady-state control problem for Markov decision processes

Sundararaman Akshay, Nathalie Bertrand, Serge Haddad, Loïc Hélouët

# The steady-state control problem for Markov decision processes

S. Akshay[1,2], Nathalie Bertrand[1], Serge Haddad[3], and Loïc Hélouët[1]

[1] Inria Rennes, France
[2] IIT Bombay, India
[3] LSV, ENS Cachan & CNRS & INRIA, France

**Abstract.** This paper addresses a control problem for probabilistic models in the setting of Markov decision processes (MDP). We are interested in the *steady-state control problem* which asks, given an ergodic MDP $\mathcal{M}$ and a distribution $\delta_{\mathsf{goal}}$, whether there exists a (history-dependent randomized) policy $\boldsymbol{\pi}$ ensuring that the steady-state distribution of $\mathcal{M}$ under $\boldsymbol{\pi}$ is exactly $\delta_{\mathsf{goal}}$. We first show that stationary randomized policies suffice to achieve a given steady-state distribution. Then we infer that the steady-state control problem is decidable for MDP, and can be represented as a linear program which is solvable in PTIME. This decidability result extends to labeled MDP (LMDP) where the objective is a steady-state distribution on labels carried by the states, and we provide a PSPACE algorithm. We also show that a related *steady-state language inclusion problem* is decidable in EXPTIME for LMDP. Finally, we prove that if we consider MDP under partial observation (POMDP), the steady-state control problem becomes undecidable.

## 1 Introduction

Probabilistic systems are frequently modeled as Markov chains, which are composed of a set of states and a probabilistic transition relation specifying the probability of moving from one state to another. When the system interacts with the environment, as is very often the case in real-life applications, in addition to the probabilistic moves, non-deterministic choices are possible. Such choices are captured by Markov Decision Processes (MDP), which extend Markov chains with non-determinism. Finally, in several applications, the system is not fully observable, and the information about the state of a system at a given instant is not precisely known. The presence of such uncertainty in observation can be captured by Partially Observable Markov Decision Processes (POMDP).

In all these settings, given a probabilistic system one is often interested in knowing whether, in the long run, it satisfies some property. For instance, one may want to make sure that the system does not, on an average, spend too much time in a faulty state. In the presence of non-deterministic choices (as in an MDP) or partial observation (as in a POMDP), a crucial question is whether we can always "control" these choices so that a long run property can be achieved.

In this paper, we are interested in control problems for Markov decision processes (MDP) and partially observable Markov decision processes (POMDP) with respect to long-run objectives. Given a Markov chain, it is well known [5,7] that one can compute its set of steady-state distributions, depending on the initial distribution. In an open setting, *i.e.*, when considering MDP, computing steady-state distributions becomes more challenging. Controlling an MDP amounts to defining a policy, that is, a function that associates, with every history of the system, a distribution on non-deterministic choices.

We tackle the *steady-state control problem*: given an MDP with a fixed initial distribution, and a goal distribution over its state space, does there exist a policy realizing the goal distribution as its steady-state distribution? (1) We prove decidability of the steady-state control problem for the class of so-called ergodic MDP, and provide a PTIME algorithm using linear programming techniques. (2) We next lift the problem to the setting of LMDP, where we add labels to states and check if a goal distribution over these labels can be reached by the system under some policy. For LMDP we show decidability of the steady-state control problem and provide a PSPACE algorithm. (3) Finally, for POMDP, we establish that the steady-state control problem becomes undecidable.

We also consider the *steady-state language inclusion problem* for LMDP. Namely, given two LMDP the question is whether any steady-state distribution over labels realizable in one process can be realized in the other. Building on our techniques for the steady-state control problem, we show that the language inclusion problem for LMDP is decidable in EXPTIME.

As already mentionned, steady-state control can be useful to achieve a given error rate, and in general to enforce quantitative fairness in a system. Steady-state language inclusion is a way to guarantee that a refinement of a system does not affect its long term behaviors. The problem of controlling a system such that it reaches a steady-state has been vastly studied in control theory for continuous models, *e.g.* governed by differential equations and where reachability should occur in finite time. There is a large body of work which addresses control problems for Markov decision processes. However, the control objectives are usually defined in terms of an optimization of a cost function (see *e.g.* [8, 10]). On the contrary, in this work, the control objective is to achieve a given steady-state distribution. In a recent line of work [3, 6], the authors consider transient properties of MDP viewed as transformers of probability distributions. Compared to that setting, we are interested rather in long run properties. Finally, in [4], the authors consider the problem of language equivalence for labeled Markov chains (LMC) and LMDP. For LMC, this problem consists of checking if two given LMC have the same probability distribution on finite executions (over the set of labels) and is shown to be decidable in PTIME. The equivalence

problem for LMDP is left open. As we are only interested in long run behaviors, we tackle a steady-state variant of this problem.

The paper is organized as follows. Section 2 introduces notations and definitions. Section 3 formalizes and studies the steady-state control problem: MDP are considered in Subsection 3.1; Subsection 3.2 extends the decidability results to LMDP and also deals with the steady-state language inclusion problem; and Subsection 3.3 establishes that partial observation entails undecidability of the steady-state control problem. We conclude with future directions in Section 4.

## 2 Preliminaries

In what follows, we introduce notations for matrices and vectors, assuming the matrix/vector size is understood from the context. We denote the identity matrix by $\mathbf{Id}$, the (row) vector with all entries equal to 1 by $\mathbb{1}$ and the (row) vector with only 0's by $\mathbf{0}$. The transpose of a matrix $\mathbf{M}$ (possibly a vector) is written $\mathbf{M}^t$. Given a square matrix $\mathbf{M}$, $\det(\mathbf{M})$ is its determinant.

### 2.1 Markov chains

We recall some definitions and results about Markov chains. Given a countable set $T$, we let $\mathsf{Dist}(T)$ denote the set of distributions over $T$, that is, the set of functions $\delta : T \to [0,1]$ such that $\sum_{t\in T} \delta(t) = 1$.

**Definition 1.** *A discrete time Markov chain (*DTMC*) is a tuple* $\mathcal{A} = (S, \Delta, \boldsymbol{s}_0)$ *where:*

- $S$ *is the finite or countable set of states.*
- $\Delta : S \to \mathsf{Dist}(S)$ *is the transition function describing the distribution over states reached in one step from a state.*
- $\boldsymbol{s}_0 \in \mathsf{Dist}(S)$ *is the initial distribution.*

As usual the transition matrix $\mathbf{P}$ of the Markov chain $\mathcal{A}$ is the $|S| \times |S|$ row-stochastic matrix defined by $\mathbf{P}[s, s'] \stackrel{\text{def}}{=} \Delta(s)(s')$, i.e., the $(s, s')^{th}$ entry of the matrix $\mathbf{P}$ gives the value defined by $\Delta$ of the probability to reach $s'$ from $s$ in one step. When the DTMC $\mathcal{A}$ is finite, one defines an directed graph $G_{\mathcal{A}}$ whose vertices are states of $\mathcal{A}$ and such that there is an arc from $s$ to $s'$ if $\mathbf{P}[s, s'] > 0$. $\mathcal{A}$ is said to be *recurrent* if $G_{\mathcal{A}}$ is strongly connected. The periodicity of a graph $p$ is the greatest integer such that there exists a partition of $S = \biguplus_{i=0}^{p-1} S_i$ such that for all $s \in S_i$ and $s' \in S$, there is an arc from $s$ to $s'$ only if $s' \in S_{(i+1 \mod p)}$. When the periodicity of $G_{\mathcal{A}}$ is 1, $\mathcal{A}$ is said to be *aperiodic*. Finally $\mathcal{A}$ is said to be *ergodic* if it is recurrent and aperiodic.

Now, consider the sequence of distributions $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots$ such that $\boldsymbol{s}_i = \boldsymbol{s}_0 \cdot \mathbf{P}^i$. This sequence does not necessarily converge (if the Markov chain is periodic)[4]. We write $\mathsf{sd}(\mathcal{A})$ when the limit exists and call it the *steady-state distribution* of $\mathcal{A}$. In case of an ergodic DTMC $\mathcal{A}$, (1) $\mathsf{sd}(\mathcal{A})$ exists, (2) it does not depend on $\boldsymbol{s}_0$ and, (3) it is the unique distribution $\boldsymbol{s}$ which fulfills $\boldsymbol{s} \cdot \mathbf{P} = \boldsymbol{s}$. When $\mathcal{A}$ is only recurrent, there is still a single distribution, called the *invariant distribution*, that fulfills this equation, and it coincides with the Cesàro limit. However it is a steady-state distribution only for a subset of initial distributions.

*Labeled Markov chains.* Let $L = \{l_1, l_2, \ldots\}$ be a finite set of labels. A *labeled Markov chain* is a tuple $(\mathcal{A}, \ell)$ where $\mathcal{A} = (S, \Delta, \boldsymbol{s}_0)$ is a Markov chain and $\ell : S \to L$ is a function assigning a label to each state. Given $(\mathcal{A}, \ell)$ a labeled Markov chain, the *labeled steady-state distribution*, denoted by $\mathsf{lsd}(\mathcal{A}, \ell)$ or simply $\mathsf{lsd}(\mathcal{A})$ when $\ell$ is clear from the context, is defined when $\mathsf{sd}(\mathcal{A})$ exists and is its projection onto the labels in $L$, via $\ell$. More formally, for every $l \in L$,

$$\mathsf{lsd}(\mathcal{A})(l) = \sum_{s \in S \mid \ell(s) = l} \mathsf{sd}(\mathcal{A})(s)$$

### 2.2 Markov decision processes

**Definition 2.** *A* Markov decision process *(*MDP*)* $\mathcal{M} = (S, \{A_s\}_{s \in S}, p, \boldsymbol{s}_0)$ *is defined by:*

- $S$*, the finite set of states;*
- *For every state* $s$*,* $A_s$*, the finite set of actions enabled in* $s$*.*
- $p : \{(s, a) \mid s \in S, a \in A_s\} \to \mathsf{Dist}(S)$ *is the transition function. The conditional probability transition* $p(s'|s, a)$ *denotes the probability to go from* $s$ *to* $s'$ *if* $a$ *is selected.*
- $\boldsymbol{s}_0 \in \mathsf{Dist}(S)$ *is the initial distribution.*

To define the semantics of an MDP $\mathcal{M}$, we first define the notion of *history*: a possible finite or infinite execution of the MDP.

**Definition 3.** *Given an* MDP $\mathcal{M}$*, a* history *is a finite or infinite sequence alternating states and actions* $\sigma = (s_0, a_0, \ldots, s_i, a_i, \ldots)$*. The number of actions of* $\sigma$ *is denoted* $\lg(\sigma)$*, and if* $\sigma$ *is finite, we write* $\mathsf{last}(\sigma)$ *for this last state. One requires that for all* $0 \leq i < \lg(\sigma)$*,* $p(s_{i+1}|s_i, a_i) > 0$*.*

[4] But it always admits a *Cesàro-limit*: the sequence $c_n = \frac{1}{n}(\boldsymbol{s}_0 + \cdots + \boldsymbol{s}_{n-1})$ converges (see *e.g.* [8, p.590]).

Compared to Markov chains, MDP contain non-deterministic choices. From a state $s$, when an action $a \in A_s$ is chosen, the probability to reach state $s'$ is $p(s'|s,a)$. In order to obtain a stochastic process, we need to fix the non-deterministic features of the MDP. This is done via (1) decision rules that select at some time instant the next action depending on the history of the execution, and (2) policies which specify which decision rules should be used at any time instant. Different classes of decision rules and policies are defined depending on two criteria: (1) the information used in the history and (2) the way the selection is performed (deterministically or randomly).

**Definition 4.** *Given an* MDP $\mathcal{M}$ *and* $t \in \mathbb{N}$*, a* decision rule $d_t$ *associates with every history* $\sigma$ *of length* $t = \lg(\sigma) < \infty$ *ending at a state* $s_t$*, a distribution* $d_t(\sigma)$ *over* $A_{s_t}$.

- *The set of all decision rules (also called* history-dependent randomized *decision rules) at time* $t$ *is denoted* $D_t^{HR}$.
- *The subset of* history-dependent deterministic *decision rules at time* $t$*, denoted* $D_t^{HD}$*, consists of associating a single action (instead of a distribution) with each history* $\sigma$ *of length* $t < \infty$ *ending at a state* $s_t$*. Thus, in this case* $d_t(\sigma) \in A_{s_t}$.
- *The subset of* Markovian randomized *decision rules at time* $t$*, denoted* $D_t^{MR}$ *only depends on the final state of the history. So one denotes* $d_t(s)$ *the distribution that depends on* $s$.
- *The subset of* Markovian deterministic *decision rules at time* $t$*,* $D_t^{MD}$ *only depends on the final state of the history and selects a single action. So one denotes* $d_t(s)$ *this action belonging to* $A_s$.

When the time $t$ is clear from context, we will omit the subscript and just write $D^{HR}, D^{HD}, D^{MD}$ and $D^{MR}$.

**Definition 5.** *Given an* MDP $\mathcal{M}$*, a* policy *(also called a* strategy*)* $\boldsymbol{\pi}$ *is a finite or infinite sequence of decision rules* $\boldsymbol{\pi} = (d_0, \ldots, d_t, \ldots)$ *such that* $d_t$ *is a decision rule at time* $t$*, for every* $t \in \mathbb{N}$.

*The set of policies such that for all* $t$*,* $d_t \in D_t^K$ *is denoted* $\Pi^K$ *for each* $K \in \{HR, HD, MR, MD\}$.

*When decisions* $d_t$ *are Markovian and all equal to some rule* $d$*,* $\boldsymbol{\pi}$ *is said* stationary *and denoted* $d^\infty$*. The set of stationary randomized (resp. deterministic) policies is denoted* $\Pi^{SR}$ *(resp.* $\Pi^{SD}$*).*

A Markovian policy only depends on the current state and the current time while a stationary policy only depends on the current state. Now, once a policy $\boldsymbol{\pi}$ is chosen, for each $n$, we can compute the probability distribution over the

histories of length $n$ of the MDP. That is, under the policy $\boldsymbol{\pi} = d_0, d_1, \ldots d_n, \ldots$ and with initial distribution $\boldsymbol{s}_0$, then, for any $n \in \mathbb{N}$, the probability of the history $\sigma_n = s_0 a_0 \ldots s_{n-1} a_{n-1} s_n$, is defined inductively by:

$$p^{\boldsymbol{\pi}}(\sigma_n) = d_n(\sigma_{n-1})(a_{n-1}) \cdot p(s_n | s_{n-1}, a_{n-1}) \cdot p^{\boldsymbol{\pi}}(\sigma_{n-1}) \ ,$$

and $p^{\boldsymbol{\pi}}(\sigma_0) = \boldsymbol{s}_0(s_0)$. Then, by summing over all histories of length $n$ ending in the same state $s$, we obtain the probability of reaching state $s$ after $n$ steps. Formally, letting $X_n$ denote the random variable corresponding to the state at time $n$, we have:

$$\mathbb{P}^{\boldsymbol{\pi}}(X_n = s) = \sum_{\sigma | \mathrm{lg}(\sigma) = n \wedge \mathsf{last}(\sigma) = s} p^{\boldsymbol{\pi}}(\sigma)$$

Observe that once a policy $\boldsymbol{\pi}$ is chosen, an $\mathsf{MDP}$ $\mathcal{M}$ can be seen as a discrete-time Markov chain ($\mathsf{DTMC}$), written $\mathcal{M}_{\boldsymbol{\pi}}$, whose states are histories. The Markov chain $\mathcal{M}_{\boldsymbol{\pi}}$ has infinitely many states in general. When a stationary policy $d^\infty$ is chosen, one can forget the history of states except for the last one, and thus consider the states of the $\mathsf{DTMC}$ $\mathcal{M}_{\boldsymbol{\pi}}$ to be those of the $\mathsf{MDP}$ $\mathcal{M}$ and the transition matrix $\mathbf{P}_d$ is defined by:

$$\mathbf{P}_d[s, s'] \stackrel{\text{def}}{=} \sum_{a \in A_s} d(s)(a) p(s' | s, a).$$

Thus, in this case the probability of being in state $s$ at time $n$ is just given by $\mathbb{P}(X_n = s) = (\boldsymbol{s}_0 \cdot \mathbf{P}_d^n)(s)$.

*Recurrence and ergodicity.* A Markov decision process $\mathcal{M}$ is called *recurrent* (resp. *ergodic*) if for every $\boldsymbol{\pi} \in \mathit{\Pi}^{SD}$, $\mathcal{M}_{\boldsymbol{\pi}}$ is recurrent (resp. ergodic). Recurrence and ergodicity of an $\mathsf{MDP}$ can be effectively checked, as the set of graphs $\{G_{\mathcal{M}_{\boldsymbol{\pi}}} \mid \boldsymbol{\pi} \in \mathit{\Pi}^{SD}\}$ is finite. Observe that when $\mathcal{M}$ is called recurrent (resp. ergodic) then for every $\boldsymbol{\pi} \in \mathit{\Pi}^{SR}$, $\mathcal{M}_{\boldsymbol{\pi}}$ is recurrent (resp. ergodic).

*Steady-state distributions.* We fix a policy $\boldsymbol{\pi}$ of an $\mathsf{MDP}\mathcal{M}$. Then, for any $n \in \mathbb{N}$, we define the distribution reached by $\mathcal{M}_{\boldsymbol{\pi}}$ at the $n$-th stage, i.e., for any state $s \in S$ as: $\delta_n^{\boldsymbol{\pi}}(s) = \mathbb{P}^{\boldsymbol{\pi}}(X_n = s)$. Now when it exists, the steady-state distribution $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}})$ of the $\mathsf{MDP}$ $\mathcal{M}$ under policy $\boldsymbol{\pi}$ is defined as: $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}})(s) = \lim_{n \to \infty} \delta_n^{\boldsymbol{\pi}}(s)$. Observe that when $\mathcal{M}$ is ergodic, for every decision rule $d$, $\mathcal{M}_{d^\infty}$ is ergodic and so $\mathsf{sd}(\mathcal{M}_{d^\infty})$ is defined.

Now, as we did for Markov chains, given a set of labels $L$, a labeled $\mathsf{MDP}$, is a tuple $(\mathcal{M}, \ell)$ where $\mathcal{M}$ is an $\mathsf{MDP}$ and $\ell : S \to L$ is a labeling function. Then, for $\mathcal{M}$ an $\mathsf{MDP}$, $\ell$ a labeling function, and $\boldsymbol{\pi}$ a strategy, we define $\mathsf{lsd}(\mathcal{M}_{\boldsymbol{\pi}}, \ell)$ or simply $\mathsf{lsd}(\mathcal{M}_{\boldsymbol{\pi}})$ for the projection of $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}})$ (when it exists) onto the labels in $L$ via $\ell$.

## 3 The steady-state control problem

### 3.1 Markov decision processes

Given a Markov decision process, the steady-state control problem asks whether one can come up with a policy to realize a given steady-state distribution. In this paper, we only consider ergodic MDP. Formally,

**Steady-state control problem for** MDP
Input: An ergodic MDP $\mathcal{M} = (S, \{A_s\}_{s\in S}, p, \boldsymbol{s}_0)$, and a distribution $\delta_{\mathsf{goal}} \in \mathsf{Dist}(S)$.
Question: Does there exist a policy $\boldsymbol{\pi} \in \Pi^{HR}$ s.t. $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}})$ exists and is equal to $\delta_{\mathsf{goal}}$?

The main contribution of this paper is to prove that, the above decision problem is decidable and belongs to PTIME for ergodic MDP. Furthermore it is effective: if the answer is positive, one can compute a witness policy. To establish this result we show that if there exists a witness policy, then there is a simple one, namely a stationary randomized policy $\boldsymbol{\pi} \in \Pi^{SR}$. We then solve this simpler question by reformulating it as an equivalent linear programming problem, of size polynomial in the original MDP. More formally,

**Theorem 1.** *Let $\mathcal{M}$ be an ergodic* MDP. *Assume there exists $\boldsymbol{\pi} \in \Pi^{HR}$ such that $\lim_{n\to\infty} \delta_n^{\pi} = \delta_{\mathsf{goal}}$. Then there exists $d^{\infty} \in \Pi^{SR}$ such that $\lim_{n\to\infty} \delta_n^{d^{\infty}} = \delta_{\mathsf{goal}}$.*

The following folk theorem states that Markovian policies (that is, policies based only on the history length and the current state) are as powerful as general history-dependent policies to achieve marginal distributions for $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ where $\{Y_n\}_{n\in\mathbb{N}}$ to denote the family of random variables corresponding to the chosen actions at time $n$. Observe that this is no more the case when considering joint distributions.

**Theorem 2 ( [8], Thm. 5.5.1).** *Let $\boldsymbol{\pi} \in \Pi^{HR}$ be a policy of an* MDP $\mathcal{M}$. *Then there exists a policy $\boldsymbol{\pi}' \in \Pi^{MR}$ such that for all $n \in \mathbb{N}$, $s \in S$ and $a \in A_s$:*

$$\mathbb{P}^{\boldsymbol{\pi}'}(X_n = s, Y_n = a) = \mathbb{P}^{\boldsymbol{\pi}}(X_n = s, Y_n = a)$$

Hence for an history-dependent randomized policy, there exists a Markovian randomized one with the same transient distributions and so with the same steady-state transient distribution if the former exists. It thus suffices to prove Theorem 1 assuming that $\boldsymbol{\pi} \in \Pi^{MR}$. To this aim, we establish several intermediate results.

Let $d \in D^{MR}$ be a Markovian randomized decision rule. $d$ can be expressed as a convex combination of the finitely many Markovian deterministic decision rules: $d = \sum_{e \in D^{MD}} \lambda_e e$. We say that a sequence $d_n \in D^{MR}$ admits a limit $d$, denoted $d_n \to_{n\to\infty} d$, if, writing $d_n = \sum_{e \in D^{MD}} \lambda_{e,n} e$ and $d = \sum_{e \in D^{MD}} \lambda_e e$, then for all $e \in D^{MD}$, $\lim_{n\to\infty} \lambda_{e,n} = \lambda_e$.

**Lemma 1.** *Let $\mathcal{M}$ be an ergodic* MDP *and $(d_n)_{n\in\mathbb{N}} \in \Pi^{MR}$. If the sequence $d_n$ has a limit $d$, then $\lim_{n\to\infty} \mathsf{sd}(\mathcal{M}_{d_n^\infty})$ exists and is equal to $\mathsf{sd}(\mathcal{M}_{d^\infty})$.*

In words, Lemma 1 states that the steady-state distribution under the limit policy $d$ coincides with the limit of the steady-state distributions under the $d_n$'s. The steady-state distribution operator is thus continuous over Markovian randomized decision rules.

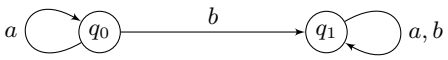
*Proof (of Lemma 1).* Consider the following equation system with parameters $\{\lambda_e\}_{e\in D^{MD}}$, and a vector of variables $\boldsymbol{X}$, obtained from

$$\boldsymbol{X} \cdot (\mathbf{Id} - \sum_{e \in D^{MD}} \lambda_e \mathbf{P}_e) = 0$$

by removing one equation (any of them), and then adding $\boldsymbol{X} \cdot \mathbb{1}^t = 1$. This system can be rewritten in the form $\boldsymbol{X} \cdot \mathbf{M} = \mathbf{b}$. Using standard results of linear algebra, the determinant $\det(\mathbf{M})$ is a rational fraction in the $\lambda_e$'s. Moreover due to ergodicity of $\mathcal{M}$, $\mathbf{M}$ is invertible for any tuple $(\lambda_e)_{e\in D^{MD}}$ with $\sum_{e \in D^{MD}} \lambda_e = 1$. Thus the denominator of this fraction does not cancel for such values. As a result, the function $f : (\lambda_e)_{e\in D^{MD}} \mapsto \mathsf{sd}(\mathcal{M}_{(\sum \lambda_e e)^\infty})$, which is a vector of rational functions, is continuous which concludes the proof. □

Note that Lemma 1 does not hold if we relax the assumption that $\mathcal{M}$ is ergodic. Indeed, consider an MDP with two states $s_0$, $s_1$ and two actions $a$, $b$, where action $a$ loops with probability 1 on the current state, whereas action $b$ moves from both states to state $q_1$, with probability 1. According to the terminology in [8, p. 348] this example models a multichain, not weakly communicating MDP. We assume the initial distribution to be the Dirac function in $q_0$. On this example, only the decision in state $q_0$ is relevant, since $q_1$ is a sink state. For every $n \in \mathbb{N}$, let $d_n \in D^{MR}$ be the Markovian randomized decision rule defined by $d_n(a) = 1 - \frac{1}{n+1}$ and $d_n(b) = \frac{1}{n+1}$. On one hand, the steady-state distribution in $\mathcal{M}$ under the stationary randomized policy $d_n^\infty$ is $\mathsf{sd}(\mathcal{M}, d_n^\infty) = (0, 1)$. On the other hand, the sequence $(d_n)_{n\in\mathbb{N}}$ of decision rules admits a limit: $\lim_{n\to\infty} d_n = d$ with $d(a) = 1$ and $d(b) = 0$, and $\mathsf{sd}(\mathcal{M}, d^\infty) = (1, 0)$. For the next lemma, we introduce further notations. For $d$ a decision rule, we define its *greatest acceptable radius*, denoted $r_d$, as

$$r_d = \max\{r \in \mathbb{R} \mid \forall v \in \mathbb{R}^{|S|},\ ||v - \mathsf{sd}(\mathcal{M}_{d^\infty})|| = r \implies \forall s \in S, v(s) \geq 0\}\ ,$$




where $||\cdot||$ is the Euclidean norm. Intuitively, $r_d$ is the greatest radius of a neighborhood around $\mathcal{M}_{d^\infty}$ such that no element inside it has negative coordinates.

Clearly enough, for a fixed decision rule $d \in D^{MR}$, $r_d > 0$. Indeed, since $\mathcal{M}$ is ergodic, $\mathcal{M}$ equipped with the stationary policy $d^\infty$ is a Markov chain consisting of a single recurrent class; hence, every state has a positive probability in the steady-state distribution $\mathsf{sd}(\mathcal{M}_{d^\infty})$. We also define the following set of distributions, that are $r$-away from a distribution $w$:

$$\mathcal{N}_{=r}(w) = \{v \mid v \in \mathsf{Dist}(S) \text{ and } ||v - w|| = r\}.$$

**Lemma 2.** *Let $\mathcal{M}$ be an ergodic* MDP. *Define*

$$\alpha \stackrel{def}{=} \inf_{d \in D^{MR}} \inf_{v \in \mathcal{N}_{=r_d}(\mathsf{sd}(\mathcal{M}_{d^\infty}))} \frac{||v \cdot (\mathbf{Id} - \mathbf{P}_d)||}{r_d}$$

*Then, $\alpha > 0$.*

*Proof (of Lemma 2).* First observe that for fixed $d \in D^{MR}$ and $v \in \mathcal{N}_{=r_d}(\mathsf{sd}(\mathcal{M}_{d^\infty}))$, $\frac{||v \cdot (\mathbf{Id} - \mathbf{P}_d)||}{r_d} > 0$. Indeed, if $v \in \mathsf{Dist}(S)$ and $||v - \mathsf{sd}(\mathcal{M}_{d^\infty})|| = r_d > 0$, then $v$ is not the steady-state distribution under $d^\infty$, so that $v \neq v \cdot \mathbf{P}_d$.
Towards a contradiction, let us assume that the infimum is 0:

$$\inf_{d \in D^{MR}} \inf_{v \in \mathcal{N}_{=r_d}(\mathsf{sd}(\mathcal{M}_{d^\infty}))} \frac{||v \cdot (\mathbf{Id} - \mathbf{P}_d)||}{r_d} = 0.$$

In this case, there exists a sequence of decisions $(d_n)_{n \in \mathbb{N}} \in D^{MR}$ and a sequence of distributions $(v_n)_{n \in \mathbb{N}}$, such that for each $n \in \mathbb{N}$, $v_n \in \mathcal{N}_{=r_{d_n}}(\mathsf{sd}(\mathcal{M}_{d_n{}^\infty}))$ and $\lim_{n \to \infty} \frac{||v_n \cdot (\mathbf{Id} - \mathbf{P}_{d_n})||}{r_{d_n}} = 0$. From these sequences $(d_n)$ and $(v_n)$, we can extract subsequences, for simplicity still indexed by $n \in \mathbb{N}$ such that:

(i) $(d_n)$ converges, and we write $d$ for its limit, and
(ii) $(v_n)$ converges, and we write $v$ for its limit.

Thanks to Lemma 1, $\lim_{n \to \infty} \mathsf{sd}(\mathcal{M}_{d_n{}^\infty}) = \mathsf{sd}(\mathcal{M}_{d^\infty})$. Moreover, using the continuity of the norm function $||\cdot||$, $\lim_{n \to \infty} r_{d_n} = r_d$, and $v \in \mathcal{N}_{=r_d}(\mathsf{sd}(\mathcal{M}_{d^\infty}))$. Still by continuity, we derive $\frac{||v \cdot (\mathbf{Id} - \mathbf{P}_d)||}{r_d} = 0$, a contradiction. □

$||v \cdot (\mathbf{Id} - \mathbf{P}_d)||$ is the distance between a distribution $v$ and the resulting distribution after applying the decision rule $d$ from $v$. Since we divide it by $r_d$, $\alpha$ roughly represents the minimum deviation "rate" (w.r.t. all $d$'s) between $v$ and its image when going away from $\mathsf{sd}(\mathcal{M}_{d^\infty})$.

**Lemma 3.** *Let $\mathcal{M}$ be an ergodic* MDP. *Assume there exists a policy $\boldsymbol{\pi} \in \Pi^{MR}$ such that $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}})$ exists and is equal to $\delta_{\mathsf{goal}}$. Then for every $\varepsilon > 0$, there exists $d \in D^{MR}$ such that $||\mathsf{sd}(\mathcal{M}_{d^\infty}) - \delta_{\mathsf{goal}}|| < \varepsilon$.*

The above lemma states that if there is a Markovian randomized policy which at steady-state reaches the goal distribution $\delta_{\mathsf{goal}}$, then there must exist a stationary randomized policy $d^\infty$ which at steady-state comes arbitrarily close to $\delta_{\mathsf{goal}}$.

*Proof.* Let us fix some arbitrary $\varepsilon > 0$. Since we assume that $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}}) = \delta_{\mathsf{goal}}$, for all $\gamma > 0$, there exists $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, $||\delta^{\boldsymbol{\pi}}_n - \delta_{\mathsf{goal}}|| < \gamma$. Let us choose $\gamma = \min\{\frac{\alpha\varepsilon}{4}, \frac{\varepsilon}{2}\}$.
Define $d \in D^{MR}$ as the decision made by $\boldsymbol{\pi}$ at the $n_0$-th step. That is, if $\boldsymbol{\pi} = ((d_i)_{i\in\mathbb{N}})$, then $d_{n_0} = d$. Now, if $\delta^{\boldsymbol{\pi}}_{n_0} = \mathsf{sd}(\mathcal{M}_{d^\infty})$ we are done since we will have $||\mathsf{sd}(\mathcal{M}_{d^\infty}) - \delta_{\mathsf{goal}}|| < \epsilon/2 < \epsilon$. Otherwise, we let $\Theta = \frac{r_d}{||\delta^{\boldsymbol{\pi}}_{n_0} - \mathsf{sd}(\mathcal{M}_{d^\infty})||}$ and $v = \Theta\delta^{\boldsymbol{\pi}}_{n_0} + (1-\Theta)\mathsf{sd}(\mathcal{M}_{d^\infty})$. Note that, under these definitions, $v \in \mathcal{N}_{=r_d}(\mathsf{sd}(\mathcal{M}_{d^\infty}))$. Observe also that $v \cdot (\mathbf{Id} - \mathbf{P}_d) = \Theta\delta^{\boldsymbol{\pi}}_{n_0} \cdot (\mathbf{Id} - \mathbf{P}_d)$, by definition of $v$ and since $\mathsf{sd}(\mathcal{M}_{d^\infty}) \cdot \mathbf{P}_d = \mathsf{sd}(\mathcal{M}_{d^\infty})$.
Thus we have

$$\begin{aligned} ||v \cdot (\mathbf{Id} - \mathbf{P}_d)|| &= ||\Theta\delta^{\boldsymbol{\pi}}_{n_0} \cdot (\mathbf{Id} - \mathbf{P}_d)|| \\ &= \Theta||\delta^{\boldsymbol{\pi}}_{n_0} - \delta^{\boldsymbol{\pi}}_{n_0} \cdot \mathbf{P}_d|| = \Theta||\delta^{\boldsymbol{\pi}}_{n_0+1} - \delta^{\boldsymbol{\pi}}_{n_0}|| \\ &\leq \Theta(||\delta^{\boldsymbol{\pi}}_{n_0+1} - \delta_{\mathsf{goal}}|| + ||\delta^{\boldsymbol{\pi}}_{n_0} - \delta_{\mathsf{goal}}||) < \frac{\Theta\alpha\varepsilon}{2}. \end{aligned}$$

By definition of $\alpha$, we have $\alpha \leq \frac{||v\cdot(\mathbf{Id}-\mathbf{P}_d)||}{r_d}$. Then, combining this with the above equation and using the fact that $r_d > 0$, we obtain:

$$r_d \cdot \alpha \leq ||v \cdot (\mathbf{Id} - \mathbf{P}_d)|| < \frac{\Theta\alpha\varepsilon}{2}$$

By Lemma 2 we have $\alpha > 0$ which implies that $r_d < \frac{\Theta\varepsilon}{2}$. Substituting the definition of $\Theta$ we get after simplification:

$$||\delta^{\boldsymbol{\pi}}_{n_0} - \mathsf{sd}(\mathcal{M}_{d^\infty})|| < \frac{\varepsilon}{2}$$

Thus, finally:

$$\begin{aligned} ||\mathsf{sd}(\mathcal{M}_{d^\infty}) - \delta_{\mathsf{goal}}|| &\leq ||\mathsf{sd}(\mathcal{M}_{d^\infty}) - \delta^{\boldsymbol{\pi}}_{n_0}|| + ||\delta^{\boldsymbol{\pi}}_{n_0} - \delta_{\mathsf{goal}}|| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

which proves the lemma. □

Theorem 1 is a consequence of Lemma 3 because stationary randomized policies form a closed set due to Lemma 1. Thanks to Theorems 2 and 1 a naive algorithm to decide the steady-state control problem for MDP is the following: build a linear program whose non negative variables $\{\lambda_e\}$ are indexed by each $e \in D^{SD}$ and check whether $\delta_{\mathsf{goal}} \cdot (\sum_e \lambda_e \mathbf{P}_e) = \delta_{\mathsf{goal}}$ admits a solution with $\sum_e \lambda_e = 1$. This algorithm runs in exponential time w.r.t. the size of $\mathcal{M}$ since there are exponentially many stationary deterministic policies. Yet, a better complexity can be obtained as stated in the following theorem.

**Theorem 3.** *The steady-state control problem for ergodic* MDP *is effectively decidable in PTIME.*

*Proof.* According to Theorem 1, finding a policy with steady-state distribution $\delta_{\mathsf{goal}}$ can be brought back to finding such a policy in $\Pi^{SR}$. Now, on the one hand, defining a randomized stationary policy for an MDP $\mathcal{M}$ consists in choosing decision rules $d_s \in D^{SR}$, or equivalently real numbers $\lambda_{s,a} \in [0,1]$ for each pair $(s,a)$ of states and action, and such that for every $s$, $\sum_{a \in A_s} \lambda_{s,a} = 1$. Intuitively, $\lambda_{s,a}$ represent the probability to choose action $a$ when in state $s$. Note that the set $\Lambda = \{\lambda_{s,a} \mid s \in S, a \in A_s\}$ is of polynomial size (in the size of $\mathcal{M}$). Note also that once we have defined $\Lambda$, we have defined at the same time a policy $\boldsymbol{\pi}_\Lambda \in \Pi^{SR}$, and that $\mathcal{M}_{\boldsymbol{\pi}_\Lambda}$ is a Markov chain with state space $S$. The transition matrix $\mathbf{P}_\Lambda$ of $\mathcal{M}_{\boldsymbol{\pi}_\Lambda}$ is such that $\mathbf{P}_\Lambda[s,s'] = \sum_{a \in A_s} \lambda_{s,a} \cdot p(s|s',a)$.

Due to ergodicity of $\mathcal{M}$, one only has to check whether $\delta_{\mathsf{goal}} \cdot \mathbf{P}_\Lambda = \delta_{\mathsf{goal}}$. Putting this altogether, we can derive a polynomial size linear programming specification of our problem: there exists a stationary randomized policy to achieve $\delta_{\mathsf{goal}}$ if and only if we can find a set of non negative reals $\Lambda = \{\lambda_{s,a} \mid s \in S, a \in A_s\}$ such that

$$\forall s \in S, \sum_{s' \in S, a \in A_{s'}} \delta_{\mathsf{goal}}(s').p(s|s',a).\lambda_{s',a} = \delta_{\mathsf{goal}}(s) \text{ and } \sum_{a \in A_s} \lambda_{s,a} = 1$$

Solving this linear program can be done in polynomial time, using techniques such as the interior point methods (see for instance [9] for details). This proves the overall PTIME complexity. □

*Discussion.* Observe that Lemmas 1, 2 and 3 hold when $\mathcal{M}$ is only recurrent substituting the steady-state distribution of $M_{d^\infty}$ by the (single) invariant distribution of this DTMC. Unfortunately, combining these lemmas in the recurrent case only provides a *necessary* condition namely: "If $\delta_{\mathsf{goal}}$ is the steady-state distribution of some policy then it is the invariant distribution of some of $M_{d^\infty}$".

### 3.2 Labeled Markov decision processes

The steady-state control problem for $\mathsf{MDP}$ describes random processes in which interactions with users or with the environment drive the system towards a desired distribution. However, the goal distribution is a very accurate description of the desired objective and, in particular, this assumes that the controller knows each state of the system. Labeling an $\mathsf{MDP}$ is a way to define higher-level objectives: labels can be seen as properties of states, and the goal distribution as a distribution over these properties. For instance, when resources are shared, a set of labels $L = \{l_1, \ldots, l_k\}$ may indicate which user (numbered from 1 to $k$) owns a particular resource. In such an example, taking $\delta_{\mathsf{goal}}$ as the discrete uniform distribution over $L$ encodes a guarantee of fairness.

In this section, we consider Markov decision processes in which states are labeled. Formally, let $\mathcal{M}$ be a Markov decision process, $L$ be a finite set of labels and $\ell : S \to L$ a labeling function. We consider the following decision problem. Given $(\mathcal{M}, \ell)$ a labeled Markov decision process and $\delta_{\mathsf{goal}} \in \mathsf{Dist}(L)$ a distribution over labels, does there exist a policy $\boldsymbol{\pi}$ such that $\mathsf{lsd}(\mathcal{M}_{\boldsymbol{\pi}}) = \delta_{\mathsf{goal}}$.

The steady-state control problem for $\mathsf{LMDP}$ is a generalization of the same problem for $\mathsf{MDP}$: any goal distribution $\delta_{\mathsf{goal}} \in \mathsf{Dist}(L)$ represents the (possibly infinite) set of distributions in $\mathsf{Dist}(S)$ that agree with $\delta_{\mathsf{goal}}$ when projected onto the labels in $L$.

**Theorem 4.** *The steady-state control problem for ergodic* $\mathsf{LMDP}$ *is decidable in PSPACE.*

*Proof.* Given an ergodic $\mathsf{MDP}$ $\mathcal{M}$ labeled with a labeling function $\ell : S \to L$ and a distribution $\delta_{\mathsf{goal}} \in \mathsf{Dist}(L)$, the question is whether there exists a policy $\boldsymbol{\pi}$ for $\mathcal{M}$ such that the steady-state distribution $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}})$ of $\mathcal{M}$ under $\boldsymbol{\pi}$ projected on labels equals $\delta_{\mathsf{goal}}$.

First, let us denote by $\Delta_{\mathsf{goal}} = \ell^{-1}(\delta_{\mathsf{goal}})$ the set of distributions in $\mathsf{Dist}(S)$ that agree with $\delta_{\mathsf{goal}}$. If $\mathbf{x} = (x_s)_{s \in S} \in \mathsf{Dist}(S)$ is a distribution, then $\mathbf{x} \in \Delta_{\mathsf{goal}}$ can be characterized by the constraints:

$$\forall l \in L, \sum_{s \in S | \ell(s) = l} x_s = \delta_{\mathsf{goal}}(l) \ .$$

We rely on the proof idea from the unlabeled case: If there is a policy $\boldsymbol{\pi} \in \Pi^{HR}$ with steady-state distribution in $\mathsf{Dist}(S)$, then, there is a policy $\boldsymbol{\pi}' \in \Pi^{SR}$ with the same steady-state distribution (thanks to Lemmas and Theorems from Subsection 3.1). This policy $\boldsymbol{\pi}'$ hence consists in repeatedly applying the same decision rule $d \in D^{SR}$: $\boldsymbol{\pi}' = d^\infty$. As the $\mathsf{MDP}$ $\mathcal{M}$ is assumed to be ergodic, there is exactly one distribution $\mathbf{x} \in \mathsf{Dist}(S)$ that is invariant under $\mathbf{P}_d$. The

question then is: Is there a policy $d \in \Pi^{SR}$, and a distribution $\mathbf{x} = (x_s)_{s \in S} \in \Delta_{\mathsf{goal}}$ such that $\mathbf{x} \cdot \mathbf{P}_d = \mathbf{x}$?

We thus derive the following system of equations, over non negative variables $\{x_s \mid s \in S\} \cup \{\lambda_{s,a} \mid s \in S, a \in A_s\}$:

$$\forall l \in L, \sum_{s \in S | \ell(s) = l} x_s = \delta_{\mathsf{goal}}(l)$$

$$\forall s \in S, \sum_{s' \in S, a \in A_{s'}} x_{s'}.p(s \mid s', a).\lambda_{s',a} = x_s \text{ and } \sum_{a \in A_s} \lambda_{s,a} = 1$$

The size of this system is still polynomial in the size of the LMDP. Note that the distribution $\mathbf{x}$ and the weights $\lambda_{s,a}$ in the decision rule $d$ are variables. Therefore, contrary to the unlabeled case, the obtained system is composed of quadratic equations (*i.e.* equations containing products of two variables). We conclude by observing that quadratic equation systems as particular case of polynomial equation systems can be solved in PSPACE [2]. □

Note that the above technique can be used to find policies enforcing *a set of distributions* $\Delta_{\mathsf{goal}}$. Indeed, if expected goals are defined through a set of constraints of the form $\delta_{\mathsf{goal}}(s) \in [a, b]$ for every $s \in S$, then stationnary randomized policies achieving a steady-state distribution in $\Delta_{\mathsf{goal}}$ are again the solutions for a polynomial equation system. Another natural steady-state control problem can be considered for LMDP. Here, the convergence to a steady-state distribution is assumed, and belongs to $\Delta_{\mathsf{goal}} = \ell^{-1}(\delta_{\mathsf{goal}})$. Alternatively, we could consider whether a goal distribution on labels can be realized by some policy $\boldsymbol{\pi}$ even when the convergence of the sequence $(\delta_n^{\boldsymbol{\pi}})_{n \in \mathbb{N}}$ is not guaranteed. This problem is more complex than the one we consider, and is left open.

Finally we study a "language inclusion problem". As mentioned in the introduction, one can define the *steady-state language inclusion problem* similar to the language equivalence problem for LMDP defined in [4]. Formally the steady-state language inclusion problem for LMDP asks whether given two LMDP $\mathcal{M}, \mathcal{M}'$, for every policy $\boldsymbol{\pi}$ of $\mathcal{M}$ such that $\mathsf{lsd}(\mathcal{M}_{\boldsymbol{\pi}})$ is defined there exists a policy $\boldsymbol{\pi}'$ of $\mathcal{M}'$ such that $\mathsf{lsd}(\mathcal{M}_{\boldsymbol{\pi}}) = \mathsf{lsd}(\mathcal{M}'_{\boldsymbol{\pi}'})$. The following theorem establishes its decidability for ergodic LMDP.

**Theorem 5.** *The steady-state language inclusion problem for ergodic* LMDP *is decidable in EXPTIME.*

### 3.3 Partially observable Markov decision processes

In the previous section, we introduced labels in MDP. As already mentioned, this allows us to talk about groups of states having some properties instead of

states themselves. However, decisions are still taken according to the history of the system and, in particular, states are fully observable. In many applications, however, the exact state of a system is only partially known: for instance, in a network, an operator can only know the exact status of the nodes it controls, but has to rely on partial information for other nodes that it does not manage.

Thus, in a partially observable MDP, several states are considered as similar from the observer's point of view. As a consequence, decisions apply to a whole class of similar states, and have to be adequately chosen so that an objective is achieved regardless of which state of the class the system was in.

**Definition 6.** *A partially observable* MDP *(*POMDP *for short) is a tuple* $\mathcal{M} = (S, \{A_s\}_{s\in S}, p, \boldsymbol{s}_0, Part)$ *where* $(S, \{A_s\}_{s\in S}, p, \boldsymbol{s}_0)$ *is an* MDP*, referred to as the* MDP *underlying* $\mathcal{M}$ *and* $Part$ *is a partition of* $S$*.*

The partition $Part$ of $S$ induces an equivalence relation over states of $S$. For $s \in S$, we write $[s]$ for the equivalence class $s$ belongs to, and elements of the set of equivalence classes will be denoted $c, c_0$, etc. We assume that for every $s, s' \in S$, $[s] = [s']$ implies $A_s = A_{s'}$, thus we write $A_{[s]}$ for this set of actions.

**Definition 7.** *Let* $\mathcal{M} = (S, \{A_s\}_{s\in S}, p, \boldsymbol{s}_0, Part)$ *be a* POMDP*.*

- *A history in* $\mathcal{M}$ *is a finite or infinite sequence alternating state equivalence classes and actions* $\sigma = (c_0, a_0, \cdots, c_i, a_i \cdots)$ *such that there exists a history* $(s_0, a_0, \cdots, s_i, a_i \cdots)$ *in the underlying* MDP *with for all* $0 \leq i \leq \lg(\sigma)$*,* $c_i = [s_i]$*.*
- *A decision rule in* $\mathcal{M}$ *associates with every history of length* $t < \infty$ *a distribution* $d_t(\sigma)$ *over* $A_{[s_t]}$*.*
- *A policy of* $\mathcal{M}$ *is finite or infinite a sequence* $\boldsymbol{\pi} = (d_0, \cdots, d_t, \cdots)$ *such that* $d_t$ *is a decision rule at time* $t$*.*

Given a POMDP $\mathcal{M} = (S, \{A_s\}_{s\in S}, p, \boldsymbol{s}_0, Part)$, any policy $\boldsymbol{\pi}$ for $\mathcal{M}$ induces a DTMC written $\mathcal{M}_{\boldsymbol{\pi}}$. The notion of steady-state distributions extends from MDP to POMDP trivially, the steady-state distribution in the POMDP $\mathcal{M}$ under policy $\boldsymbol{\pi}$ is written $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}})$. Contrary to the fully observable case, the steady-state control problem cannot be decided for POMDP.

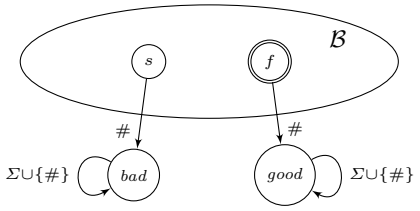
**Theorem 6.** *The steady-state control problem is undecidable for* POMDP*.*

*Proof.* The proof is by reduction from a variant of the emptiness problem for probabilistic finite automata. We start by recalling the definition of probabilistic finite automata (PFA). A PFA is a tuple $\mathcal{B} = (Q, \Sigma, \tau, F)$, where $Q$ is the finite set of states, $\Sigma$ the alphabet, $\tau : Q \times \Sigma \to \mathsf{Dist}(Q)$ defines the probabilistic transition function and $F \subseteq Q$ is the set of final states. The threshold language

emptiness problem asks if there exists a finite word over $\Sigma$ accepted by $\mathcal{B}$ with probability exactly $\frac{1}{2}$. This problem is known to be undecidable [1].
From a PFA $\mathcal{B} = (Q, \Sigma, \tau, F)$, we define a POMDP $\mathcal{M} = (S, \{A_s\}_{s \in S}, p, Part)$:

- $S = Q \cup \{good, bad\}$
- $A_s = \Sigma \cup \{\#\}$, for all $s \in S$
- $p(s'|s, a) = \tau(s, a)(s')$ if $a \in \Sigma$; $p(good|s, \#) = 1$ if $s \in F$; $p(bad|s, \#) = 1$ if $s \in Q \setminus F$; $p(good|good, a) = 1$ for any $a \in A_{good}$ and $p(bad|bad, a) = 1$ for any $a \in A_{bad}$.
- $Part = S$, that is $Part$ consists of a single equivalence class, $S$ itself.

The construction is illustrated below.




Assuming states in $S$ are ordered such that $good$ and $bad$ are the last states, we then let $\delta_{\mathsf{goal}} = (0, \cdots, 0, 1/2, 1/2)$ be the goal distribution (thus assigning probability mass $1/2$ to both $good$ and $bad$). This construction ensures that the answer to the steady-state control problem on $\mathcal{M}$ with $\delta_{\mathsf{goal}}$ is yes if and only if there exists a word $w \in \Sigma^*$ which is accepted in $\mathcal{B}$ with probability $1/2$.

Observe that in the POMDP $\mathcal{M}$ we built, all states are equivalent, so that a policy in $\mathcal{M}$ can only base its decision on the number of steps so far, and thus simply corresponds to a word on $A = \Sigma \cup \{\#\}$. Let us now prove the correctness of the reduction.

($\Longleftarrow$) Given a word $w$ such that $\mathbb{P}_{\mathcal{B}}(w) = 1/2$, in $\mathcal{M}$ we define a policy $\boldsymbol{\pi}$ such that $\boldsymbol{\pi} = w\#^\omega$. Then we can infer that $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}}) = (0, \cdots, 0, 1/2, 1/2)$.

($\Longrightarrow$) First observe that $\boldsymbol{\pi}$ must contain a $\#$-action, otherwise $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}}) = (\_, \cdots, \_, 0, 0)$. Thus we may write $\boldsymbol{\pi} = w\#\rho$ with $w \in \Sigma^*$ and $\rho \in A^\omega$. So we obtain that $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}}) = (0, \cdots, 0, \mathbb{P}_{\mathcal{B}}(w), 1 - \mathbb{P}_{\mathcal{B}}(w))$. From the assumption $\mathsf{sd}(\mathcal{M}_{\boldsymbol{\pi}}) = (0, \cdots, 0, 1/2, 1/2)$, this implies that $\mathbb{P}_{\mathcal{B}}(w) = 1/2$.

This completes the undecidability proof. □

*Remark 1.* In the above undecidability proof, the constructed POMDP is not ergodic. Further, to the best of our knowledge, the undecidability proofs (see for e.g., [1]) for the emptiness of PFA with threshold do not carry over to the ergodic setting. Thus, the status of the steady-state control problem for ergodic POMDP and ergodic PFA are left open in this paper.

## 4 Conclusion

In this paper, we have defined the steady-state control problem for MDP, and shown that this question is decidable for (ergodic) MDP in polynomial time, and for labeled MDP in polynomial space, but becomes undecidable when observation of states is restricted. It is an open question whether our algorithms are optimal and to establish matching lower-bounds or improve the complexities. Further, implementing our decision algorithm is an interesting next step to establish the feasibility of our approach on case studies. We would also like to extend the results to MDP that are not necessarily ergodic, and treat the case of ergodic POMDP. Another possible extension is to consider the control problem with a finite horizon: given an MDP $\mathcal{M}$, a goal distribution $\delta_{\mathsf{goal}}$, and a threshold $\varepsilon$, does there exist a strategy $\boldsymbol{\pi}$ and $k \in \mathbb{N}$ such that $||\delta_k^{\boldsymbol{\pi}} - \delta_{\mathsf{goal}}|| \leq \varepsilon$?

Finally, the results of this paper can have interesting potential applications in diagnosability of probabilistic systems [11]. Indeed, we would design strategies forcing the system to exhibit a steady-state distribution that depends on the occurrence of a fault.

**Acknowledgments.** We warmly thank the anonymous reviewers for their useful comments.

## References

1. A. Bertoni. The solution of problems relative to probabilistic automata in the frame of the formal languages theory. In *Proc. of the 4th GI Jahrestagung*, volume 26 of *LNCS*, pages 107–112. Springer, 1974.
2. J. F. Canny. Some algebraic and geometric computations in PSPACE. In *20th ACM Symp. on Theory of Computing*, pages 460–467, 1988.
3. R. Chadha, V. Korthikanti, M. Vishwanathan, G. Agha, and Y. Kwon. Model checking MDPs with a unique compact invariant set of distributions. In *QEST'11*, 2011.
4. L. Doyen, T. A. Henzinger, and J.-F. Raskin. Equivalence of labeled Markov chains. *Int. J. Found. Comput. Sci.*, 19(3):549–563, 2008.
5. J. G. Kemeny and J. L. Snell. *Finite Markov chains*. Princeton university press, 1960.
6. V. A. Korthikanti, M. Viswanathan, G. Agha, and Y. Kwon. Reasoning about MDPs as transformers of probability distributions. In *QEST*. IEEE Computer Society, 2010.
7. J. R. Norris. *Markov chains*, volume 2 of *Cambridge series on statistical and probabilistic mathematics*. Cambridge University Press, 1997.
8. M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.
9. C. Roos, T. Terlaky, and J.-P. Vial. *Theory and Algorithms for Linear Optimization: An interior point approach*. John Wiley & Sons, 1997.
10. O. Sigaud and O. Buffet. *(editors) Markov decision processes in artifical intelligence*. John Wiley & Sons, 2010.
11. D. Thorsley and D. Teneketzis. Diagnosability of stochastic discrete-event systems. *IEEE Trans. Automat. Contr.*, 50(4):476–492, 2005.