Notions of State Equivalence under Partial Observability

Pablo Samuel Castro, Prakash Panangaden and Doina Precup
School of Computer Science
McGill University
{pcastr,prakash,precup}@cs.mcgill.ca

Abstract

We explore equivalence relations between states in Markov Decision Processes and Partially Observable Markov Decision Processes. We focus on two different equivalence notions: bisimulation (Givan et al, 2003) and a notion of trace equivalence, under which states are considered equivalent roughly if they generate the same conditional probability distributions over observation sequences (where the conditioning is on action sequences). We show that the relationship between these two equivalence notions changes depending on the amount and nature of the partial observability. We also present an alternate characterization of bisimulation based on trajectory equivalence.

1 Introduction

Probabilistic systems are very useful modeling tools in many fields of science and engineering. In order to understand the behavior of existing models, or to provide compact models, notions of equivalence between states in such systems are necessary. Equivalence relations have to be defined in such a way that important properties are preserved, e.g., the *long-term behavior* of equivalent states should be the same. However, there are different ways in which "long-term behavior" could be defined, leading to different equivalence notions. In this paper, we focus on two equivalence relations which have been explored in depth in the process algebra literature: bisimulation [Milner, 1980; Larsen and Skou, 1991] and trace equivalence [Hoare, 1980]. Roughly speaking, two states are bisimilar if they have the same immediate behavior, and they transition with the same probabilities to equivalence classes of states. Two states are trajectory equivalent if they generate the same (conditional) probability distribution over observable system trajectories. At first glance, these notions are quite similar; however, they are not the same, and in particular bisimulation has stronger theoretical guarantees for certain classes of processes.

In this paper, we focus on bisimulation and trace equivalence in the context of Markov Decision Processes (MDPs) [Puterman, 1994] and Partially Observable Markov Decision Processes [Kaelbling *et al.*, 1998]. Bisimulation has been defined for MDPs by Givan et al [2003] and has generated several pieces of follow-up work and extensions (e.g. Dean & Givan[1997], Ferns et al. [2004], Taylor et al. [2009]). Comparatively little work has focused on bisimulation for POMDPs, except for a basic definition of a bisimulation notion for POMDP states [Pineau, 2004] (though the terminology of "bisimulation" is not used there). To our knowledge, trace equivalence has not really been explored in either MDPs or POMDPs. However, using traces holds the potential of offering a more efficient and natural way of computing and approximating state equivalence through sampling methods (rather than the global, model-based process used typically to compute bisimulation).

In this paper we investigate the relationship between bisimulation and trajectory equivalence, focusing on partially observable systems. We show that these two notions are not equivalent in MDPs, but they can be equivalent in POMDPs. We also present a different characterization of bisimulation based on trajectory equivalence. This could potentially yield new algorithms for computing or approximating bisimulation.

The paper is organized as follows. In Sec. 2, we present the definitions and theoretical analysis of the relationship between bisimulation and trajectory equivalence in MDPs. The analysis reveals the surprising fact that trajectory equivalence makes unnecessary distinctions in MDPs. In Sec. 3 we present a weaker version of trajectory equivalence that does not suffer from this problem. In Sec. 4, we consider these equivalence relations in the context of POMDPs, under two reasonable definitions of bisimulation. Finally, in Sec. 5, we discuss our findings and present ideas for future work.

2 Fully Observable States

Definition 2.1. *A* **Markov Decision Process (MDP)** *is a 4-tuple* $M = \langle \mathscr{S}, \mathscr{A}, \mathscr{P}, \mathscr{R} \rangle$*, where* $\mathscr{S}$ *is the set of states;* $\mathscr{A}$ *is the set of actions;* $\mathscr{P} : \mathscr{S} \times \mathscr{A} \to Dist(\mathscr{S})$ *is the next state transition dynamics;* $\mathscr{R} : \mathscr{S} \times \mathscr{A} \to Dist(\mathbb{R})$ *is the reward function.*

We note that most often in the MDP literature, the reward function is defined as a deterministic function of the current state and action. The reward distribution is not explicitly considered because, for the purpose of computing value functions, only the expected value of the reward matters. However, in order to analyze state equivalences, we need to consider the entire distribution, because its higher-order moments (e.g. the variance) may be important. In what follows, we will assume for simplicity that the rewards only take values in a *finite* subset of $\mathbb{R}$, denoted $\mathfrak{R}$. This is done for simplicity of exposition, and all results can be extended beyond this case.

Bisimulation for MDPs is defined in [Givan *et al.*, 2003] for the case in which rewards are deterministic; here, we give the corresponding definition for reward distributions.

Definition 2.2. *Given an MDP $M = \langle \mathscr{S}, \mathscr{A}, \mathscr{P}, \mathscr{R} \rangle$, an equivalence relation $R : \mathscr{S} \times \mathscr{S} \to \{0,1\}$ is defined to be a bisimulation relation if whenever sRt the following properties hold:*

1. $\forall a \in \mathscr{A}. \forall r \in \mathbb{R}. \mathscr{R}(s,a)(r) = \mathscr{R}(t,a)(r)$
2. $\forall a \in \mathscr{A}. \forall c \in \mathscr{S}/R. \mathscr{P}(s,a)(c) = \mathscr{P}(t,a)(c)$, *where* $\mathscr{P}(s,a)(c) = \sum_{s' \in c} \mathscr{P}(s,a)(s')$,

where $\mathscr{S}/R$ denotes the partition of $\mathscr{S}$ into R-equivalence classes. Two states s and t are **bisimilar**, *denoted $s \sim t$, if there exists a bisimulation relation R such that sRt.*

We will now define the notion of trajectory equivalence for MDP states, in a similar vein to the notion of trace equivalence for labelled transition systems [Hoare, 1980]. Intuitively, two states are trajectory equivalent if they produce the same trajectories. In MDPs, in order to define an analogous notion, we will need to give a similar, probabilistic definition *conditional* on action sequences (since actions can be independently determined by a controller or policy).

Definition 2.3. *An* **action sequence** *is a function $\theta : \mathbb{N}^+ \mapsto \mathscr{A}$ mapping a time step to an action. Let Θ be the set of all action sequences. Let $N : \Theta \mapsto \Theta$ be a function which returns the tail of any sequence of actions: $\forall i \in \mathbb{N}^+. \theta(i+1) = N(\theta)(i)$.*

Consider any *finite* reward-state trajectory $\alpha \in (\mathbb{R} \times \mathscr{S})^*$ and let $Pr(\alpha|s,\theta)$ be the probability of observing α when starting in state $s \in \mathscr{S}$ and choosing the actions specified by θ.

Definition 2.4. *Given an MDP, the states $s,t \in \mathscr{S}$ are* **trajectory equivalent** *if and only if $\forall \theta \in \Theta$ and for any finite reward-state trajectory α,*

$$Pr(\alpha|s,\theta) = Pr(\alpha|t,\theta).$$

We note that conditioning on state-independent (open-loop) sequences of actions may be considered non-standard for MDPs, where most behavior is generated by state-conditional policies (in which the choice of action depends on the state). We focus here on open-loop sequences because this is the closest match to trace equivalence. We conjecture that a very similar analysis can be performed for closed-loop policies, but we leave this for future work.

We are now ready to present our main results relating trajectory equivalence and bisimulation in MDPs. The following lemma can be proved easily by considering one-step trajectories.

Lemma 2.5. *Trajectory equivalence implies model equivalence.*

The following theorem is a direct consequence of Lemma 2.5.

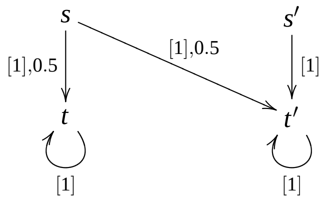

Figure 1: Example showing that bisimulation does not imply trajectory equivalence.

Theorem 2.6. *Trajectory equivalence implies bisimulation.*

Theorem 2.7. *Bisimulation does not imply trajectory equivalence.*

Proof. Consider the MDP depicted in Figure 1, with 4 states and only one action. In this, as well as in all subsequent examples, the annotations on the links represent the rewards received (in brackets) and the transition probabilities. In this MDP, t and t' are bisimilar, and thus, s and s' are also bisimilar. Note that there is only one possible infinite action sequence θ, since there is only one action. Let $\alpha = \langle (1,t) \rangle$. Then $Pr(\alpha|s,\theta) = 0.5 \neq 0 = Pr(\alpha|s',\theta)$. Thus, s and s' are not trajectory equivalent. □

These results show that trajectory equivalence is a sufficient but not necessary condition for bisimulation. This result seems counterintuitive, as bisimulation is considered perhaps the strongest equivalence notion in the process algebra literature. Upon closer inspection, one can notice that this result is due to the full state observability in an MDP. More precisely, because the identity of the state is fully observable, and is included in the trajectory, very fine distinctions are made between trajectories. This is undesirable if one wants an equivalence notion that is useful, for example, in reducing the state space of an MDP. With the current definition of trajectory equivalence, even completely disjoint but otherwise identical subsets of the MDP would be considered distinct, as long as their states are numbered differently. Hence, we will now consider a weaker version of trajectory equivalence, which is closer in spirit to bisimulation, and has more desirable properties.

3 A Different Notion of Trajectory Equivalence

In order to define a more appropriate notion of trajectory equivalence, we need to allow the exact state identity to not appear in the trajectory. In bisimulation, the equivalence relation R is essentially used to partition the state space into partitions. Afterwards, essentially the identity of a state is replaced by the partition to which it belongs (as follows from the second condition in Definition 2.2). To exploit this idea, we will consider now trajectory equivalence when the state space is partitioned, and the identity of a state is replaced by the identity of the partition to which it belongs.

Let $\Psi(\mathscr{S})$ be a partitioning of the state space into disjoint subsets and $\psi : \mathscr{S} \to \Psi(\mathscr{S})$ be the function mapping each state to its corresponding partition in $\Psi(\mathscr{S})$. Consider any finite reward-partition trajectory $\kappa \in (\mathfrak{R} \times \Psi(\mathscr{S}))^*$ and let $Pr(\kappa|s,\theta)$ be the probability of observing κ when starting in state $s \in \mathscr{S}$ and choosing the actions specified by θ.

Definition 3.1. *Given an MDP $M = \langle \mathscr{S}, \mathscr{A}, \mathscr{P}, \mathscr{R} \rangle$ and a decomposition $\Psi(\mathscr{S})$, two states $s,t \in \mathscr{S}$ are* **Ψ-trajectory equivalent** *if and only if $\psi(s) = \psi(t)$ and $\forall \theta \in \Theta$ and for any finite reward-partition trajectory κ, $Pr(\kappa|s,\theta) = Pr(\kappa|t,\theta)$.*

If $\Psi(\mathscr{S}) = \mathscr{S}$ and ψ is the identity function, we have trajectory equivalence as defined in Sec. 2. Note, however, that if Ψ is defined in an arbitrary way, this notion of equivalence may not be useful at all.

Given that bisimulation distinguishes states with different rewards, it is natural to define a clustering $\Psi_R(\mathscr{S})$ such that $\psi_R(s) = \psi_R(s')$ if and only if $\forall a \in \mathscr{A}. \forall r. \mathscr{R}(s,a)(r) = \mathscr{R}(s',a)(r)$. Let ψ_R be its equivalent membership function.

Theorem 3.2. *Ψ_R-trajectory equivalence does not imply bisimulation.*

Proof. Consider the MDP in Figure 2, in which there is again only one action. We can see that $\Psi_R(\mathscr{S}) = \{c_0, c_1, c_2\}$, where $c_0 = \{s, s'\}$, $c_1 = \{t_1, t_2, t', u_1, u_1'\}$ and $c_2 = \{u_2, u_2'\}$. Both s and s' observe c_1 w.p.1 in the first step. For any trajectories of length $n > 1$, $\langle 0, c_1\rangle\langle 1, c_1\rangle^{n-1}$ and $\langle 0, c_1\rangle\langle 1, c_1\rangle\langle 2, c_2\rangle^{n-2}$ are observed w.p. 0.5 each. Thus, s and s' are Ψ_R-trajectory equivalent. However, they are not bisimilar since neither t_1 nor t_2 is bisimilar to t'. □

Lemma 3.3. *For all $c \in \mathscr{S}/_{\sim}$ and $d \in \Psi_R(\mathscr{S})$, either $c \subseteq d$ or $c \cap d = \emptyset$.*

Proof. Without loss of generality assume $c \cap d \neq \emptyset$. If c contains only one state s, then $c \subseteq \psi(s)$. Now suppose that c has at least two states. For any two states $s, s' \in c$, from Def. 2.2, we have that $\forall a \in \mathscr{A}, r. \mathscr{R}(s,a)(r) = \mathscr{R}(s',a)(r) \Rightarrow \psi_R(s) = \psi_R(s')$, so $s, s' \in d$. □

Lemma 3.4. *For all $d \in \Psi_R(\mathscr{S})$ there exists a set $C \subseteq \mathscr{S}/_{\sim}$ such that $\bigcup_{c \in C} c = d$.*

Proof. Immediate from Lemma 3.3 and the fact that $\bigcup_{c \in \mathscr{S}/_{\sim}} c = \bigcup_{d \in \Psi_R(\mathscr{S})} d = \mathscr{S}$. □

Theorem 3.5. *Bisimulation implies Ψ_R-trajectory equivalence.*

Proof. Assume $s_0 \sim t_0$. Take any $\theta \in \Theta$ and any finite trajectory κ. The proof is by induction on the length of κ.
Base case: $|\kappa| = 1$. Say $\kappa = \langle d \rangle$. Let $a = \theta(0)$. By Lemma 3.4 there exists $C \subseteq \mathscr{S}/_{\sim}$ such that $\bigcup_{c \in C} c = d$. Therefore:

$$\begin{aligned} Pr(\kappa|s_0,\theta) &= \sum_{s' \in d} \mathscr{P}(s_0,a)(s') = \sum_{c \in C} \sum_{s' \in c} \mathscr{P}(s_0,a)(s') \\ &= \sum_{c \in C} \mathscr{P}(s_0,a)(c) = \sum_{c \in C} \mathscr{P}(t_0,a)(c), \text{ because } s_0 \sim t_0 \\ &= Pr(\kappa|t_0,\theta) \end{aligned}$$

Induction step: Assume that the claim holds up to $|\kappa| = n-1$. Let $\kappa = \langle d_1, \cdots, d_n \rangle$ and $\kappa' = \langle d_2, \cdots, d_n \rangle$. As before, let $a = \theta(0)$. Again, by Lemma 3.4, there exists C such that $\bigcup_{c \in C} c = d$. We have:

$$\begin{aligned} Pr(\kappa|s_0,\theta) &= \sum_{s1 \in d_1} \mathscr{P}(s_0,a)(s_1) Pr(\kappa'|s_1, N(\theta)) \\ &= \sum_{c \in C} \sum_{s_1 \in c} \mathscr{P}(s_0,a)(s_1) Pr(\kappa'|s_1, N(\theta)) \end{aligned}$$

From the induction hypothesis, $Pr(\kappa'|s_1, N(\theta))$ is the same $\forall s_1 \in c$, so we can denote this by $Pr(\kappa'|c, N(\theta))$, Hence, continuing from above, we have:

$$\begin{aligned} &= \sum_{c \in C} Pr(\kappa'|c, N(\theta)) \sum_{s_1 \in c} \mathscr{P}(s_0,a)(s_1) \\ &= \sum_{c \in C} \mathscr{P}(s_0,a)(c) Pr(\kappa'|c, N(\theta)) \\ &= \sum_{c \in C} \mathscr{P}(t_0,a)(c) Pr(\kappa'|c, N(\theta)), \text{because } s_0 \sim t_0 \\ &= Pr(\kappa|t_0,\theta) \end{aligned}$$

which concludes the proof. □

Theorems 3.2 and 3.5 are closer to what we would normally expect for these notions. The fact that trajectory equivalence is weaker is not surprising, since bisimulation has a "recursive" nature that is lacking in Ψ_R-trajectory equivalence. We now proceed by iteratively strengthening Ψ_R-trajectory equivalence to bring it closer to bisimulation.

Let Γ be an operator that takes a partitioning $\Psi(\mathscr{S})$ and returns a more refined decomposition as follows. For any subset $d \in \mathscr{S}$, $d \in \Gamma(\Psi(\mathscr{S}))$ if and only if, for any two states $s, t \in d$, we have:

1. For any $a \in \mathscr{A}$ and $\forall r$, $\mathscr{R}(s,a)(r) = \mathscr{R}(t,a)(r)$;
2. s and t and Ψ-trajectory equivalent.

Let $\Gamma^{(n)}$ denote the n-th iterate of Γ. It is clear that $\Gamma(\Psi_R(\mathscr{S}))$ equivalence is Ψ_R-trajectory equivalence. Using Theorem 3.5, it is easy to prove that bisimulation implies $\Gamma^{(n)}(\Psi_R(\mathscr{S}))$ equivalence by induction. Similarly, it can be shown that for every n, $\Gamma^{(n)}(\Psi_R(\mathscr{S}))$ does not imply bisimulation. The counterexamples are similar in spirit to the one from Theorem 3.2, but they grow linearly in height and exponentially in width with n.

Theorem 3.6. *The iterates Γ^n have a fixed point, Γ^*.*

Proof. Define a binary relation $\sqsupseteq$ on the set of partitionings of $\mathscr{S}$, where for any $D_1(\mathscr{S})$ and $D_2(\mathscr{S})$, $D_1(\mathscr{S}) \sqsupseteq D_2(\mathscr{S})$ if and only if for any $d_1 \in D_1(\mathscr{S})$ and $d_2 \in D_2(\mathscr{S})$, either $d_1 \cap d_2 = \emptyset$ or $d_2 \subseteq d_1$. It is easy to see that the set of all possible partitions of $\mathscr{S}$ along with $\sqsupseteq$ constitute a complete partial order with bottom, where bottom is simply $\Psi_R(\mathscr{S})$. It then follows from Theorem 5.11 in [Winskel, 1993] that Γ^* exists and is well defined. □

From the results so far, it is easy to see that bisimulation implies Γ^* equivalence. We now show that the reverse is also true.

Theorem 3.7. *Γ^*-equivalence implies bisimulation.*

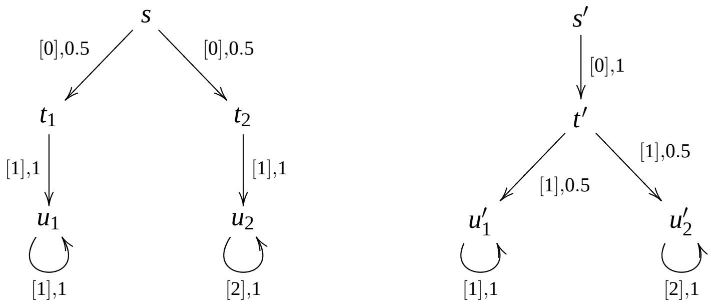

Figure 2: Counterexample showing that Ψ_R-trajectory equivalence does not imply bisimulation

Proof. Let R be the Γ^*-equivalence relation. Given s and t with sRt, we will show $s \sim t$ by checking the conditions of Def 2.2. The first condition follows from the definition of Γ. The second condition follows from the definition of Γ and the fact that Γ^* is a fixed point. □

Hence, we have obtained a new fixed-point characterization of bisimulation in terms of this new notion of trajectory equivalence.

4 Equivalences in Partially Observable Markov Decision Processes

We now turn our attention to the case of partial observability.

Definition 4.1. *A* **Partially Observable Markov Decision Process (POMDP)** *is a 6-tuple $M = \langle \mathscr{S}, \mathscr{A}, \mathscr{P}, \mathscr{R}, \Omega, \mathscr{O} \rangle$, where $\langle \mathscr{S}, \mathscr{A}, \mathscr{P}, \mathscr{R} \rangle$ define an MDP; Ω is a finite set of observations; and $\mathscr{O} : \mathscr{S} \times \mathscr{A} \mapsto Dist(\Omega)$ is the observation distribution function, with $\mathscr{O}(s,a)(\omega) = Pr(o_{t+1} = \omega | s_{t+1} = s, a_t = a)$.*

A **belief state** b is a distribution over $\mathscr{S}$, quantifying the uncertainty in the system's internal state. Let $\mathscr{B}$ be the set of all belief states over $\mathscr{S}$. After performing an action $a \in \mathscr{A}$ and witnessing observation $\omega \in \Omega$ from belief state b, the function $\tau : \mathscr{B} \times \mathscr{A} \times \Omega \mapsto \mathscr{B}$ computes the new belief state $b' = \tau(b,a,\omega)$ as follows, $\forall s' \in \mathscr{S}$:

$$b'(s') = Pr(s'|\omega,a,b) = \frac{\mathscr{O}(s',a)(\omega)\sum_{s\in\mathscr{S}}\mathscr{P}(s,a)(s')b(s)}{Pr(\omega|a,b)}$$

$$\text{where } Pr(\omega|b,a) = \sum_{s'\in\mathscr{S}} \mathscr{O}(s',a)(\omega) \sum_{s\in\mathscr{S}} \mathscr{P}(s,a)(s')b(s)$$

Many standard approaches replace the POMDP with a corresponding, continuous-state **belief MDP** $\langle \mathscr{B}, \mathscr{A}, \mathscr{T}, \rho \rangle$, where $\mathscr{B}$ is the (continuous) state space; $\mathscr{A}$ is the action set; the transition probability function $\mathscr{T} : \mathscr{B} \times \mathscr{A} \mapsto Dist(\mathscr{B})$ is defined as $\mathscr{T}(b,a)(b') = \sum_{\omega\in\Omega} Pr(b'|b,a,\omega)Pr(\omega|a,b)$ where $Pr(\omega|a,b)$ is as defined above and $Pr(b'|b,a,\omega) = \mathbb{1}_{b'=\tau(b,a,\omega)}$; and the reward function $\rho : \mathscr{B} \times \mathscr{A} \mapsto Dist(\mathbb{R})$ is defined as: $\rho(b,a)(r) = \sum_{s\in\mathscr{S}} b(s)\mathscr{R}(s,a)(r)$

Consider any finite reward-observation trajectory $\beta \in (\mathbb{R} \times \Omega)^*$ and let $Pr(\beta|b,\theta)$ be the probability of observing β when starting in belief state b and choosing the actions dictated by θ.

Definition 4.2. *Given a POMDP, two belief states b,c are* **belief trajectory equivalent** *if and only if $\forall \theta \in \Theta$ and for any finite reward-observation trajectory β, $Pr(\beta|b,\theta) = Pr(\beta|c,\theta)$.*

Lemma 4.3. *Belief trajectory equivalence implies model equivalence.*

Proof. Assume that $b,c \in \mathscr{B}$ are belief trajectory equivalent. Take any $a \in \mathscr{A}$ and $r \in \mathbb{R}$. Take any $\theta \in \Theta$ with $\theta(0) = a$. From belief trajectory equivalence, we have:

$$\begin{aligned}\rho(b,a)(r) &= \sum_{\omega\in\Omega} Pr(\langle(r,\omega)\rangle|b,\theta)\\ &= \sum_{\omega\in\Omega} Pr(\langle(r,\omega)\rangle|c,\theta) = \rho(c,a)(r)\end{aligned}$$

Similarly, $\forall\omega \in \Omega . Pr(\omega|b,a) = Pr(\omega|c,a)$. □

Lemma 4.4. *If $b,c \in \mathscr{B}$ are belief trajectory equivalent, then for any $a \in \mathscr{A}$ and $\omega \in \Omega$, $\tau(b,a,\omega)$ and $\tau(c,a,\omega)$ are belief trajectory equivalent.*

Proof. We need to show that for any finite reward-observation trajectory α, $\theta \in \Theta$, $a \in \mathscr{A}$ and $\omega \in \Omega$ we have that $Pr(\alpha|\tau(b,a,\omega),\theta) = Pr(\alpha|\tau(c,a,\omega),\theta)$.

Let θ' be a new action sequence s.t. $\theta'(0) = a$ and $N(\theta') = \theta$. Taking an arbitrary reward r, construct a new reward-observation trajectory α' where $\alpha' = \langle(r,\omega),\alpha\rangle$. We know $Pr(\alpha'|b,\theta') = Pr(\alpha'|c,\theta')$ since b and c are belief trajectory equivalent. We also know that

$$Pr(\alpha'|b,\theta') = \rho(b,a)(r)Pr(\omega|b,a)Pr(\alpha|\tau(b,a,\omega),\theta) \text{ and}$$
$$Pr(\alpha'|c,\theta') = \rho(c,a)(r)Pr(\omega|c,a)Pr(\alpha|\tau(c,a,\omega),\theta)$$

From Lemma 4.3, $\rho(b,a)(r) = \rho(c,a)(r)$ and $Pr(\omega|b,a) = Pr(\omega|c,a)$. So $Pr(\alpha|\tau(b,a,\omega),\theta) = Pr(\alpha|\tau(c,a,\omega),\theta)$, and since α,a,θ,ω were all chosen arbitrarily, the proof concludes. □

Previous work on POMDPs defines bisimulation between internal POMDP states. Instead, here we want to define bisimulation between belief states. However, there are two possible, reasonable definitions that one could adopt, which we present below.

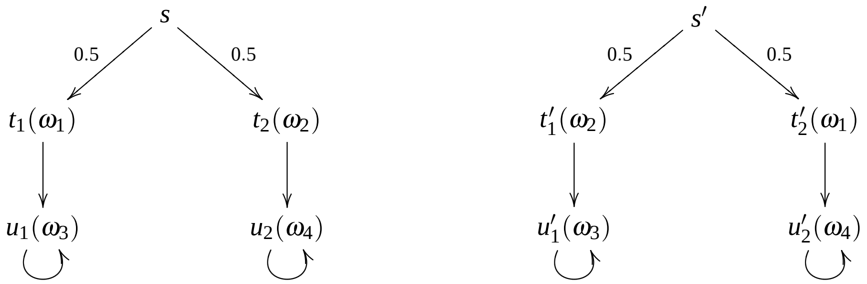

Figure 3: Example showing that weak belief bisimulation does not imply trajectory equivalence.

Definition 4.5. *A relation $R \subseteq \mathscr{B} \times \mathscr{B}$ is defined to be a* **weak belief bisimulation relation**[1] *if whenever bRc, the following properties hold:*

1. *$\forall a \in \mathscr{A}.\forall r.\rho(b,a)(r) = \rho(c,a)(r)$*
2. *$\forall a \in \mathscr{A}.\forall \omega \in \Omega.Pr(\omega|b,a) = Pr(\omega|c,a)$*
3. *For any $a \in \mathscr{A}$ and $B \in \mathscr{B}/R$, $Pr(B|b,a) = Pr(B|c,a)$, where*

$$Pr(B|b,a) = \sum_{b' \in B} \mathscr{T}(b,a)(b')$$

Two belief states b,c are **weakly belief bisimilar**, *denoted $b \approx_w c$, if there exists a weak belief bisimulation relation R such that bRc.*

Definition 4.6. *A relation $\mathscr{R} \subseteq \mathscr{B} \times \mathscr{B}$ is a* **strong belief bisimulation relation** *if it respects the first two conditions of Def. 4.5, and the following third condition:*

3. *$\forall a \in \mathscr{A}.\forall \omega \in \Omega$, $\tau(b,a,\omega)$ and $\tau(c,a,\omega)$ are strongly belief bisimilar.*

Two belief states b,c are **strongly belief bisimilar**, *denoted $s \approx t$, if there exists a strong belief bisimulation relation R such that bRc.*

Since both bisimulation definitions are quite similar in spirit, one would expect them to be equivalent. However, as we will now show, this is not the case.

Lemma 4.7. *Any strong belief bisimulation is also a weak belief bisimulation.*

Proof. Let R be a strong belief bisimulation. Take any two belief states b and c such that bRc. The first two conditions in Def. 4.5 and Def. 4.6 are identical, so we only need to prove that the third condition in Def. 4.5 holds. Consider an arbitrary $B \in \mathscr{B}/_R$ and $a \in \mathscr{A}$. We have:

$$\begin{aligned}
Pr(B|b,a) &= \sum_{b' \in B} \mathscr{T}(b,a)(b') = \sum_{b' \in B} \sum_{\omega \in \Omega} Pr(b'|b,a,\omega)Pr(\omega|b,a) \\
&= \sum_{\omega \in \Omega} Pr(\omega|b,a) \sum_{b' \in B} Pr(b'|b,a,\omega) \\
&= \sum_{\omega \in \Omega} Pr(\omega|b,a) \sum_{b' \in B} \mathbb{1}_{b'=\tau(b,a,\omega)} \\
&= \sum_{\omega \in \Omega} Pr(\omega|c,a) \sum_{b' \in B} \mathbb{1}_{b'=\tau(b,a,\omega)} \\
&= \sum_{\omega \in \Omega} Pr(\omega|c,a) \sum_{b' \in B} \mathbb{1}_{b'=\tau(c,a,\omega)} \text{ (from Def. 4.6)} \\
&= Pr(B|c,a)
\end{aligned}$$

The last step follows because $\tau(b,a,\omega)R\tau(c,a,\omega)$ implies that $\forall B \in \mathscr{B}/R$, $\tau(b,a,\omega) \in B$ if and only if $\tau(c,a,\omega) \in B$. □

Lemmas 4.3 and 4.4 are sufficient conditions for strong belief bisimilarity. This observation, combined with Lemma 4.7 yields the following corollary.

Corollary 4.8. *Belief trajectory equivalence implies strong and weak belief bisimulation.*

Theorem 4.9. *Strong belief bisimulation implies belief trajectory equivalence.*

The proof uses the definition of strong belief bisimilarity and by induction on the length of the trajectory. It is very similar to previous proofs, and we omit it for succinctness.

Theorem 4.10. *Weak belief bisimulation does not imply belief trajectory equivalence.*

Proof. Consider the POMDP in Figure 3. There is only one available action and we assume all transitions yield the same reward. The observation received upon entering a state is indicated in parentheses next to the state name.

Let θ be the only available action sequence, and denote by δ_s the belief state concentrated at state s. We have: $Pr(\omega_1|\delta_s,\theta) = Pr(\omega_1|\delta_{s'},\theta) = 0.5$ and $Pr(\omega_2|\delta_s,\theta) = Pr(\omega_2|\delta_{s'},\theta) = 0.5$. Furthermore, $\delta_{u_1} \approx_w \delta_{u'_1}$, $\delta_{u_2} \approx_w \delta_{u'_2}$, implying that $\delta_{t_1} \approx_w \delta_{t'_1}$ and $\delta_{t_2} \approx_w \delta_{t'_2}$, and hence $\delta_s \approx_w \delta_{s'}$. However, δ_s and $\delta_{s'}$ are not belief trajectory equivalent since

$$Pr(\langle\omega_1,\omega_3\rangle|\delta_s,\theta) = 0.5 \neq 0 = Pr(\langle\omega_1,\omega_3\rangle|\delta_{s'},\theta)$$

□

[1] We do not use 'weak' and 'strong' here in the same sense as [Milner, 1980].

Note that this result is due mainly to the fact that the observation is obtained upon *entering* a state, and past observations are in some sense not taken into account.

5 Discussion and Future Work

We analyzed the relationship between bisimulation and trajectory equivalence in MDPs and POMDPs. When the state is fully observable, trajectory equivalence is stronger than bisimulation, because it distinguishes between differences in transition probabilities to individual states. Bisimulation, on the other hand, can only distinguish between differences in transition probabilities to classes of bisimilar states.

By considering partitions over states, we obtained a weaker notion than bisimulation. We showed that bisimulation can be characterized as the fixed point of a sequence of iterates in which states are initially aggregated according to their immediate reward. K-moment equivalence, presented in [Zhioua, 2008], is somewhat similar to our method and bisimulation is only reached in the limit. Their method, however, still requires replicating states at the end of trajectories to compute the equivalence, whereas our approach only uses normal trajectories.

We gave two definitions of bisimulation over belief states for POMDPs, which at first sight seem very similar, but they are not. The fact that strong belief bisimulation is equivalent to belief trajectory equivalence is not surprising, because the belief MDP is deterministic: from a belief state b, for a given action a and observation ω, there is exactly *one* reachable belief state. It is well known in the process algebra literature that trace equivalence and bisimulation are identical for deterministic automata. If we did not consider the belief states, but rather, the underlying states, we would be in a situation similar to the one presented in Sec. 3, considering that states that yield the same observations upon arrival would be grouped together.

The Γ iterative operator provides an alternative way of computing bisimulation classes. It would be interesting to analyze the number of iterations required to reach the fixed point Γ^*. This approach could yield an alternative algorithm for computing bisimulation classes, and could potentially be extended to a metric, in the spirit of [Ferns *et al.*, 2004]. The advantage of our method compared to other bisimulation constructions is that one can accumulate a set of trajectories from action sequences and then approximate Ψ_R-trajectory equivalence, and further $\Gamma^{(n)}(\Psi_R(\mathscr{S}))$-equivalence. This would not require knowing the system model, and performance should improve as the number of trajectories gathered increases. We plan to study this idea, as well as algorithms for efficiently gathering trajectories, in future work.

Note that two belief states are belief trajectory equivalent if and only if they have the same probability of witnessing all linear PSR tests [Littman *et al.*, 2002], since linear PSR tests are essentially finite reward-observation trajectories. This means that one can compute trajectory equivalence by means of a PSR model. This idea will be further studied in future work.

Acknowledgments

This work was funded in part by NSERC. The authors would like to thank the reviewers for their helpful comments and for suggesting the relationship between belief trajectory equivalence and linear PSR tests.

References

[Dean and Givan, 1997] Thomas Dean and Robert Givan. Model minimization in Markov Decision Processes. In *Proceedings of AAAI*, pages 106–111, 1997.

[Ferns *et al.*, 2004] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence*, pages 162–169, 2004.

[Givan *et al.*, 2003] Robert Givan, Thomas Dean, and Matthew Greig. Equivalence Notions and Model Minimization in Markov Decision Processes. *Artificial Intelligence*, 147(1–2), 2003.

[Hoare, 1980] C.A.R. Hoare. *Communicating Sequential Processes. On the Construction of Programs–an Advanced Course*. Cambridge University Press, 1980.

[Kaelbling *et al.*, 1998] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

[Larsen and Skou, 1991] Kim Guldstrand Larsen and Arne Skou. Bisimulation through probabilistic testing. *Information and Computation*, 94(1):1–28, 1991.

[Littman *et al.*, 2002] Michael Littman, Richard Sutton, and Satinder Singh. Predictive representations of state. In *Proceedings of the 14th Conference on Advances in Neural Information Processing Systems (NIPS-02)*, pages 1555–1561, 2002.

[Milner, 1980] Robin Milner. *A Calculus of communicating systems*. Springer-Verlag, New York, NY, 1980.

[Pineau, 2004] Joelle Pineau. *Tractable Planning Under Uncertainty: Exploiting Structure*. PhD thesis, Carnegie Mellon University, 2004.

[Puterman, 1994] Martin L. Puterman. *Markov Decision Processes*. John Wily & Sons, New York, NY, 1994.

[Taylor *et al.*, 2009] Jonathan Taylor, Doina Precup, and Prakash Panangaden. Bounding performance loss in approximate MDP homomorphisms. In *Advances in Neural Information Processing Systems 21*, page In press, 2009.

[Winskel, 1993] Glynn Winskel. *The Formal Semantics of Programming Languages*. MIT Press, Cambridge, Massachusetts, 1993.

[Zhioua, 2008] Sami Zhioua. *Stochastic systems divergence through reinforcement learning*. PhD thesis, Université Laval, 2008.