On Average Reward Policy Evaluation in Infinite–State Partially Observable Systems

Yuri Grinberg
School of Computer Science, McGill University
ygrinb@cs.mcgill.ca

Doina Precup
School of Computer Science, McGill University
dprecup@cs.mcgill.ca

Abstract

We investigate the problem of estimating the average reward of given decision policies in discrete-time controllable dynamical systems with finite action and observation sets, but possibly infinite state space. Unlike in systems with finite state spaces, in infinite–state systems the expected reward for some policies might not exist, so policy evaluation, which is a key step in optimal control methods, might fail. Our main analysis tool is Ergodic theory, which allows learning potentially useful quantities from the system without building a model. Our main contribution is three-fold. First, we present several dynamical systems that demonstrate the difficulty of learning in the general case, without making additional assumptions. We state the necessary condition that the underlying system must satisfy to be amenable for learning. Second, we discuss the relationship between this condition and state-of-the-art predictive representations, and we show that there are systems that satisfy the above condition but cannot be modeled by such representations. Third, we establish sufficient conditions for average-reward policy evaluation in this setting.

1 Introduction

The problem of learning from interaction with an unknown system (i.e., environment) is central both to artificial intelligence and to control theory. In this paper, we focus on a well-studied setting, in which the interaction happens in discrete time, assuming that actions can be chosen from a finite set and discrete observations from a finite set are received from the system. This setting is common, for example, in reinforcement learning (RL) problems, where many algorithms rely on estimating a quantity of interest (e.g., value function [16, 15, 8]) conditioned on a particular way of choosing actions (i.e., policy). Under certain assumptions on the underlying system (e.g., the system is a Fully or Partially Observable Markov Decision Process, often with a finite state space), the quantities of interest can be estimated in various ways and used to compare policies, with the goal of improving the behavior. In this paper, we are interested in this scenario, but without making traditional assumptions about the system; in particular, the system may have infinite state space, even though the set of actions and observations is finite. This setting brings difficulties to the estimation procedure; specifically the expectation of returns (which represents the value function) might not exist. This obstacle does not exist in systems with finite state spaces [7].

We focus on the problem of estimating the average reward for policies in partially observable environments, which is a key ingredient in solving an average reward RL problem ([6]). For the analysis, we draw on the well established Ergodic theory [3], which focuses on systems without control. Suppose that we observe an unknown process. The conventional way to estimate the probability of a particular event is to find the empirical estimate through a long running average. Ergodic theory provides the necessary and sufficient condition for this procedure to be sound: the system has to be *Asymptotically Mean Stationary* (AMS, [2]). Our strategy will be to examine what types of controllable systems, together with a fixed policy, satisfy the AMS property. This property, in particular, guarantees the convergence of the running average of collected rewards.

After some preliminaries, we present three examples of systems that highlight the difficulty of obtaining such a characterization in a general case. Then, we state a necessary condition for the controllable system to be learnable in this setting, and discuss its relationship to current predictive state representation models. In particular, we give an example of an AMS system that cannot be modelled by current state-of-the-art representations. Finally, we show that this condition together with a couple of other auxiliary assumptions allows for policy evaluation by averaging collected rewards.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

2 Background

For the rest of the paper we assume a finite action set $\mathcal{A}$ (from which an agent interacting with the system can choose) and a finite observation set $\mathcal{O}$, with the corresponding random variables A and O that take values from these sets. We define a controllable dynamical system as a tuple $(\Omega, \mathcal{F}, T, \mu_0)$, where $(\Omega, \mathcal{F})$ is a measurable space, μ_0 is a probability kernel representing the effect of actions, and T is a shift operator. For example, the elements of Ω can be infinite sequences of action–observation pairs and $\mathcal{F}$ the σ-algebra generated by all finite–dimensional cylinder sets. In this case, μ_0 will represent the initial distribution over observations given actions, and the operator T shifts pairs of actions and observations, meaning that for $a_i \in \mathcal{A}$ and $o_i \in \mathcal{O}$:

$$T(\langle a_1, o_1, a_2, o_2, ...\rangle) = \langle a_2, o_2, a_3, o_3 ...\rangle, \text{ and}$$

$$T^{-1}(\langle a_1, o_1, ...\rangle) = \{\langle a, o, a_1, o_1, ...\rangle | a \in \mathcal{A};\ o \in \mathcal{O}\}.$$

Given a policy π, the distribution of trajectories is fully characterized by:

$$P\{A_1, O_1, A_2, O_2, ..., A_n, O_n\} = \prod_{i=1}^{n} \mu_0(O_i | A_1, ...A_i, O_1, ..., O_{i-1}) \times \prod_{i=1}^{n} \pi(A_i | A_1, ...A_{i-1}, O_1, ..., O_{i-1}).$$

Thus, once the policy is fixed, a traditional dynamical system $(\Omega, \mathcal{F}, T, P)$ is induced, which has no actions, and whose observations are action–observation pairs of the original system.

Recall that a dynamical system is *asymptotically mean stationary* (AMS) if and only if

$$\forall F \in \mathcal{F} : \bar{P}(F) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P\{T^{-i} F\} \quad \text{exists;}$$

and $\bar{P}$ is called *stationary mean*. If the dynamical system is AMS, then for any bounded function f:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) = E(f|\mathcal{G}) \text{ a.s.},$$

where $\mathcal{G}$ is the σ-field of invariant sets [3]. This is a very useful property in the context of learning from data, because it guarantees the convergence of any empirical estimate to the expected value conditioned on the component from which the system started. From a reinforcement learning perspective, if a policy gives rise to an AMS system, averaging collected rewards converges to a conditional expected reward[1]. If *all* policies of interest (when enacted in the same environment) give rise to AMS systems, then the averages of collected rewards converge for any of these policies. Clearly, this is a necessary requirement for average reward reinforcement learning.

Policies with infinite memory can easily violate the AMS assumption. For example, imagine a two-room environment and a policy that stays in one room for some time, then changes the room and stays exponentially longer, then goes back to the first room and stays there exponentially longer than in the previous room, etc. As a result, the empirical estimate of the probability of staying in room one will always oscillate; therefore the process is not AMS. Since we have control over the policy class, we will rule out such examples, by considering only policies with finite memory. The other compelling reason is that we have, simply, no means to implement an infinite memory policy. Hence, from now on we will be interested only in the behavior of dynamical systems induced by finite memory policies. A major question of interest is whether one can characterize easily the class of controllable systems in which all finite memory policies induce AMS systems. Moreover, can one characterize a class of policies that induce AMS systems for any controllable process? As we will see in the next section, the answers are non-trivial.

[1] Throughout the paper we assume that the reward function is measurable.

3 Finite memory policies inducing non-AMS systems

From the theory of Markov chains, any chain with a finite number of states creates an AMS process (for more details see Section 4). Hence, we will focus on the case of controllable systems with a (countable) infinite state space, and we analyze several examples in which finite-memory policies lead to non-AMS dynamical systems. In order to facilitate the proofs for the following examples we will show that any transient Markov chain induces a non-AMS process.

Proposition 1. *Let $(\Omega, \mathcal{F}, P)$ be a stochastic process induced by a Markov chain with a countable state space $\mathcal{S}$. Then, for any $F \in \mathcal{F}$ and any $k \in \mathrm{N}$ there exists an increasing sequence of finite sets of states $\{C_i \subset \mathcal{S}\}_{i\in\mathrm{N}}$ such that*

$$P\Big\{F \cap \{\forall_{1\le j\le k} : X_j \in C_i\}\Big\} \to P\{F\} \text{ as } i \to \infty, \quad (1)$$

where $X_j \in \mathcal{S}$ represents a state visited at time j.

Proof. For convenience, we identify the state space $\mathcal{S}$ with a set of elements from N. Define $\forall i \in \mathrm{N} : C_i = \{1, ..., i\}$, so that $\{C_i\}_{i\in\mathrm{N}}$ is an increasing sequence of finite sets, and $\cup_{i=1}^{\infty} C_i = \mathcal{S}$.

Therefore, for any fixed $k \in \mathrm{N}$,

$$\Big\{P\big(F \cap \{\forall_{1\le j\le k} : X_j \in \cup_{i=1}^{n} C_i\}\big)\Big\}_{n\in\mathrm{N}}$$

is an increasing sequence with the limit

$$P\big(F \cap \{\forall_{1\le j\le k} : X_j \in \cup_{i=1}^{\infty} C_i\}\big) = P\big(F \cap \{\forall_{1\le j\le k} : X_j \in \mathcal{S}\}\big) = P(F). \quad \square$$

Proposition 2. *A stochastic process induced by a transient Markov chain with a countably infinite state space is not asymptotically mean stationary.*

Proof. For a finite set of states $S \subset \mathcal{S}$ we define an event $\bar{S}^{(n)} = \cup_{i=n}^{\infty}\{X_i \in S\}$ which represents visiting states in S at least once after n transitions. The transient chain is the one satisfying

$$\forall s \in \mathcal{S} : P\{X_i = s \text{ infinitely often } | X_0 = s\} = 0.$$

Therefore we get

$$P\{(X_k \in S) \cap \bar{S}^{(n+k)}\} \to 0 \text{ as } n \to \infty \qquad (2)$$

for any $k \in \mathrm{N}$.
Pick finite $S_0 \in \mathcal{S}$ such that $P\{X_1 \in S_0\} = p_0 > 0$.
Waiting step: Due to (2), for any $\epsilon_0 > 0$ there exists N_0 such that $P\{(X_1 \in S_0) \cap \bar{S}_0^{(N_0)}\} < \epsilon_0$, implying that

$$P\{F_0\} > p_0 - \epsilon_0,$$

where $F_0 \triangleq (X_1 \in S_0) \cap \big[\bar{S}_0^{(N_0)}\big]^c$.
Approximation step: Due to (1), for any $\beta_0 > 0$ we can find a finite set $S_1 \in \mathcal{S}$ such that $S_0 \subset S_1$ and

$$P\{G_0\} > P\{F_0\} - \beta_0,$$

where $G_0 \triangleq F_0 \cap \{X_1, ..., X_{3N_0} \in S_1\}$.
We construct the entities S_{i+1}, F_i, G_i in the following inductive fashion. (*Waiting step*): Due to (2), for any $\epsilon_i > 0$ there exists N_i such that

$$P\{F_i\} > P\{G_{i-1}\} - \epsilon_i,$$

where $F_i \triangleq G_{i-1} \cap \big[\bar{S}_i^{(N_i)}\big]^c$. (*Approximation step*): Due to (1), for any $\beta_i > 0$ we can find a finite set $S_{i+1} \in \mathcal{S}$ such that $S_i \subset S_{i+1}$ and

$$P\{G_i\} > P\{F_i\} - \beta_i,$$

where $G_i \triangleq F_i \cap \{X_1, ..., X_{3N_i} \in S_{i+1}\}$.
Now, denote $G = \cap_{i=0}^{\infty} G_i$. Since we have $P\{G\} > p_0 - \sum_{i=0}^{\infty}(\epsilon_i + \beta_i)$, we can choose $\{\epsilon_i\}_{i\in\mathrm{N}}$ and $\{\beta_i\}_{i\in\mathrm{N}}$ so that $P\{G\} > 0$.
To complete the proof, we set

$$f = 1\big[\cup_{i=1}^{\infty}(S_{2i} \setminus S_{2i-1})\big]$$

and observe that $\forall x = \langle x_1, x_2, ...\rangle \in G$:

$$\frac{1}{t}\sum_{n=0}^{t-1} f(x_n) \le \frac{1}{3} \quad \forall t \in \{3N_{2j} | j \in \mathrm{N}\}$$

$$\frac{1}{t}\sum_{n=0}^{t-1} f(x_n) \ge \frac{2}{3} \quad \forall t \in \{3N_{2j+1} | j \in \mathrm{N}\}$$

So the function f does not satisfy the ergodic property, hence the process is not asymptotically mean stationary. $\square$

3.1 Infinite queue

Consider the countably infinite Markov chain depicted in the Fig. 1, and suppose each state leads to an observation from a finite set. The long term behavior of this chain differs drastically for different values of $p > 0$. If $p < \frac{1}{2}$ then the chain is ergodic with unique limiting stationary distribution [2], so the system is AMS. If $p > \frac{1}{2}$ then each state in this chain is transient (intuitively, the chain drifts right; see e.g. [1] for an analysis). Hence Proposition 2 is applicable and the system is not AMS. To change the example above into a controllable system, suppose there are two actions, a-arrive and d-depart, each succeeding with probability 1. If a is chosen, the size of the queue increases by one; if d is chosen the size of the queue decreases by one. Then, a policy that chooses a with probability 0.4 induces an AMS system, while a policy that chooses d with probability 0.4 induces a non-AMS system.

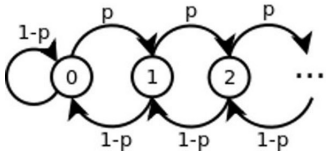

Figure 1: The infinite queue (the initial state is 0), also called the random walk on non-negative integers.

3.2 Infinite binary graph

The previous example considered two memoryless stochastic policies that yield different types of dynamical systems. However, that example has a continuous range of policies that induce AMS processes: all policies that choose action a with some probability in the range $[0, \frac{1}{2})$. This leads to the following question: if two policies lead to AMS systems, does their convex combination also lead to an AMS system? This property holds in the previous example, but we will now see that it is not always true.

Consider a controllable dynamical system induced by an infinite binary graph depicted in Fig. 2. The actions are l - descend left and r - descend right. Suppose that these actions always succeed. Following arguments similar to the previous example, it is possible to show that the policies π_1 - choose l w.p. 0.8, and π_2 - choose r w.p. 0.8, create two AMS systems. However, the policy π_3 - choose l w.p. 0.5, will result in a non-AMS dynamical system.

To see this, first observe that the probability of crossing the left $\frac{1}{4}$ border under policy π_1 decreases as we descend, resulting in $P\{X = 1 \ i.o.\} = 0$ (this can be shown using any concentration bound, like Chebyshev or Hoeffding). Therefore, the distribution $\hat{P}$ defined by

$$\forall i : \hat{P}\{X_i\} = \begin{cases} 1 & \text{for } X_i = 0 \\ 0 & \text{for } X_i = 1 \end{cases}$$

is stationary and asymptotically dominates P, implying

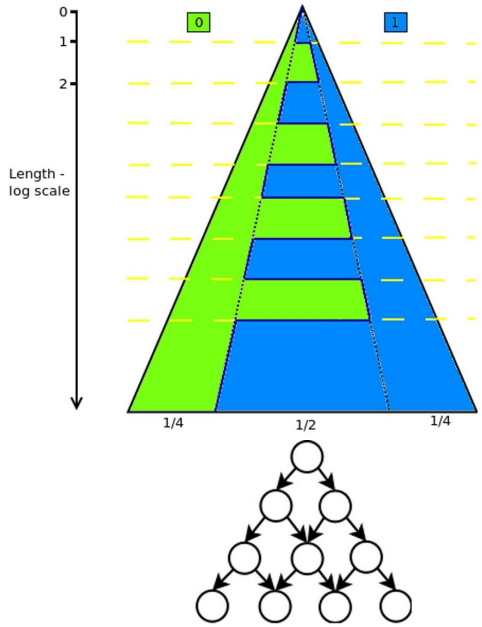

Figure 2: The infinite binary graph (top). The initial state is the root. Each node has two children, but most of the children are shared (the bottom image shows the first few layers). The nodes that fall in the green region are labeled 0 and those in the blue region are labeled 1. Note that the X axis (the depth of the tree) is displayed on a log-scale, meaning that the green/blue regions in the middle grow exponentially as we go deeper.

that the resulting process is AMS (For details on asymptotic dominance and its link to the AMS property, see [2]). Similarly, while following the policy π_2, the probability of crossing the right $\frac{1}{4}$ border decreases, resulting in an AMS process. Thus both π_1 and π_2 lead to an AMS system. However, by following policy π_3 the process will eventually stay in the middle between two borders. Hence, the running average of the sequence of observations will not converge, and we conclude that the process is not AMS.

3.3 Fully connected infinite queue

Both previous examples suggest that a weak connectivity of the underlying Markov chain is a possible cause for such a phenomenon: in both cases, the number of edges is of the order of the number of nodes, so the system can always reach a state that is "arbitrarily far away" from any fixed state. In this section we present a fully connected system in which finite policies still may not induce AMS systems. It is an extension of the queue example, in which all the states are connected and there is a positive probability of going from any state to any other state (see Fig. 3).

The probability that the queue grows is always p since $\sum_{i=0}^{\infty} ap^i = a\frac{1}{1-p} = p$ where $a \triangleq p(1-p)$. For $p < \frac{1}{2}$ it is easy to show that this Markov chain is positive recurrent[2] and therefore AMS [3]. For $p > \frac{1}{2}$, using similar tools one can show that the chain is transient, therefore Proposition 2 is again applicable.

4 The AMS property and predictive representations

The previous section suggests that it is difficult to establish a priori what policies induce AMS processes, without introducing some other restrictions on the original controllable system. Therefore, we proceed by considering only controllable systems that induce AMS processes for any finite memory policy.

Definition 1. *A controllable dynamical system is called* ***absolutely asymptotically mean stationary*** *if the (uncontrolled) dynamical system induced by any finite memory policy is AMS.*

Thus, a controllable system that is absolutely AMS can be used to evaluate any quantity of interest induced by any finite memory policy. In particular, if the quantity of interest is the reward, then according to Section 2, the time averages of the reward function R induced by a policy π will converge to $E_\pi(R|\mathcal{G})$. To obtain $E_\pi(R)$ instead, which is required for comparison of policies in the average reward RL case, additional conditions should be enforced. Those are discussed in the next section.

In this section, we examine the question of how restrictive the absolute AMS assumption is, specifically by looking at existing models of dynamical systems (especially predictive state representations). We find that many existing models are absolutely AMS, but there are absolutely AMS systems that cannot be modeled by any such representations.

4.1 Existing absolutely AMS models

Finite Markov chains and finite Hidden Markov Models were shown to be AMS by Kieffer and Rahe [7]. At the same time, every finite *Markov decision process* (MDP) and *partially observable Markov decision process* (POMDP) [8] induce a finite Markov chain and a finite HMM respectively, once the policy is fixed. Therefore, finite-state MDP and POMDP are absolutely AMS according to Definition 1. More recently, a new modeling framework, called *predictive state representations* (PSR) was proposed by Littman et. al. [9]. Most of the work in this framework focuses on so-called linear PSR models, which can represent a strictly larger set of systems than POMDPs

[2] First, observe that $E(J_i|X_i = k) < -\epsilon$ for almost all k and $\epsilon > 0$ small enough, where J_i is the jump at time i ($J_i = X_{i+1} - X_i$), $\epsilon \in (0, 1)$. Then, using Chebyshev's bound we obtain that $P\{X > k \text{ almost always}\} = 0$ for almost all k. From here it is straightforward to show positive recurrence for all states.

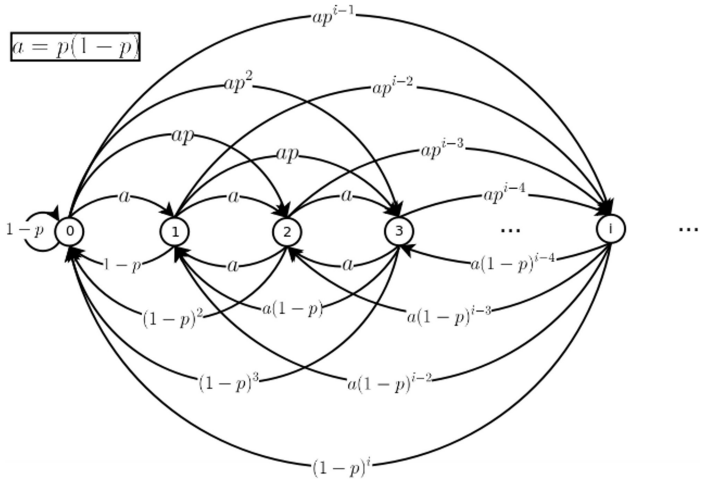

Figure 3: The fully connected infinite queue (the initial state is 0).

[9]. Singh et. al. [10] showed that linear PSR models without actions are equivalent in their representational power to *observable operator models* (OOM) developed by Jaeger [11]. Finally, Faigle and Schoenhuth [4] showed that OOMs can only model AMS processes. Combining this result with the work of Wiewiora [12], who showed that a linear PSR coupled with a finite memory policy can still be represented by a linear PSR without actions, we conclude that all controllable systems that can be represented by a linear PSR are absolutely AMS.

4.2 Non-linear predictive state representations and AMS systems

As it was mentioned above, most of the literature on predictive state representations focuses on linear PSRs. To the best of our knowledge, the only exception is the work of Rudary and Singh [13], which gives an example of a system that can be represented by a nonlinear PSR with dimension exponentially smaller than that of a linear PSR. However, it was not known whether the representational power of nonlinear PSR is larger than that of linear PSR, until Faigle and Schoenhuth [4] and Schoenhuth and Jaeger [5] provided counterexamples that answer this question as a side effect. Both examples consider AMS systems (in fact, the example from [4] is stationary) that cannot be represented by a linear PSR. Yet, these systems can be represented by a nonlinear PSR. This observation has not been reported earlier in the literature. For completeness, we will restate the example from Schoenhuth and Jaeger [5] with a slight modification of the proof and show that the system is representable by a nonlinear PSR.

Nonlinear PSR Example

Consider an independent but nonstationary process (without actions) $\{O_t, t \in \mathrm{N}\}$ defined on the observation space $\mathcal{O} = \{a, b\}$ by the following distribution at time t:

$$P\{O_t = a\} = \alpha^{t+1}, \quad P\{O_t = b\} = 1 - \alpha^{t+1},$$

for some $\alpha \in (0, 1)$.
We will show that this process cannot be represented by a linear PSR, but it is AMS and can be represented by a nonlinear PSR.

Proof. First, consider a fragment of the system dynamics matrix of this process presented in Figure 4. One can see that a column corresponding to a test of size m cannot be represented as a linear combination of columns of shorter tests. Therefore the system dynamics matrix has infinite rank, hence this system cannot be represented by a linear PSR [10].
Now, consider a distribution

$$\hat{P}\{s\} = \begin{cases} 1 & \text{if } s = b^{|s|} = b \cdots b \\ 0 & \text{otherwise} \end{cases}.$$

Note that this is a stationary distribution that asymptotically dominates P, so the system is AMS.
Finally, denote $p(h) = P\{a|h\}$ for a history $h \in \mathcal{O}^*$. Note that for any test $t \in \mathcal{O}^*$, $P\{t|h\}$ can be computed from $p(h)$ and α. For example, $P\{abab|h\} = p(h) \cdot [1 - \alpha p(h)] \cdot \alpha^2 p(h) \cdot [1 - \alpha^3 p(h)]$. Also, we can update $p(h)$ on each time step using the equation $p(hx) = \alpha p(h), x \in \mathcal{O}$. Hence we constructed a nonlinear PSR that models the system above. □

Finally, we would like to highlight, with an example, the fact that nonlinear PSR systems and AMS systems are not comparable (neither is more general).

Non-AMS nonlinear PSR example

Consider an independent nonstationary process $\{O_t, t \in \mathrm{N}\}$ defined on the observation space $\mathcal{O} = \{a, b, c\}$ with the

	a	b	aa	ab	...	aaa	...	$\overbrace{a\cdots a}^{m}$
0	α	$1-\alpha$	α^3	$\alpha(1-\alpha^2)$	...	α^6	...	$\alpha^{\frac{m(m+1)}{2}}$
1	α^2	$1-\alpha^2$	α^5	$\alpha^2(1-\alpha^3)$	...	α^9	...	$\alpha^{\frac{m(m+3)}{2}}$
...	...	...	...	...	...	...	...	...
k	α^k	$1-\alpha^k$	α^{2k+3}	$\alpha^{k+1}(1-\alpha^{k+2})$	...	α^{3k+6}	...	$\alpha^{\frac{m(m+2k+1)}{2}}$

Figure 4: The fragment of the system dynamics matrix. The columns represent the tests and the rows represent lengths of the histories (the prediction does not depend on the history in this case, only on its length). The entries represent conditional probabilities of tests given histories. For more details on the system dynamics matrix see [10].

following distribution:

$$\begin{cases} P\{O_t = a\} = \frac{1}{t+1} \\ P\{O_t = b\} = [1 - P\{O_t = a\}] \cdot 1_{\sin\log(t)\geq 0} \\ P\{O_t = c\} = [1 - P\{O_t = a\}] \cdot 1_{\sin\log(t)<0} \end{cases}.$$

Hence, $\frac{1}{n}\sum_{t=0}^{n-1} P\{O_t = b\}$ does not converge as $n \to \infty$ since $\sin\log(t)$ is a periodic function with exponentially increasing periods. However, this process can be represented by the following nonlinear PSR. The state of the PSR is defined by only one prediction $p(t) = P\{O_t = a\}$, and it is updated by $p(t+1) = \frac{1}{\frac{1}{p(t)}+1}$. The other one step predictions can be computed from $p(t)$ by:

$$\begin{cases} P\{O_t = b\} = [1 - P\{O_t = a\}] \cdot 1_{\sin\log(\frac{1}{p(t)}-1)\geq 0} \\ P\{O_t = c\} = 1 - P\{O_t = b\} - p(t) \end{cases}.$$

Since the process is independent, the above equations are enough to compute the prediction of a test of any length.

Non-PSR AMS System Example

Consider an independent non-stationary process $\{O_t, t \in \mathrm{N}\}$ defined on the observation space $\mathcal{O} = \{a, b\}$ with a following distribution:

$$\begin{cases} P\{O_t = a\} = 2^{-t} & \text{if } \exists k \in \mathrm{N} \text{ s.t. } t = 2^k \\ P\{O_t = a\} = 0 & \text{otherwise} \end{cases}.$$

The distribution $\hat{P}$ defined in the first example also asymptotically dominates P, therefore the process is AMS.
Now, suppose that there exists a nonlinear PSR representing this process. According to [9] there must be a set of core tests $Q = \{q_1, ..., q_n | q_i \in \mathcal{O}^*\}$ such that for any test $t \in \mathcal{O}^*$ there exists a projection $f_t : [0,1]^n \to [0,1]$ with the property:

$$\forall h \in \mathcal{O}^* : P\{t|h\} = f_t\Big(P\{Q|h\}\Big),$$

where $P\{Q|h\} = \Big(P\{q_1|h\}, ..., P\{q_n|h\}\Big)$ is a PSR state for a history h. Denote $k = \max_{1\leq i\leq n} |q_i|$ the length of the longest core test and assume WLOG that $k > 2$. Pick any history h_1 of length 2^k, and any history h_2 of length $2^{k+1} - (k+1)$. Denote $q^* = bb\cdots b$ the test of size $k+1$. Observe that $P\{Q|h_1\} = P\{Q|h_2\}$ but

$$P\{q^*|h_1\} = 1 \neq 1 - 2^{-(k+1)} = P\{q^*|h_2\},$$

implying that a projection f_{q^*} does not exist; therefore, this process cannot be modeled with a nonlinear PSR.

5 Average reward RL in absolutely AMS systems

As we have mentioned before, the absolute AMS condition guarantees convergence of time averages of the reward function R to a conditional expectation $E_\pi(R|\mathcal{G})$ where $\mathcal{G}$ is the σ-algebra of invariant events and π is the policy inducing the underlying stochastic process. It is not difficult to image a scenario when this conditional expectation is not constant for a given policy. For example, assume a two-state Markov chain and two policies π_1 and π_2, where following π_1 ensures always staying at the current state and following π_2 switches the state. We get that the time average of the reward for π_1 can converge to two different values, depending on the initial state. Thus, executing π_2 in between two evaluations of π_1 can make the evaluation process appear inconsistent. In fact, without any further assumptions, the only conclusion that we can reach is that the current policy "now" is same/better/worse than the previous policy "then". The goal of this section is to address this concern for absolutely AMS systems in general.

Denote by μ the probability kernel that completely describes the future dynamics of the underlying system. We will call μ the state of the system. Let $\mathcal{M}$ be a class of states, and all finite convex combinations thereof, that are reachable in finite time by following a uniformly random policy starting from an initial state μ_0[3]. Clearly, adding convex combinations of states does not ruin the fact that $\mathcal{M}$ is a collection of probability kernels. Let $\bar{\mathcal{M}}$ be the completion of $\mathcal{M}$ by including the limits of sequences of probability kernels in $\mathcal{M}$ (the Vitali–Hahn theorem ensures that the limits are probability kernels themselves). Note that the stationary mean of an AMS stochastic process generated from the state μ_0 and finite memory policy π can be

[3] I.e, system states that correspond to all finite histories with non-zero probability of occurring under such policy will be in $\mathcal{M}$.

decomposed into the corresponding *stationary mean kernels*, $\bar{\mu}$ and $\bar{\pi}$. Thus, the stationary mean kernels ($\bar{\mu}$) that correspond to these policies are in $\bar{\mathcal{M}}$.

Proposition 3. *If a policy π together with all states in $\bar{\mathcal{M}}$ generates a class of AMS and ergodic stochastic processes, then they have the same stationary mean kernels.*

Proof. Suppose that policy π together with states $\mu_1, \mu_2 \in \bar{\mathcal{M}}$ generate two stochastic processes with corresponding (different) stationary means $\bar{P}_1$ and $\bar{P}_2$. $\bar{P}_1$ and $\bar{P}_2$ cannot agree on $\mathcal{G}$, the σ-field of the invariant events[4]; therefore we can find two invariant sets I_1, I_2 such that

$$\begin{aligned} \bar{P}_1(I_1) &= \bar{P}_2(I_2) = 1 \text{ and} \\ \bar{P}_1(I_2) &= \bar{P}_2(I_1) = 0 \end{aligned}$$

since both processes are ergodic. Now consider $\mu = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2$. Then, $\bar{P} = \frac{1}{2}\bar{P}_1 + \frac{1}{2}\bar{P}_2$ is the stationary mean of the process generated by μ and π. Observe that the corresponding stationary mean kernel $\bar{\mu} \in \mathcal{M}$. However, this is a contradiction since $\bar{P}(I_1) = \bar{P}(I_2) = \frac{1}{2}$, implying that $\bar{P}$ is not ergodic. Therefore, $\bar{P}_1 = \bar{P}_2$, which implies equality in their stationary mean kernels. □

We now show the first main result of this section.

Theorem 4. *If the class of policies Π together with the states in $\bar{\mathcal{M}}$ generate AMS and ergodic processes, then these policies are comparable in the following sense: given a measurable bounded reward function R,*

$$\forall \pi \in \Pi : \ \frac{1}{t}\sum_{n=0}^{t-1} R(T^n x) \to E_{\bar{P}}(R) \text{ a.s. } \bar{P},$$

where $\bar{P}$ is the stationary mean of the process induced by any $\mu \in \mathcal{M}$ and the corresponding π.

Proof. Fix a policy $\pi \in \Pi$. According to Proposition 3, no matter what the initial system state μ is (as long as $\mu \in \bar{\mathcal{M}}$), the resulting process, P_μ, possesses the same stationary mean $\bar{P}$. The ergodicity of P_μ and Birkhoff's almost-sure ergodic theorem together imply that

$$\lim_{t\to\infty} \frac{1}{t}\sum_{n=0}^{t-1} R(T^n x) = E_{\bar{P}}(R) \text{ a.s. } \bar{P}, P_\mu. \qquad \square$$

Now, recall that $\bar{\mathcal{M}}$ contains all system states reachable in a finite time. Therefore, Theorem 4 guarantees that evaluating the same π again, after running a different policy for an arbitrary but finite time will lead to the same result.

To reassure the reader that the assumptions of Theorem 4 are reasonable, assume that our dynamical system is a POMDP with finite state space. If the underlying state space represents an irreducible Markov chain for some ϵ-soft ($\epsilon > 0$) n-th order Markov memory policy (for some finite n), then one can verify that the same holds for any other such policy. Hence, if we define Π to be the set of these policies and $\mathcal{M}$ to be all possible initial distributions over the underlying state space, the assumptions of Theorem 4 are satisfied.

Yet, from a practical perspective, there is one more component that needs attention. Although convergence is guaranteed, it is likely that more assumptions are needed if one wants to have a theoretically sound convergence test[5]. It is worth mentioning that if the underlying model of a system (e.g. linear PSR) is known, it is easy to estimate the rate of convergence to a stationary mean given any policy. This rate can be directly used to estimate the rate of convergence of reward averages. For the rest of the discussion we assume that such a convergence test, formally defined below, is provided.

Definition 2. *Let f be a measurable function. The convergence test $C_f(\epsilon, \delta, t, \omega)$ is a binary valued measurable function for any $t \in \mathrm{N}$, $\epsilon > 0$ and $\delta > 0$, such that with probability $1 - \delta$ ($\omega \in \Omega$):*

$$\begin{aligned} \forall t \in \mathrm{N} : & C_f(\epsilon, \delta, t, \omega) = 1 \implies \\ & \left| \frac{1}{t}\sum_{n=0}^{t-1} f(T^n\omega) - E(f|\mathcal{G})(\omega) \right| \le \epsilon \quad \textit{and} \\ & \forall k > t : C_f(\epsilon, \delta, k, \omega) = 1. \end{aligned}$$

Dynamical systems with reset Sometimes, the environment has the ability to *reset*, i.e. put itself in a state that does not depend on the previous history. Formally speaking, we can model this behavior by viewing it as a special action. Moreover, the effect of the reset, in terms of the controllable dynamical system, is that the next state is μ_0, i.e. the state from which the system has initially started. In this case, we show that other than the existence of a sound convergence test presented in Definition 2, no further assumptions are required to evaluate the policy at hand.

Theorem 5. *Given that the special reset action is available and the system possesses a convergence test C_R, any policy π that induces an AMS process can be evaluated at any time, and its value will be equal to $E_{\bar{P}}(R)$, where $\bar{P}$ is the stationary mean of the process induced by μ_0 and π, and R is a bounded measurable reward function.*

Proof: Due to the AMS assumption,

$$\lim_{t\to\infty} \frac{1}{t}\sum_{n=0}^{t-1} R(T^n\omega) = E_{\bar{P}}(R|\mathcal{G})(\omega) \text{ a.s. } \bar{P}.$$

Since every reset puts the system in a state that is independent of the past, after each reset we obtain an independent

[4] It is a classical result that stationary distributions are uniquely identified by their restriction on $\mathcal{G}$.

[5] Assumptions that guarantee some sort of convergence rate would be sufficient.

sample from the distribution $\bar{P}_R$ induced by the random variable $E_{\bar{P}}(R|\mathcal{G})$.

To formalize this, denote by P the probability induced by μ_0 and π. We will define a new probability space $(X, \mathcal{X}, Q)$ as a countable product space of $(\Omega, \mathcal{F}, P)$, i.e. $X = \times_{i=1}^{\infty}\Omega$, etc. We equip this space with two shift operators T_{in}, T_{out} defined $\forall x \in X$ as:

$$\begin{aligned} T_{in}(x) &= T_{in}(\omega_1, \omega_2, \omega_3, ...) = (T(\omega_1), \omega_2, \omega_3, ...), \\ T_{out}(x) &= T_{out}(\omega_1, \omega_2, \omega_3, ...) = (\omega_2, \omega_3, \omega_4, ...). \end{aligned}$$

One can verify that these are measurable. Also, with the slight abuse of notation we define $\forall x \in X$:

$$R(x) = R(\omega_1, \omega_2, \omega_3, ...) \triangleq R(\omega_1).$$

$$\text{Denote } \forall i \in \mathrm{N} : R_i^{(t)} \triangleq \frac{1}{t}\sum_{n=0}^{t-1} R \circ T_{in}^n \circ T_{out}^i,$$

$$\bar{R}_i \triangleq \lim_{t\to\infty} R_i^{(t)} \sim \bar{P}_R.$$

Observe that $\forall t_1, t_2, i \neq j : R_i^{(t_1)}$ is independent of $R_j^{(t_2)}$ and $\bar{R}_i$ is independent of $\bar{R}_j$. The strong law of large numbers guarantees that (a.s. Q)

$$\frac{1}{n}\sum_{i=1}^{n} \bar{R}_i \to \int_{\mathrm{R}} x\, d\bar{P}_R = E_{\bar{P}}(R) \text{ as } n \to \infty. \quad (3)$$

The convergence rate of (3) can be bounded by the Berry-Esseen theorem, for example, since the existence of finite moments for $\bar{R}_1$ is guaranteed due to the boundedness of R. Therefore, for any $0 < \gamma < 1$ and $\epsilon > 0$ we can calculate M such that, with probability $\geq \gamma$:

$$\left|\frac{1}{M}\sum_{i=1}^{M} \bar{R}_i - E_{\bar{P}}(R)\right| \leq \frac{\epsilon}{2}. \quad (4)$$

After the i-th reset (which is represented by a T_{out} shift) we can use the convergence test C_{R_i}[6] to obtain an estimate of $\bar{R}_i$. Now, choose $\{N_i\}_{1\leq i\leq M}$ such that

$$\forall 1 \leq i \leq M : C_{R_i}\left(\frac{\epsilon}{2}, 1-\gamma^{\frac{1}{M}}, N_i, \omega\right) = 1,$$

which implies, with probability $\geq [1-(1-\gamma^{\frac{1}{M}})]^M = \gamma$:

$$\left|\frac{1}{M}\sum_{i=1}^{M} R_i^{(N_i)} - \frac{1}{M}\sum_{i=1}^{M} \bar{R}_i\right| \leq \frac{\epsilon}{2}. \quad (5)$$

Combining 4 and 5 we have, with probability $\geq 2\gamma - 1$:

$$\begin{aligned} \left|\frac{1}{M}\sum_{i=1}^{M} R_i^{(N_i)} - E_{\bar{P}}(R)\right| &\leq \left|\frac{1}{M}\sum_{i=1}^{M} R_i^{(N_i)} - \frac{1}{M}\sum_{i=1}^{M} \bar{R}_i\right| \\ &+ \left|\frac{1}{M}\sum_{i=1}^{M} \bar{R}_i - E_{\bar{P}}(R)\right| \leq \epsilon. \end{aligned}$$

□

[6]Use the definition 2 w.r.t. the shift T_{in} and $\mathcal{G}$, which is the invariant σ-algebra with respect to the first sequence of $x \in X$.

Thus, Theorem 5 tells us that if we perform the reset and subsequent estimation of $E_{\bar{P}}(R|\mathcal{G})$ enough times, their average will provide an accurate estimate of $E_{\bar{P}}(R)$.

6 Conclusions and future work

We proposed the notion of absolutely AMS systems as an interesting class of controllable dynamical systems for which the computation of running averages for quantities of interest leads to meaningful conditional expectations. While typical finite-state models used in the literature are absolutely AMS, as we have shown, this property can easily be violated in infinite systems, even with finite action and observation sets. We have also shown examples of absolutely AMS systems that cannot be captured by existing types of models, even nonlinear PSRs; linear PSRs, however, are absolute AMS systems. Hence, this seems to be a rich and interesting class of systems to study. Figure 5 summarizes the relationships illustrated by the examples in the paper.

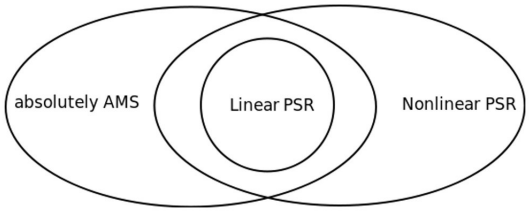

Figure 5: The relationship between categories of controllable systems. All inclusions are proper.

The absolutely AMS property is a necessary condition for the averages of collected rewards to converge, given any (reasonable) policy of interest. In Sec. 5 we discussed additional assumptions that, together with this property, guarantee the validity of average-reward policy evaluation in partially observable systems.

Much work remains to be done towards learning in absolutely AMS systems. The policy improvement step –an inherent part of any RL algorithm– is a main direction of future research. We are currently investigating properties of controllable dynamical systems that allow a gradient search in policy space. Finally, the strong relationship between the AMS property and information theory [2] should be explored in the controllable systems context.

Acknowledgments

This research was funded by ONR. We thank the anonymous reviewers as well as Andre Barreto for helpful comments.

References

[1] Lawer, F. G. (1995), *Introduction to stochastic processes*. Chapman and Hall, New York.

[2] Gray, R. M., Kieffer, J. C. (1980), *Asymptotically mean stationary measures*. Annals of Probability 8, pp. 962–973.

[3] Gray, R. M. (2009), *Probability, Random Processes, and Ergodic Properties*, 2nd edition. Springer Publishing Company.

[4] Faigle, U., Schoenhuth, A. (2007), *Asymptotic mean stationarity of sources with finite evolution dimension*. IEEE Trans. Inf. Theory 53, pp. 2342–2348.

[5] Schonhuth, A., Jaeger, H. (2009), *Characterization of ergodic hidden Markov sources*. IEEE Transactions on Information Theory 55, pp. 2107–2118.

[6] Mahadevan, S. (1996), *Average reward reinforcement learning: foundations, algorithms, and empirical results*. Machine Learning 22, pp. 159–196.

[7] Kieffer, J. C., Rahe, M. (1981), *Markov channels are asymptotically mean stationary*, SIAM J. Math. Anal. 12, pp. 293–305.

[8] Kaelbling, L. P., Littman, M. L., Cassandra, A. R. (1998), *Planning and acting in partially observable stochastic domains*. Artificial Intelligence Journal 101, pp. 99–134.

[9] Littman, M. L., Sutton R. S., Singh, S. (2002), *Predictive Representations of State*. Advances in Neural Information Processing Systems 14 (NIPS), pp. 1555–1561.

[10] Singh, S., James, M. R., Rudary, M. R. (2004), *Predictive State Representations: A New Theory for Modeling Dynamical Systems*. Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI), pp. 512–519.

[11] Jaeger, H. (2000), *Observable operator models for discrete stochastic time series*. Neural Computation 12, pp. 1371–1398.

[12] Wiewiora, E. (2007), *Modeling Probability Distributions with Predictive State Representations*. PhD thesis, The University of California at San Diego.

[13] Rudary, M. R., Singh, S. (2004), *A nonlinear predictive state representation*. Advances in Neural Information Processing Systems 16 (NIPS), pp. 79–798.

[14] Resnick, S., (2005), *Adventures in Stochastic Processes*. Birkhauser, Boston.

[15] Smallwood, R.D., Sondik, E.J. (1973), *The optimal control of partially observable Markov processes over a finite horizon*, Operations Research, pp. 1071–1088.

[16] Sutton, R.S., Barto, A.G. (1998), *Reinforcement learning: An introduction*, MIT Press, Cambridge, MA.