Average Reward Optimization Objective In Partially Observable Domains

Yuri Grinberg YGRINB@CS.MCGILL.CA
School of Computer Science, McGill University, Canada

Doina Precup DPRECUP@CS.MCGILL.CA
School of Computer Science, McGill University, Canada

Abstract

We consider the problem of average reward optimization in domains with partial observability, within the modeling framework of linear predictive state representations (PSRs) (Littman et al., 2001). The key to average-reward computation is to have a well-defined stationary behavior of a system, so the required averages can be computed. If, additionally, the stationary behavior varies smoothly with changes in policy parameters, average-reward control through policy search also becomes a possibility. In this paper, we show that PSRs have a well-behaved stationary distribution, which is a rational function of policy parameters. Based on this result, we define a related reward process particularly suitable for average reward optimization, and analyze its properties. We show that in such a predictive state reward process, the average reward is a rational function of the policy parameters, whose complexity depends on the dimension of the underlying linear PSR. This result suggests that average reward-based policy search methods can be effective when the dimension of the system is small, even when the system representation in the POMDP framework requires many hidden states. We provide illustrative examples of this type.

1. Introduction

Partial observability is prevalent in practical domains, in which one has to model systems based on noisy observations. In domains with partial observability and finite action and observation spaces, a popular framework for modeling and control is that of finite state *partially observable Markov decision processes* (POMDP) (Kaelbling et al., 1998). POMDPs generalize the well-known framework of Markov decision processes (MDPs); hence, they inherit many useful properties from MDPs. In particular, the results of Schweitzer (1968) imply that the stationary distribution of an ergodic Markov chain is a rational function of the changes in the transition matrix. Under mild conditions, this result can be applied immediately to finite state POMDPs: the stationary distribution of the Markov chain induced by some finite memory policy exists and is a rational function of policy parameters. This is important when one has to estimate expectations of different quantities (such as returns) with respect to the policy, since these expectations are taken with respect to the stationary behavior of the process. In particular, one can show that the average reward in finite state POMDPs is a linear function of the stationary distribution. Thus, knowing that the stationary distribution is robust to small changes in the policy implies that the estimates of different quantities based on the stationary distribution are robust as well.

About a decade ago, a new modeling framework for finite action and observation spaces was introduced, named (linear) *predictive state representations* (PSR) (Littman et al., 2001; Singh et al., 2004). PSRs alleviate model learning challenges encountered in finite state POMDP models, because their state representation is grounded in measurable quantities. More importantly, PSRs are capable of representing some POMDPs with smaller dimension than the number of hidden states. Moreover, there are systems which can be represented by a linear PSR but not a POMDP (Jaeger, 2000).

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

PSRs are formulated in terms of actions and observations, but rewards do not play as special role in the original framework, as they do in MDPs and POMDPs. The planning problem using linear PSRs has already been tackled by several authors, e.g. James et al. (2004); Boots et al. (2010); Izadi & Precup (2003) but little was said about how their specific assumptions on the reward function affect the combined reward-PSR process. The main motivation for our work is to provide a theoretical framework for reward processes based on PSRs, and analyze the behavior of the average reward in this framework. There are two practical implications of this idea. First, this would enable the development of learning and planning algorithms which are directly based on evaluating reward averages, rather than learning a full predictive representation first, and then using it to estimate policy values. In the MDP and POMDP framework, methods based on value functions and/or policy search are known to scale better to very large problems than methods based on a full model estimation. We would like to bring this advantage to linear PSRs as well. Second, defining an appropriate reward process based on PSRs enables an easier theoretical analysis of any control methods.

In Section 4, we present a (linear) predictive state reward process (PRP), which is built on top of a (linear) PSR and provides a suitable framework for average reward optimization. We analyze some of its properties and discuss its relationship to the POMDP framework. Our most important result connects the size of the linear PSR representation underlying the linear PRP to the complexity of the average reward as a function of policy parameters. We prove that a compact PSR representation will induce a reward process with a simple average reward function. We provide some examples in which the linear PRP representation is much smaller than the corresponding POMDP representation; hence, using the PRP for policy search, for example, could potentially be much easier and yield a solution faster.

The key ingredient in analyzing the PRP is to analyze the behavior of its stationary distribution as a function of the policy used, similar to the case of a POMDP. This stationary distribution is in fact the stationary distribution of the underlying linear PSR, which has been shown to exist (Faigle & Schonhuth, 2007) given a fixed policy. However, its form has not been stated before. In Section 3, we show that the stationary distribution induced by a finite memory policy in a linear PSR is a rational function of the policy parameters, as long as the process is ergodic, similar to a finite state POMDP framework. This result is a novel contribution in itself, and can be useful if a stability / perturbation analysis of a linear PSR is desired. Throughout the paper, we omit the proofs, for brevity; they are included in the appendix provided as supplementary material.

2. Background and notation

We consider systems with control having finite action and observation/percept spaces, denoted as $\mathcal{A}$ and $\mathcal{O}$ respectively, and rewards coming from a bounded set $\mathcal{R} \subset \mathbb{R}$. We define a reward process as a tuple $(\Omega, \mathcal{F}, \mathrm{T}, \mu_0)$, 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. In our setting, the elements of Ω can be viewed as infinite sequences of action-observation-reward triples, $\omega \in \Omega$:

$$\omega = \langle a_1, o_1, r_1, a_2, o_2, r_2, ... \rangle, a_i \in \mathcal{A}, o_i \in \mathcal{O}, r_i \in \mathcal{R}.$$

In this setting, T is defined by:

$$\mathrm{T}(\langle a_1, o_1, r_1, a_2, o_2, r_2, ... \rangle) = \langle a_2, o_2, r_2, a_3, o_3, r_3 ... \rangle,$$
$$\mathrm{T}^{-1}(\omega) = \{\omega' \mid \mathrm{T}(\omega') = \omega\}.$$

At each time step, the reward process waits for the agent to take an action and outputs an observation-reward pair. We are interested in the average reward setting, where the goal is to find a behavior (or policy) π maximizing $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} R_t$ where R_t is the random variable denoting the reward observed at time t, when following policy π. The average reward setting is in fact preferable to discounted rewards in partially observable systems when the dynamics are not known[1] (Singh et al., 1994). We consider the classical policy definition, in which the action choice depends (stochastically) only on actions and observations, but not rewards: $\forall k \in \mathbb{N}, \pi : \langle A, O \rangle^k \to [0,1]^{|\mathcal{A}|}$. Representing any arbitrary policy of this kind is infeasible, since it requires infinite memory. Therefore, only policies having a finite memory representation are considered. Such policies can be implemented through (stochastic) finite state controllers, whose parameters represent the transition probabilities between internal states of the policy, and the probabilities of actions conditioned on the policy state. We call this a *direct policy parametrization*. In particular, if the policy is memoryless and open-loop (i.e., it chooses actions with a fixed distribution regardless of the time step and history), its direct parametrization will simply be the vector representing the distribution over actions.

[1] See also Sutton (1998), last bullet, for details; this addresses the function approximation in MDPs, which is essentially equivalent to POMDPs.

Fixing a policy π generates a stochastic process whose elements are action-observation-reward triples. We denote the distribution of this process by P_π or P_θ where θ is the parametrization of π, and the expectation with respect to P_π by E_π or E_θ. P_π is *asymptotically mean stationary* (AMS) (Gray & Kieffer, 1980) if

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

$\bar{\mathrm{P}}_\pi$ is called *stationary mean*; let $\mathrm{E}_{\bar{\pi}}$ be the expectation with respect to $\bar{\mathrm{P}}_\pi$. The AMS property is a necessary and sufficient condition for the running averages of any bounded quantity to converge; in particular, it guarantees the existence of the average reward. For example, if the reward process is a finite state Markov decision process (MDP), the stationary mean is the stationary distribution of the Markov chain induced by policy π. Moreover, P_π is *ergodic* if

$$\forall G \in \mathcal{G} : \mathrm{P}_\pi(G) \in \{0, 1\},$$

where $\mathcal{G} \subset \mathcal{F}$ is a set of invariant events ($\mathrm{T}^{-1}G = G$). Thus, if P_π is AMS and ergodic,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} R_t = \mathrm{E}_{\bar{\pi}}(R) \quad a.s.\ (\mathrm{P}_\pi, \bar{\mathrm{P}}_\pi),$$

meaning that the average reward is constant almost surely (for more details see Gray (2009)). In a finite state MDP, the ergodicity of the process corresponds to the Markov chain induced by π being irreducible.

2.1. Predictive State Representations

The idea of a predictive state of a system is rooted in the following observation: knowing the distribution of any finite sequence of observations given any sequence of actions describes completely and accurately the future dynamics of that system. In the setting we consider, the action and observation spaces are finite, meaning that one can represent the predictive state of the system with a countably infinite vector enumerating probabilities of finite sequences of observations given sequences of actions. We will refer to this vector as the *state* of the system. In general, it might not be possible to represent this vector concisely.

In Littman et al. (2001), a new representation for the action-observation controllable process (defined without rewards) was proposed, called predictive state representation (PSR). It captures processes in which predictions of a finite number of sequences are enough to compute the prediction of any other sequence. Formally, let $h, t \in \langle A \times O \rangle^*$ be sequences of action-observation pairs of finite length, where h is a *history*, i.e., a sequence observed until now, and t is a sequence that might occur in the future, called *test*. Let $\mathcal{Q} = \{q_1, ..., q_n\}$ be a set of tests, called *core tests*, and $\mathbf{p}(h) \triangleq [\mathrm{P}_{psr}(q_1|h), ..., \mathrm{P}_{psr}(q_n|h)]^\top$ be the prediction vector where $\mathrm{P}_{psr}(q|h) \triangleq \frac{\mu_0(hq)}{\mu_0(h)}$ represents the probability of observing percepts from q given that the agent has seen history h and is going to take actions from q. Then, $\mathbf{p}(h)$ is a predictive state representation of dimension n if and only if

$$\mathrm{P}_{psr}(t|h) = f_t(\mathbf{p}(h)), \tag{1}$$

for any test t and history h, where $f_t : [0,1]^n \to [0,1]$ is any function independent of history. If f_t is linear for all t, then the representation is called linear PSR, and

$$\mathrm{P}_{psr}(t|h) = \mathbf{m}_t^\top \mathbf{p}(h),$$

where the function f_t is replaced by a linear projection $\mathbf{m}_t$. It has been shown (Singh et al., 2004; Wiewiora, 2007) that a linear PSR possesses a finite parametrization of the form

$$\{\mathbf{m}_* \in \mathbb{R}^n, \{\mathbf{M}_{ao} \in \mathbb{R}^{n\times n}\}_{a\in\mathcal{A}, o\in\mathcal{O}}, \mathbf{p}_0 \in \mathbb{R}^n\},$$

satisfying the following properties:

1. $\mathbf{m}_*^\top \mathbf{p}_0 = 1$,
2. $\forall a \in \mathcal{A} : \left[\sum_{o\in\mathcal{O}} \mathbf{M}_{ao}\right] \mathbf{m}_* = \mathbf{m}_*$,
3. $\mathrm{P}_{psr}(a_1, o_1, ..., a_k, o_k) = \mathbf{m}_*^\top \mathbf{M}_{a_k,o_k}^\top \cdots \mathbf{M}_{a_1,o_1}^\top \mathbf{p}_0$,

where $\mathbf{p}_0$ is the starting state of the linear PSR. Using property 3 one can verify that

$$\mathbf{p}(hao) \triangleq \frac{\mathbf{M}_{ao}^\top \mathbf{p}(h)}{\mathbf{m}_*^\top \mathbf{M}_{ao}^\top \mathbf{p}(h)}$$

represents the PSR state after observing ao, given the previous PSR state $\mathbf{p}(h)$. Predictions of future sequences can be calculated using the new PSR state and the rest of the linear PSR parameters only. For better readability, from now on we drop the prefix "linear" from linear PSR and explicitly state "non-linear" when referring to the general PSR framework.

Any finite memory policy π in a PSR induces an action-observation stochastic process, which can also be represented by a PSR, possibly of a larger dimension (Wiewiora, 2007). This process is often called PSR without control. In PSRs without control, property 2 reduces to $\left[\sum_{ao\in\mathcal{A}\times\mathcal{O}} \mathbf{M}_{ao}\right] \mathbf{m}_* = \mathbf{m}_*$, and in property 3, "$\mathrm{P}_{psr}$" is replaced with "$\mathrm{P}_\pi$", where π is the policy at hand. A PSR without control can be represented in at least two more frameworks: *observable operator models* (OOM) (Jaeger, 2000) and *transformed PSRs* (TPSR) (Boots et al., 2010; Rosencrantz

et al., 2004). These frameworks can be seen as different parametrizations of the same process, while keeping the same dimension to represent its state (Singh et al., 2004; Boots et al., 2010; James, 2005). Therefore, one can use any of these frameworks equivalently to analyze properties of the system. In particular, our analysis uses results obtained under the OOM framework (Faigle & Schonhuth, 2007).

Finally, the key property of these frameworks is that certain systems can be represented more compactly compared to the classical finite state hidden Markov model (HMM) representation (Jaeger, 2000). Moreover, some of these systems might not be representable using finite-state HMMs at all. This is the reason why we focus on this type of representation.

3. The stationary distribution of a PSR

In this section, we investigate how the stationary distribution of a PSR behaves as the policy changes, where by stationary distribution we mean the stationary state of the system represented by a PSR. For this purpose, we focus on the stationary distribution of a PSR without control induced by some policy, and see how changes in this policy affect this distribution.
Let $\boldsymbol{\rho}_\pi$ be the stationary distribution of a PSR without control induced by policy π[2]. Let $\forall t \in \mathbb{N} : \mathbf{p}_t$ be the expected state of the PSR at time t. Assuming that $\mathbf{p}_0 = \boldsymbol{\rho}_\pi$, we have, by definition:

$$\forall t \in \mathbb{N} : \mathbf{p}_t = \boldsymbol{\rho}_\pi.$$

Now, let

$$\mathbf{M}_* \triangleq \sum_{ao \in \mathcal{A}\times\mathcal{O}} \mathbf{M}_{ao}$$

be the PSR evolution matrix. This is called evolution matrix due to the following, which can be easily seen from property 3:

$$\forall t \in \mathbb{N} : \mathbf{p}_t = \mathbf{M}_*^\top \mathbf{p}_{t-1}.$$

The matrix $\mathbf{M}_*$ will in fact define the stationary distribution of the PSR. Under the appropriate basis, it satisfies the conditions of Lemma 1 (Faigle & Schonhuth, 2007), and it can be seen as a generalized Markov transition matrix. More precisely, its spectral properties are the same as those of a usual Markov transition matrix, and the rows sum to 1, but some of the values might be negative (see Faigle & Schonhuth (2007) Section IV). The following Lemma shows that several important properties of Markov transition matrices generalize to evolution matrices as well (represented under the appropriate basis).

[2]Note that $\boldsymbol{\rho}_\pi$ represents $\bar{\mathrm{P}}_\pi$, the stationary mean of the action-observation stochastic process.

Lemma 1. *Let $\mathbf{E} \in \mathbb{R}^{n\times n}$ be such that*

$$\mathbf{E}^\infty \triangleq \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbf{E}^t$$

exists, and the rows of $\mathbf{E}$ sum to 1. Also assume that $(\mathbf{E} - \mathbf{I})$ has rank $n-1$, where $\mathbf{I}$ is an identity matrix of appropriate size. Then the following holds:

1. $\mathbf{E}\mathbf{E}^\infty = \mathbf{E}^\infty\mathbf{E} = \mathbf{E}^\infty\mathbf{E}^\infty = \mathbf{E}^\infty$.
2. $(\mathbf{E} - \mathbf{E}^\infty)^n = \mathbf{E}^n - \mathbf{E}^\infty$.
3. $\mathbf{E}^\infty = \mathbf{1}\boldsymbol{\rho}^\top$, *where* $\mathbf{1}$ *is a column vector of ones;* $\boldsymbol{\rho}$ *is the unique column vector satisfying* $\boldsymbol{\rho}^\top\mathbf{E} = \boldsymbol{\rho}^\top$, $\boldsymbol{\rho}^\top\mathbf{1} = 1$.
4. *Let* $\mathbf{Z} \triangleq [\mathbf{I} - (\mathbf{E} - \mathbf{E}^\infty)]^{-1}$, *then* $\mathbf{Z}$ *is well defined and*
$$\mathbf{Z} = \mathbf{I} + \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n-1} \sum_{k=1}^{t} (\mathbf{E}^k - \mathbf{E}^\infty). \qquad (2)$$
5. $\boldsymbol{\rho}^\top\mathbf{Z} = \boldsymbol{\rho}^\top$.

The next theorem extends Theorem 1 of Schweitzer (1968) to the case in which the two matrices at hand are evolution matrices.

Theorem 2. *Let $\mathbf{E}_1, \mathbf{E}_2$ be as in Lemma 1, and let $\mathbf{E}_1^\infty, \mathbf{E}_2^\infty, \boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \mathbf{Z}_1, \mathbf{Z}_2$ be the quantities defined as in Lemma 1 with respect to $\mathbf{E}_1$ and $\mathbf{E}_2$ respectively. Then:*

$$\mathbf{H}_{1\to2} \triangleq [\mathbf{I} - (\mathbf{E}_2 - \mathbf{E}_1)\mathbf{Z}_1]^{-1} \qquad (3)$$

exists and is given by $\mathbf{H}_{1\to2} = \mathbf{Z}_1^{-1}[\mathbf{I} - \mathbf{E}_1^\infty + \mathbf{E}_2^\infty]\mathbf{Z}_2$. Moreover,

$$\boldsymbol{\rho}_1^\top\mathbf{H}_{1\to2} = \boldsymbol{\rho}_2^\top.$$

Theorem 2 lies in the heart of our work, since it establishes an exact relationship between the eigenvectors corresponding to eigenvalue 1 of two evolution matrices. Similarly to the theory of Markov chains, these eigenvectors represent the stationary distributions of the corresponding PSRs. Specifically, let $\mathbf{E}_1, \mathbf{Z}_1$ and $\boldsymbol{\rho}_1$ be fixed, while $\mathbf{E}_2$ varies. Theorem 2 shows that $\boldsymbol{\rho}_2$ is a rational function of the entries in $\mathbf{E}_2$. This is due to the fact that the matrix inverse can be calculated analytically through Cramer's rule.

The next theorem is the main result of this section.

Theorem 3. *Let an action-observation controllable process $\mathcal{S}$ be representable by a finite dimensional linear PSR, and let $\theta \in \Theta$ be a direct parametrization of a stochastic finite memory policy. If the stochastic process generated from $\mathcal{S}$ induced by any policy $\theta \in \Theta$ is ergodic, then the stationary state of $\mathcal{S}$ is a rational function of θ.*

The proof (included in the supplementary material) shows how to construct the parametrized evolution matrix in an appropriate basis, given a PSR without control parametrized by a policy. The construction guarantees that the entries of the evolution matrix are rational functions of the policy parameters. We show that this construction is well defined, and then invoke Theorem 2 to complete the proof.

As mentioned, Theorem 3 requires the ergodicity condition for any policy under consideration. A violation of this condition implies the existence of a policy for which the initial state of the system affects its asymptotic behavior. In the theory of Markov chains, for example, this translates into the irreducibility condition on the chains induced by any policy from the set of policies considered. In particular, the irreducibility condition(s) is either satisfied for all strictly stochastic policies or not satisfied for any policy. Hence, this assumption is considered standard in the policy search literature (Baxter & Bartlett, 2001; Peters & Schaal, 2008).

4. Predictive-state reward processes

In this section we define a (not necessarily linear) PSR-based reward process. The goal is to construct a reward process that inherits some of the useful properties of the PSR framework on one hand, while being general enough and suitable for average reward optimization on the other hand. In particular, we let the reward be continuous. Although we keep the definition quite general, we are interested in systems that are based on the linear PSR framework, and only this setting is treated afterwards.

Let $h^{(t)}$ represent the sequence of action-observation-reward triplets of length t, and $\bar{h}^{(t)}$ represent the corresponding sequence of action-observation pairs only.

Definition 1. *A predictive–state reward process (PRP) is a reward process whose action-observation component can be represented by a (possibly non-linear) PSR with some state vector $\mathbf{s} \in \mathbb{R}^n$, and whose reward R_t at time t satisfies $\forall t \in \mathbb{N}$:*

$$\mathrm{E}[R_t|h^{(t-1)}, a_t, o_t] = \mathrm{E}[R|\mathbf{s}(\bar{h}^{(t)}), a_t] \triangleq f_{R,a_t}[\mathbf{s}(\bar{h}^{(t)})],$$

where $\mathbf{s}(\bar{h}^{(t)})$ is the PSR state after observing $\bar{h}^{(t)}$, and $\{f_{R,a}\}_{a\in\mathcal{A}}$ are functions independent of $\bar{h}^{(t)}$.
A linear PRP is a PRP whose action-observation component can be represented by a linear PSR, and in which f_{R,a_t} are linear functions:

$$\forall a_t \in \mathcal{A} : f_{R,a_t}[\mathbf{s}(\bar{h}^{(t)})] = \mathbf{r}_{a_t}^\top \mathbf{s}(\bar{h}^{(t)}),$$

where $\forall a \in \mathcal{A} : \mathbf{r}_a \in \mathbb{R}^n$. The dimension of the PRP is defined to be the dimension of the underlying PSR.

Note that Definition 1 in fact formalizes a setting under which most of the planning problem in linear PSRs has been tackled. For example, in Boots et al. (2010) the reward function is obtained by applying a linear regression from PSR states to the observed rewards. James et al. (2004) assumes that the rewards are discrete and incorporates them directly into the observation vector. Izadi & Precup (2003) defines the reward signal itself to be a linear function of the state of a linear PSR. Thus, current PSR–based planning algorithms either implicitly assume that the underlying reward process is a linear PRP, or have more restrictive explicit assumptions.

The rest of this section is devoted to the analysis of the behavior of the average reward as a function of a policy in systems represented by a linear PRP. Two questions arise in this context. First, it is not clear whether the average reward is well defined for any finite memory policy. It has been shown that systems represented by a linear PSR are AMS for any finite memory policy (Grinberg & Precup, 2012). This fact guarantees that averages of different observable quantities from action–observation pairs are well defined, but not necessarily the average reward itself. Second, when/if the average reward is well defined, we want to characterize its behavior as a function of the policy parameters. Both questions are addressed in Theorem 4. We show that under mild conditions the average reward exists and is linear in a quantity derived from the stationary distribution of the underlying PSR.

Theorem 4. *Let a n-dimensional linear PRP be ergodic for a collection of stochastic finite state policies of size m given by direct parametrization $\theta \in \Theta$. Let $\boldsymbol{\rho}_\theta$ be the stationary state of the reward process without control induced by policy θ. Denote by $\boldsymbol{\rho}_\theta^{PRP} \in \mathbb{R}^n$ and $\boldsymbol{\rho}_\theta^{Pol} \in \mathbb{R}^m$ the stationary states of the PRP and the policy correspondingly, obtained from $\boldsymbol{\rho}_\theta$. Also, let $\forall a \in \mathcal{A} : \boldsymbol{\rho}_\theta^{PRP}(a) \in \mathbb{R}^n$ be the PRP state obtained from starting the PRP at state $\boldsymbol{\rho}_\theta^{PRP}$ and taking action a. Then, the average reward is a rational function of the policy parameters, defined by:*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{t=0}^{k-1} R_t = \sum_{a\in\mathcal{A}} \mathbf{r}_a^\top \boldsymbol{\rho}_\theta^{PRP}(a) \cdot \boldsymbol{\theta}_a^\top \boldsymbol{\rho}_\theta^{Pol} \quad a.s.,$$

where the rewards are collected by following policy θ, and $\forall a \in \mathcal{A} : \boldsymbol{\theta}_a$ represents the vector of probabilities of taking action a in each state of the policy.

The form of the average reward in a linear PRP is not surprising, since the average reward takes the same form in the MDP/POMDP setting. However, this result is important not only due to the generalization

of the average reward behavior from POMDPs to linear PRPs, but because of its implications on the complexity of the average reward function. The following corollary connects the complexity of the average reward function to the dimensionality of the PSR representation and the number of hidden states in the corresponding POMDP representation. Since the average reward is a rational function of policy parameters in both cases, we measure the complexity of this function in terms of the degree of the multivariate polynomials appearing in either numerator or denominator, whichever is larger.

Corollary 5. *Consider a system that can be represented by both POMDP with m hidden states and n-dimensional linear PRP. Let k be the size of the stochastic finite state controllers under consideration. The degree of the average reward function, when analyzed using the linear PRP representation is $O(kn)$, which can be (significantly) smaller compared to the function of degree $O(km)$ obtained in the POMDP framework.*

In the following subsections, we address two points that were not yet discussed in enough detail. First, we discuss the representational power of a linear PRP, compared to that of a POMDP. Second, we provide several synthetic examples of linear PRPs whose dimension is smaller than that of a corresponding POMDP. These examples highlight the benefits of the linear PRP framework.

4.1. Representation power of a linear PRP

As outlined above, existing planning algorithms in PSRs fall under the linear PRP representation assumptions. However, certain reward processes can be represented with a POMDP but not with a linear PRP. Although this might seem as a disadvantage, we argue that such systems are ill–modeled from the perspective of reinforcement learning.

The discrepancy arises from the fact that although the belief state of a POMDP is always sufficient to predict the expected reward, the predictive state for the future action–observation sequences is not necessarily sufficient. For example, one can think of a POMDP with aliased hidden states that only differ in terms of their reward function. However, this setting undermines the classical approach of constructing policies solely based on actions and observations, since it is clear that knowing the reward signal allows making better predictions of future rewards in this case. If the model of the system is not known a priori, it is even less clear how to learn such a model. To the best of our knowledge, current state-of-the-art model learning techniques estimate the reward function *after* the action–observation process has been modeled. Even when the model is known, peculiar behavior might take place if the policy ignores the reward signal, e.g., the average reward can depend on the initial hidden state of the POMDP, despite the hidden dynamics being policy independent (Yu & Bertsekas, 2008). Hence, in reinforcement learning, it is reasonable to assume that the observations provide enough information about the reward, as is the case in the linear PRP setting.

4.2. Linear PRP examples

We now present a few synthetic examples of dynamical systems with control that can be represented exactly by a PRP significantly more compactly, compared to the most compact representation of the system in the POMDP framework. The examples are based on the *probability clock* example without control from Jaeger (2000), described in the OOM framework. Although these are synthetic examples, such "clocks" are in fact common in nature, e.g. in bistable biological systems (Chaves et al., 2008). The shared property among our examples is the fact that at least one of the PSR matrices is a rotation matrix. The angle of the rotation will, in fact, determine the minimum number of states that the POMDP needs to represent the system. We first describe the probability clock example, since all the following systems are based on the same concept.

Consider a system with two observations $\mathcal{O} = \{o_1, o_2\}$, and 3-dimensional state. The initial state is a vector $\mathbf{s}_0 = (0.75, 0, 0.25)^\top$, and the following two matrices represent the change of state corresponding to each of the observations:

$$\mathbf{M}_{o_1} = 0.5 \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & -\sin(\alpha) & \cos(\alpha) \end{pmatrix},$$

$$\mathbf{M}_{o_2} = \mathbf{s}_0 \cdot \begin{pmatrix} 1 - 0.5 & & \\ 1 - 0.5[\cos(\alpha) - \sin(\alpha)] & & \\ 1 - 0.5[\cos(\alpha) + \sin(\alpha)] & & \end{pmatrix}^\top.$$

$\mathbf{M}_{o_1}$ performs the rotation of the state by an angle α around the first axis, and $\mathbf{M}_{o_2}$ resets the state back to $\mathbf{s}_0$. The predictions of the system are given by :

$$\mathrm{P}(o^{(1)}o^{(2)}, ..., o^{(k)}) = \mathbf{1}^\top \mathbf{M}_{o^{(k)}} \cdots \mathbf{M}_{o^{(1)}} \mathbf{s}_0.$$

The name of this example is due to the interesting behavior of the one step prediction of o_1: if we observe only sequences of o_1-s this prediction oscillates (the pattern is as in Fig. 1). The minimum number of states required for an HMM to represent this system exactly is the minimum k such that $k \cdot \alpha$ is a multiple of

2π (Jaeger, 2000). Hence, depending on α, one might require an arbitrarily large number of states in the corresponding HMM for an exact representation; an infinite number of states is needed if α is an irrational degree.

We now present several systems with control and reward functions that generalize the idea of the probability clock example to POMDPs. All the examples consider two-action ($\mathcal{A} = \{a, b\}$) two-observation ($\mathcal{O} = \{o_1, o_2\}$) systems, such that policies with small memory perform (at times significantly) better than constant policies. The first and perhaps the most interesting scenario is a simple extension of the three dimensional probability clock. Let

$$\mathbf{M}_{a,o_1} \triangleq \mathbf{M}_{o_1}, \quad \mathbf{M}_{a,o_2} \triangleq \mathbf{M}_{o_2},$$

$$\mathbf{M}_{b,o_1} \triangleq \mathbf{M}_{b,o_2} \triangleq 0.5 \cdot \mathbf{I},$$

where $\mathbf{I}$ is an identity matrix of appropriate dimension. Hence, taking action a will result in the same behavior as that of the probability clock example. Taking action b, though, will result in an i.i.d sequence of coins flips - observations o_1, o_2 - but will not change the state. We equip this system with a reward signal and let its expectation be equal to $\mathrm{P}(o_1|\mathbf{s}, a)$, where $\mathbf{s}$ is the current system's state. This specification satisfies the requirement of a linear PRP, since the expected reward is a linear function of the system's state. The optimal behavior is to take action a until the system reaches the state with largest $\mathrm{P}(o_1|\mathbf{s}, a)$, then choose action b thereafter.

The above example should be interpreted as the system that can operate under different modes. Action b runs the system in a given mode while action a changes the mode of the system in a stochastic fashion. Clearly, one can generalize this example to the setting in which the system's operation itself requires several hidden states to represent, more actions are involved, etc. Operating the system in different modes does not affect the dynamics between these states, but potentially affects the reward obtained from each state-action pair. Following the same reasoning as above, it is clear that such a system might require significantly more hidden states than number of dimensions, if represented in the POMDP framework. We note that the lac operon (see e.g. Chaves et al. (2008)), as well as other genetic networks related to metabolizing different types of nutrients, exhibit this type of "multiple mode" operation, with rewards dependent on the mode.

The rest of the examples illustrate the flexibility of the probability clock with respect to the choice of parameters. The reward function is the same for all cases: 1 for observation o_1 and 0 for o_2; hence, the objective is to maximize the average probability of o_1 per step. As before, we let o_1 be the observation that rotates the state and o_2 be the reset. Figure 1 illustrates a system in which both actions behave as probability clocks starting from the same state, but having different rotation angles. The policy that chooses either action a or b at all times obtains an average reward of ≈ 0.69, while the policy that takes action a three times given that no reset occurred, and then b until the next reset, obtains an average reward of ≈ 0.72. Figure 2 illustrates an example in which two actions can have different resetting states, rotation angles and magnitudes of the cycles. A constant policy choosing one action at all times will obtain average reward less than 0.36. Yet, the policy that "climbs the hill" using action b and, once reset, chooses action a until the next reset, achieves an average reward of ≈ 0.66. Such a policy requires only two internal states. The behavior of the average reward as a function of some of the two-state policy parameters is also presented in Figure 2. Another example of a system whose actions rotate the state in opposite directions and have slightly different resetting states can be found in the supplementary material depicted in Figure 3. A constant policy choosing one action at all times will obtain average reward ≈ 0.43, while the policy that changes its action right after observing a reset achieves average reward of ≈ 0.66. As in the previous example, such a policy requires only two internal states, with average reward behaving smoothly as shown in Figure 3.

All these examples are meant to give a sense of a type of systems that can benefit from the linear PRP representation. Although these systems require a much larger number of hidden states to be represented in the POMDP framework, they only require a 3-dimensional linear PRP representation. This guarantees that the average reward is a simple function of a policy with small memory, suggesting that appropriate policy search techniques can be very efficient.

5. Conclusion and future work

We proposed a formal definition for the reward process based on the linear PSR framework, and analyzed some of its properties. We proved that the average reward is a rational function of the policy parameters, and its complexity depends on the dimension of the underlying linear PSR. This suggests that systems represented by a small linear PRP should be amenable to effective policy search techniques. For example, one can use gradient-based policy search methods such as GPOMDP (Baxter & Bartlett, 2001) to efficiently find

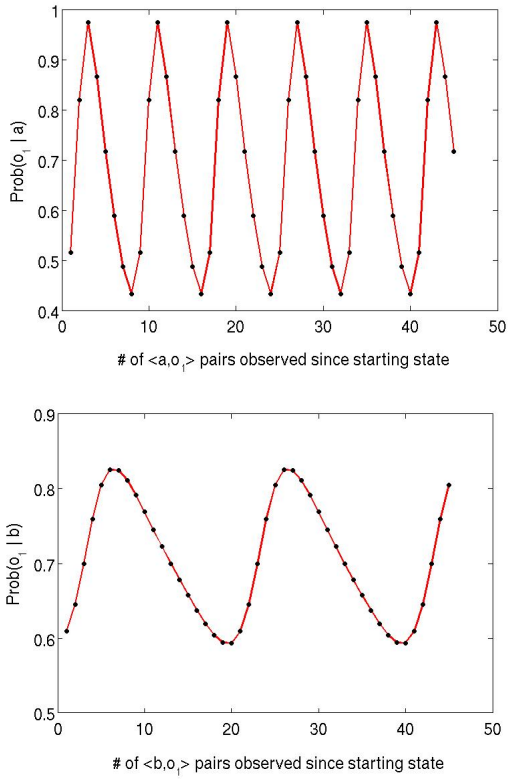
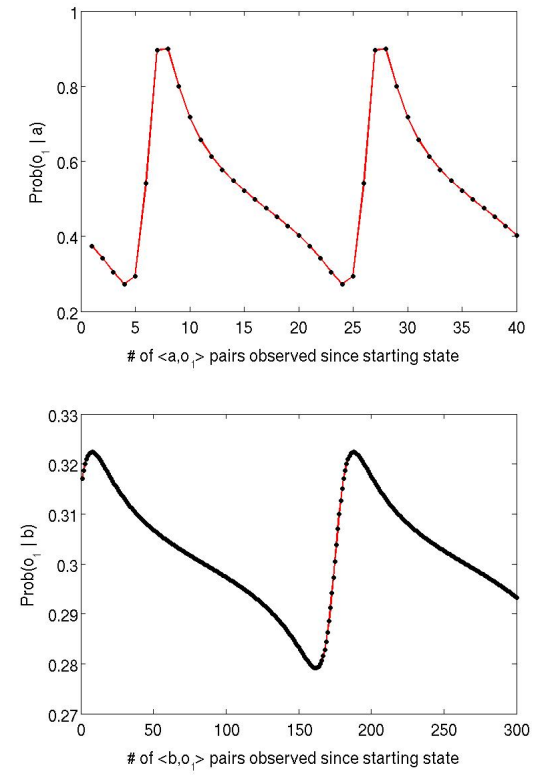


Figure 1. A system which rotates the state by a different angle for each action: $a \rightarrow 45^\circ, b \rightarrow 18^\circ$.

a policy with an acceptable performance. However, previous policy search methods do not seem to exploit the shape of the average reward function derived in this paper. As a result, this is a particularly interesting avenue for global policy search methods since properties like smoothness of the function are typically easy to exploit using this type of search. Moreover, the knowledge about the possible dimensionality of the system can boost the policy search even more. The development of suitable approaches and their analysis remains the main direction for future work.

Another direction currently under investigation is the analysis of the behavior of the average reward function for policies dependent on the predictive sufficient statistic of the process, i.e. the underlying state of the linear PSR (Aberdeen et al., 2007). This analysis could further shed light on how the error in an approximated PSR state affects the policy performance.

Finally, the established result on the stationary distribution of a linear PSR can also be useful in the future development of stability guarantees for linear PSRs.



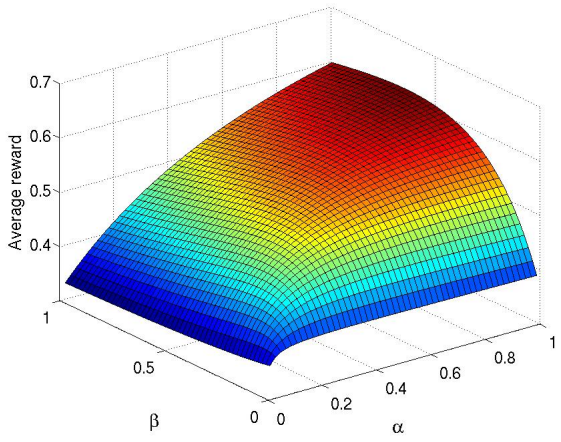


Figure 2. The first two plots describe the behavior of the system: 1) it rotates the state by a different angle for each action: $a \rightarrow 18^\circ, b \rightarrow 2^\circ$; 2) the resetting state for action a corresponds to a basin of the cycle, while the resetting state for action b corresponds to a near top of the cycle; 3) the magnitudes of the periods are different between the actions.

The bottom plot demonstrates how the average reward changes as a function of α and β, where the policies having 2 hidden states ($S \in \{1, 2\}$) are parametrized as:
$\mathrm{P}_\pi(a|S = 1) = \mathrm{P}_\pi(b|S = 2) = 1$,
$\mathrm{P}_\pi(S = 2|S = 1, a) = \mathrm{P}_\pi(S = 1|S = 2, a) = \alpha$,
$\mathrm{P}_\pi(S = 1|S = 1, b) = \mathrm{P}_\pi(S = 2|S = 2, b) = \beta$.

References

Aberdeen, D., Buffet, O., and Thomas, O. Policy-gradients for PSRs and POMDPs. In *Proceedings of international conference on artificial intelligence and statistics*, 2007.

Baxter, J. and Bartlett, P.L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Boots, B., Siddiqi, S.M., and Gordon, G.J. Closing the learning-planning loop with predictive state representations. In *Proceedings of the 9th international conference on autonomous agents and multiagent systems*, pp. 1369–1370, 2010.

Chaves, M., Eissing, T., and Allgower, F. Bistable biological systems: A characterization through local compact input-to-state stability. *Automatic Control, IEEE Transactions on*, 53(Special Issue):87–100, 2008.

Faigle, U. and Schonhuth, A. Asymptotic mean stationarity of sources with finite evolution dimension. *Information Theory, IEEE Transactions on*, 53(7): 2342–2348, 2007.

Gray, R.M. *Probability, random processes, and ergodic properties*. Springer, 2009.

Gray, R.M. and Kieffer, JC. Asymptotically mean stationary measures. *The Annals of Probability*, 8(5): 962–973, 1980.

Grinberg, Y. and Precup, D. On average reward policy evaluation in infinite–state partially observable systems. In *Proceedings of international conference on artificial intelligence and statistics*, 2012.

Izadi, M.T. and Precup, D. A planning algorithm for predictive state representations. In *Proceedings of the 18th international joint conference on artificial intelligence*, pp. 1520–1521. Morgan Kaufmann Publishers Inc., 2003.

Jaeger, H. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6): 1371–1398, 2000.

James, M.R. *Using predictions for planning and modeling in stochastic environments*. PhD thesis, The University of Michigan, 2005.

James, M.R., Singh, S., and Littman, M.L. Planning with predictive state representations. In *The 2004 international conference on machine learning and applications*, 2004.

Kaelbling, L.P., Littman, M.L., and Cassandra, A.R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998.

Littman, M.L., Sutton, R.S., and Singh, S. Predictive representations of state. *Advances in neural information processing systems*, 14:1555–1561, 2001.

Peters, J. and Schaal, S. Natural actor-critic. *Neurocomputing*, 71(7):1180–1190, 2008.

Rosencrantz, M., Gordon, G., and Thrun, S. Learning low dimensional predictive representations. In *Proceedings of the twenty-first international conference on machine learning*, pp. 88, 2004.

Schweitzer, P.J. Perturbation theory and finite Markov chains. *Journal of Applied Probability*, pp. 401–413, 1968.

Singh, S., James, M.R., and Rudary, M.R. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th conference on uncertainty in artificial intelligence*, pp. 512–519. AUAI Press, 2004.

Singh, S.P., Jaakkola, T., and Jordan, M.I. Learning without state-estimation in partially observable Markovian decision processes. In *Proceedings of the eleventh international conference on machine learning*, volume 31, pp. 37. Citeseer, 1994.

Sutton, R. `http://www.incompleteideas.net/sutton/book/errata.html`, 1998.

Wiewiora, E.W. *Modeling probability distributions with predictive state representations*. PhD thesis, The University of California at San Diego, 2007.

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