Average Reward Optimization Objective In Partially Observable Domains - Supplementary Material

Another example of a controlled system (see Sec. 4.2)

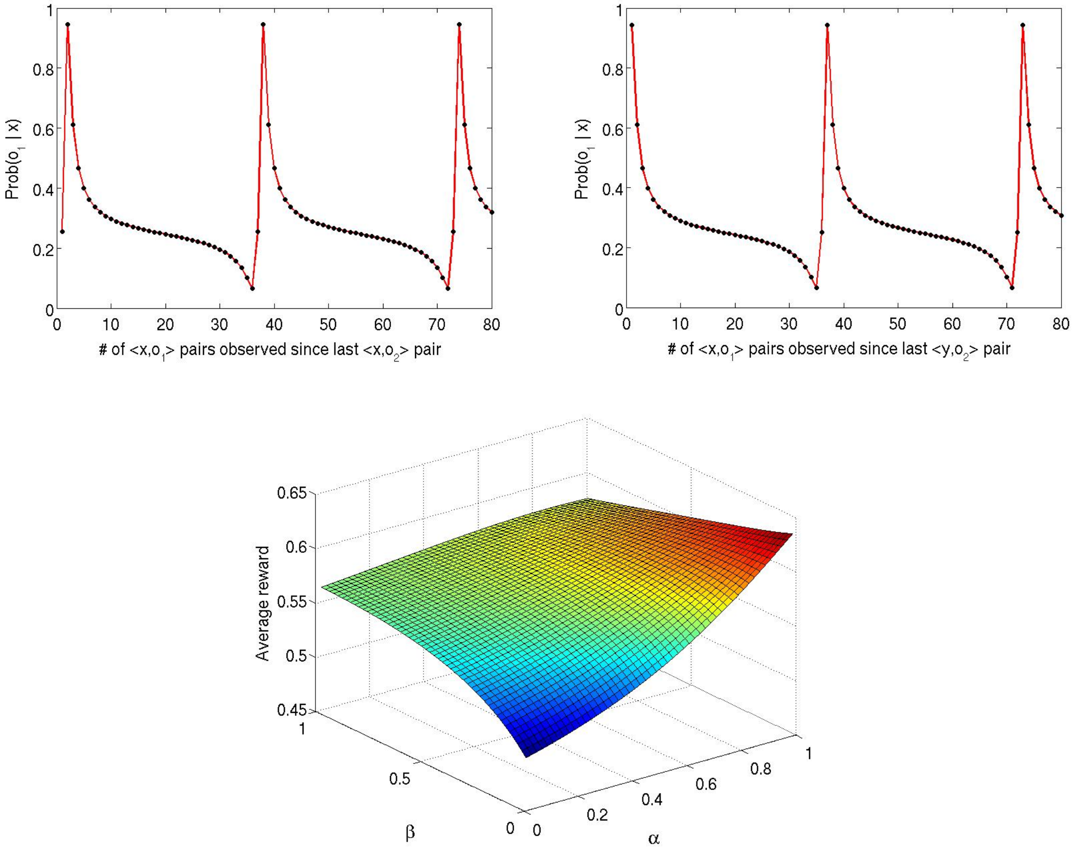

Figure 3. The first two plots describe the behavior of the system which rotates the state by $a \rightarrow 10^\circ$ and $b \rightarrow -10^\circ$, with the reset states aligned in a way that taking the opposite action brings the system to its topmost point. The x and y in the plots represent possible actions such that $x \neq y$.

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:

$\mathrm{P}_\pi(a|S = 1) = \mathrm{P}_\pi(b|S = 2) = 1$,
$\mathrm{P}_\pi(S = 1|S = 1, a, o_1) = \mathrm{P}_\pi(S = 2|S = 2, a, o_1) = \alpha$,
$\mathrm{P}_\pi(S = 2|S = 1, a, o_2) = \mathrm{P}_\pi(S = 1|S = 2, a, o_2) = \alpha$,
$\mathrm{P}_\pi(S = 1|S = 1, b, o_1) = \mathrm{P}_\pi(S = 2|S = 2, b, o_1) = \beta$,
$\mathrm{P}_\pi(S = 2|S = 1, b, o_2) = \mathrm{P}_\pi(S = 1|S = 2, b, o_2) = \beta$.

Proof of Lemma 1

1.: $\mathbf{E}\mathbf{E}^\infty = \mathbf{E}^\infty\mathbf{E} = \mathbf{E}^\infty$ is immediate from the definition of $\mathbf{E}^\infty$. Therefore $[\frac{1}{n}\sum_{t=0}^{n-1}\mathbf{E}^t]\mathbf{E}^\infty = \mathbf{E}^\infty$ for any n, implying that also $\mathbf{E}^\infty\mathbf{E}^\infty = \mathbf{E}^\infty$.
2.:

$$\begin{aligned}(\mathbf{E}-\mathbf{E}^\infty)^n &= (\mathbf{E}-\mathbf{E}\mathbf{E}^\infty)^n = \mathbf{E}^n(\mathbf{I}-\mathbf{E}^\infty)^n = \mathbf{E}^n\left[\sum_{i=0}^{n}(-1)^i\binom{n}{i}(\mathbf{E}^\infty)^i\right]\\ &= \mathbf{E}^n\left[\mathbf{I}+\left(\sum_{i=0}^{n}(-1)^i\binom{n}{i}-1\right)\mathbf{E}^\infty\right] = \mathbf{E}^n-\mathbf{E}^\infty,\end{aligned}$$

where some of the equalities follow from property 1, and the last equality holds since the alternating sum of binomial coefficients equals to 0.
3.: Since $(\mathbf{E}-\mathbf{I})$ has rank $n-1$, the eigenspace of $\mathbf{E}$ corresponding to the eigenvalue 1 is one–dimensional. So we have a unique $\boldsymbol{\rho}$ such that $\boldsymbol{\rho}^\top\mathbf{E} = \boldsymbol{\rho}^\top$ and $\boldsymbol{\rho}^\top\mathbf{1} = 1$. Also, if $\mathbf{E}\mathbf{v} = \mathbf{v}$ then $\mathbf{v}$ is a constant vector because the rows of $\mathbf{E}$ sum to 1. Hence also, $\boldsymbol{\rho}^\top\mathbf{E}^\infty = \boldsymbol{\rho}^\top$ and $\mathbf{E}^\infty\mathbf{1} = \mathbf{1}$. Now, from property 1 we have $\mathbf{E}^\infty(\mathbf{E}-\mathbf{I}) = 0$, meaning that the rows of $\mathbf{E}^\infty$ are multiples of $\boldsymbol{\rho}$. Also, we have $(\mathbf{E}-\mathbf{I})\mathbf{E}^\infty = 0$, meaning that the columns of $\mathbf{E}^\infty$ are constant vectors. Combining all together we get $\mathbf{E}^\infty = \mathbf{1}\boldsymbol{\rho}^\top$.
4.: Observe that $(\mathbf{E}-\mathbf{E}^\infty)^n$ is Cesaro summable to 0, i.e.

$$\frac{1}{n}\sum_{t=0}^{n-1}(\mathbf{E}-\mathbf{E}^\infty)^t = \left[\frac{1}{n}\sum_{t=0}^{n-1}\mathbf{E}^t\right]-\mathbf{E}^\infty \overset{n\to\infty}{\longrightarrow} 0,$$

due to property 2. So, by Kemeny & Snell (1960) (Thm. 1.11.1) and property 2,

$$[\mathbf{I}-(\mathbf{E}-\mathbf{E}^\infty)]^{-1} = \lim_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}\sum_{k=0}^{t}(\mathbf{E}-\mathbf{E}^\infty)^t = \mathbf{I}+\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n-1}\sum_{k=1}^{t}(\mathbf{E}^k-\mathbf{E}^\infty).$$

5.: Follows from replacing $\mathbf{Z}$ with the right-hand side of Eq. (2).

□

Proof of Theorem 2

The proof is essentially identical to the proof of Theorem 1 in Schweitzer (1968), which is, however, restricted to Markov chain transition matrices. We have,

$$\begin{aligned}\mathbf{I}-(\mathbf{E}_2-\mathbf{E}_1)\mathbf{Z}_1 &= (\mathbf{Z}_1^{-1}-\mathbf{E}_2+\mathbf{E}_1)\mathbf{Z}_1 = (\mathbf{I}-\mathbf{E}_2+\mathbf{E}_1^\infty)\mathbf{Z}_1\\ &= (\mathbf{Z}_2^{-1}+\mathbf{E}_1^\infty-\mathbf{E}_2^\infty)\mathbf{Z}_1 = \mathbf{Z}_2^{-1}(\mathbf{I}+\mathbf{E}_1^\infty-\mathbf{E}_2^\infty)\mathbf{Z}_1,\end{aligned}$$

where all (except for the last) inequalities follow from the definition of $\mathbf{Z}_i$. The last equality holds since $\mathbf{Z}_2^{-1}\mathbf{E}_1^\infty = \mathbf{E}_1^\infty$ and $\mathbf{Z}_2^{-1}\mathbf{E}_2^\infty = \mathbf{E}_2^\infty$, which in turn is due to $\mathbf{Z}_2^{-1}$ having rows sum to 1, and property 3 in Lemma 1. Now, observe that $(\mathbf{E}_2^\infty-\mathbf{E}_1^\infty)^2 = 0$, meaning that Theorem 1.11.1 in Kemeny & Snell (1960) can be applied on $(\mathbf{E}_2^\infty-\mathbf{E}_1^\infty)$, and we get

$$[\mathbf{I}-(\mathbf{E}_2^\infty-\mathbf{E}_1^\infty)]^{-1} = \mathbf{I}-\mathbf{E}_1^\infty+\mathbf{E}_2^\infty,$$

which proves Eq. (3). Finally,

$$\boldsymbol{\rho}_1^\top\mathbf{H}_{1\to 2} = \boldsymbol{\rho}_1^\top\mathbf{Z}_1^{-1}[\mathbf{I}-\mathbf{E}_1^\infty+\mathbf{E}_2^\infty]\mathbf{Z}_2 = \boldsymbol{\rho}_1^\top\mathbf{E}_2^\infty\mathbf{Z}_2 \overset{\star}{=} \boldsymbol{\rho}_2^\top\mathbf{Z}_2 = \boldsymbol{\rho}_2^\top,$$

where $(\star)$ is due to property 3 in Lemma 1 and $\boldsymbol{\rho}_1^\top\mathbf{1} = 1$, while the last equality is due to property 5 in Lemma 1.

□

Proof of Theorem 3

Recall the concept of a *system dynamics matrix* (SDM) (Singh et al., 2004), also known as *prediction matrix* (Faigle & Schonhuth, 2007). This is an infinite dimensional matrix, whose columns correspond to possible histories (ordered in an increasing length) and rows correspond to possible tests (ordered in an increasing length). The elements of this matrix represent conditional probabilities of observing each test given each history (for more details see Singh et al. (2004); Faigle & Schonhuth (2007)). If the SDM has rank k then the system can be represented by a k-dimensional linear PSR, and the square submatrix of SDM of dimension $|\mathcal{A} \times \mathcal{O}|^k$ will be of rank k (e.g. see James (2005)).

According to Wiewiora (2007), the system resulting from combining a k-dimensional linear PSR with control and a policy with memory of size l can be represented with a linear PSR without control whose dimension is at most $k \times l$. We will denote the distribution over future sequences in such a system with P_θ.

Let n be the largest rank of a SDM induced by some policy θ, implying that a minimal dimensional PSR for this system has rank n (Wiewiora, 2007). Let $\mathfrak{g}_\theta^{(i)}$ represent the *expected* state of the system after i time steps, i.e. $\mathfrak{g}_\theta^{(i)}$ is an infinite-dimensional vector such that $\mathfrak{g}_\theta^{(i)}(t) = \sum_{h \in \langle \mathcal{A} \times \mathcal{O} \rangle^i} \mathrm{P}_\theta(t|h)\mathrm{P}_\theta(h)$. Since $\mathfrak{g}_\theta^{(i)}$ are linear combinations of columns of SDM, we have that $\mathfrak{G}_\theta \triangleq \mathrm{span}\{\mathfrak{g}_\theta^{(0)}, \mathfrak{g}_\theta^{(1)}, ...\}$ is at most n dimensional. Let $m \leq n$ be the maximum dimension of $\mathfrak{G}_\theta$ for any $\theta \in \Theta$. We will use the following results from Faigle & Schonhuth (2007), which consider a stochastic process represented by an OOM **for a fixed** $\theta \in \Theta$:

1. If $\mathfrak{G}_\theta$ is m-dimensional, $\{\mathfrak{g}_\theta^{(0)}, ..., \mathfrak{g}_\theta^{(m-1)}\}$ is a basis for $\mathfrak{G}_\theta$.

2. the state *evolution operator* can be represented by the matrix $\mathbf{E}_\theta \in \mathbb{R}^{m \times m}$ relative to this basis, such that

$$\mathbf{E}_\theta^\top = \begin{bmatrix} 0 & 0 & ... & 0 & c_\theta^{(0)} \\ 1 & 0 & ... & 0 & c_\theta^{(1)} \\ 0 & 1 & ... & 0 & c_\theta^{(2)} \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & ... & 1 & c_\theta^{(m-1)} \end{bmatrix},$$

   where $\sum_{j=0}^{m-1} c_\theta^{(j)} = 1$, and all $c_\theta^{(j)}$-s are unique. $\mathbf{E}_\theta$ is the PSR evolution matrix represented under a different basis since $\forall j \in \mathbb{N} : \mathfrak{g}_\theta^{(j)} = \sum_{i=0}^{m-1} a_i^{(j)} \mathfrak{g}_\theta^{(i)}$, where $\mathbf{a}^{(j)} \triangleq (a_0^{(j)}, ..., a_{m-1}^{(j)}) = (1, 0, ..., 0) \cdot \mathbf{E}_\theta^j$. In other words, it outputs the next expected state provided the current state.

3. Furthermore, $\mathbf{E}_\theta^\infty \triangleq \lim_{k \to \infty} \frac{1}{k} \sum_{t=0}^{k-1} \mathbf{E}_\theta^t$ exists.

The remaining of the proof concerns the construction of $c_\theta^{(i)}$-s as functions of θ and showing that these are well defined rational functions of θ for all $\theta \in \Theta$. Once this is done, Theorem 2 can be applied on two arbitrary matrices $\mathbf{E}_{\theta_1 \in \Theta}, \mathbf{E}_{\theta_2 \in \Theta}$, concluding the proof.

Due to the properties of SDM mentioned above, we can identify infinite-dimensional vectors $\mathfrak{g}_\theta^{(i)}$ with their finite dimensional counterparts $\mathbf{g}_\theta^{(i)} \in \mathbb{R}^{|\mathcal{A} \times \mathcal{O}|^n}$, where $\forall t \in \{\langle \mathcal{A} \times \mathcal{O} \rangle^i\}_{i=1,...,n} : \mathbf{g}_\theta^{(i)}(t) = \mathfrak{g}_\theta^{(i)}(t)$, such that

$$\dim \left[\mathrm{span}\{\mathbf{g}_\theta^{(0)}, \mathbf{g}_\theta^{(1)}, ...\}\right] = \dim \left[\mathrm{span}\{\mathfrak{g}_\theta^{(0)}, \mathfrak{g}_\theta^{(1)}, ...\}\right],$$

for all $\theta \in \Theta$. Note that the same matrix $\mathbf{E}_\theta$ represents the unique[1] evolution matrix relative to $\{\mathbf{g}_\theta^{(0)}, ..., \mathbf{g}_\theta^{(m-1)}\}$. Let $\mathbf{G}_\theta \in \mathbb{R}^{|\mathcal{A} \times \mathcal{O}|^n \times m}$ be defined as

$$\mathbf{G}_\theta = \left[\mathbf{g}_\theta^{(0)}, \mathbf{g}_\theta^{(1)}, ..., \mathbf{g}_\theta^{(m-1)}\right].$$

Note that θ is a direct parametrization of a policy, meaning that elements of $\mathbf{G}_\theta$ are polynomials in θ. Thus, Proposition 1 is applicable to $\mathbf{G}_\theta$, and from there we obtain a well defined orthogonal basis $\{\mathbf{b}_\theta^{(0)}, ..., \mathbf{b}_\theta^{(m-1)}\}$ for

[1] As long as $\{\mathbf{g}_\theta^{(0)}, ..., \mathbf{g}_\theta^{(m-1)}\}$ is linearly independent set of vectors.

the column space of $\mathbf{G}_\theta$, such that these basis vectors are rational functions of θ.
Now we can define $c_\theta^{(i)}$ in the following recursive way:

$$\begin{array}{ll} i = m-1: & c_\theta^{(i)} = \left[\mathbf{g}_\theta^{(m)}\right]^\top \left[\frac{\mathbf{b}_\theta^{(m-1)}}{\|\mathbf{b}_\theta^{(m-1)}\|_2^2}\right] \\ i < m-1: & c_\theta^{(i)} = \left[\mathbf{g}_\theta^{(m)} - \sum_{j=m}^{i+1} c_\theta^{(j)} \mathbf{g}_\theta^{(j)}\right]^\top \left[\frac{\mathbf{b}_\theta^{(i)}}{\|\mathbf{b}_\theta^{(i)}\|_2^2}\right] \end{array} .$$

It is easy to verify that indeed we get $\mathbf{g}_\theta^{(m)} = \sum_{i=0}^{m-1} c_\theta^{(i)} \mathbf{g}_\theta^{(i)}$ for θ-s for which $c_\theta^{(i)}$-s are well defined. We also know that these coefficients are unique as long as $\mathbf{G}_\theta$ has full rank. Since these coefficients are rational functions of θ, by the same argument as in Proposition 1 we get that around any potential singularity point in $c_\theta^{(i)}$ we can find a sequence of θ-s converging to the singularity point while $c_\theta^{(i)}$ is well defined on this sequence. However, we know that for all θ-s where $c_\theta^{(i)}$-s are well defined, they must be bounded. This is due to the fact that $\mathbf{E}_\theta^\infty$ exists for all such θ-s, implying that the spectral radius of $\mathbf{E}_\theta$ is bounded (and equals to 1). Hence, $\forall 0 \le i < m : c_\theta^{(i)}$ do not have singularity points, or in other words, are well defined for all $\theta \in \Theta$.
Finally, we note that every $\mathbf{E}_\theta$ satisfies $\text{rank}(\mathbf{E}_\theta - \mathbf{I}) = n-1$ by construction, so the conditions of Theorem 2 for any pair of these matrices are satisfied. The left eigenvector corresponding to eigenvalue 1 of $\mathbf{E}_\theta$ identifies the stationary mean of the stochastic process $\mathcal{S}$ induced by policy θ due to the following: this vector represents the invariant state of the system with respect to column vectors in $\mathbf{G}_\theta$, which are by themselves linear combinations of any fixed column basis of SDM matrix. Recall that due to Theorem 2, the entries of this eigenvector are rational functions of policy parameters. Therefore, we get that each entry of the stationary distribution of PSR is also a rational function of policy parameters.
To complete the proof, note that the stationary distribution of PSR that we have obtained coincides with the empirical distribution of sequences observed from data since the stationary distribution is unique, which it turn is due to the ergodicity assumption.

□

Proposition 1. *Let $\mathbf{G}_\theta \in \mathbb{R}^{m \times n}, m \ge n$, be a rational matrix–valued function of $\theta \in \Theta$, such that $\mathbf{G}_\theta$ is bounded in $\theta \in \Theta$, and Θ be a subspace of some finite–dimensional Euclidean space. Assume that $\mathbf{G}_\theta$ has full rank for some $\theta \in \Theta$. Let $\{\mathbf{g}_\theta^{(i)}\}_{0 \le i < n}$ be the columns of $\mathbf{G}_\theta$. Then, $\{\mathbf{b}_\theta^{(0)}, ..., \mathbf{b}_\theta^{(n-1)}\}$ defined by*

$$\begin{array}{ll} i = 0: & \mathbf{b}_\theta^{(i)} = \mathbf{g}_\theta^{(i)} \\ i > 0: & \mathbf{b}_\theta^{(i)} = \mathbf{g}_\theta^{(i)} - \sum_{j=0}^{i-1} \frac{[\mathbf{g}_\theta^{(i)}]^\top \mathbf{b}_\theta^{(j)}}{\|\mathbf{b}_\theta^{(j)}\|_2^2} \mathbf{b}_\theta^{(j)} \end{array} ,$$

is a well defined orthogonal basis for the column space of $\mathbf{G}_\theta$ for all $\theta \in \Theta$.

Proof. Since $\mathbf{G}_\theta$ is rational in θ, $\{\mathbf{b}_\theta^{(i)}\}_{0 \le i < n}$ are rational functions (element–wise) in θ. For any θ, in particular where the function is not defined/singular[2], we can find a sequence $\theta_1, \theta_2, ... \longrightarrow \theta$ where this function is well defined. Suppose that such a singularity occurs to some element of $\mathbf{b}_{\theta_0}^{(i)}$ for some $\theta_0 \in \Theta$, and without loss of generality assume that $\mathbf{b}_{\theta_0}^{(j)}$ are well defined for $0 \le j < i$. Since $\forall j : \mathbf{g}_\theta^{(j)}$-s are bounded, at least one of the elements of the following terms $(0 \le j < i)$ must approach singularity as $\theta \to \theta_0$:

$$\frac{[\mathbf{g}_\theta^{(i)}]^\top \mathbf{b}_\theta^{(j)}}{\|\mathbf{b}_\theta^{(j)}\|_2^2} \mathbf{b}_\theta^{(j)} \longrightarrow \pm\infty.$$

But

$$\frac{[\mathbf{g}_\theta^{(i)}]^\top \mathbf{b}_\theta^{(j)}}{\|\mathbf{b}_\theta^{(j)}\|_2^2} \mathbf{b}_\theta^{(j)} = \left([\mathbf{g}_\theta^{(i)}]^\top \frac{\mathbf{b}_\theta^{(j)}}{\|\mathbf{b}_\theta^{(j)}\|_2}\right) \frac{\mathbf{b}_\theta^{(j)}}{\|\mathbf{b}_\theta^{(j)}\|_2},$$

meaning that $\mathbf{g}_\theta^{(i)}$ must have a singularity point at θ_0 since $\frac{\mathbf{b}_\theta^{(j)}}{\|\mathbf{b}_\theta^{(j)}\|_2}$ is bounded around θ_0, leading to a contradiction. □

[2] The function $f(x)$ is singular at points $\{x \in \mathcal{X} | \lim_{\hat{x} \to x} f(\hat{x}) = \pm\infty\}$. For rational functions, these are the roots of the polynomial in the denominator, which do not coincide with the roots of the polynomial in the numerator.

Proof of Theorem 4

The proof of this theorem uses two results that are given in Propositions 2 and 3 with their proofs attached. Proposition 2 provides a clear way of defining parameters of a PSR without control given a PSR (with control) and the parameters of a finite memory policy. Proposition 3 shows that one can obtain the state of the PSR (with control) from the state of the PSR without control by applying a linear operator.

First, we start by showing that $\lim_{k\to\infty}\frac{1}{k}\sum_{t=0}^{k-1} R_t$ exists almost surely for any policy $\theta \in \Theta$. Recall that we do not assume that the process is AMS (otherwise there is nothing to prove), only that the process is ergodic.
Let $\mathcal{R}_k = \frac{1}{k}\sum_{t=0}^{k-1} R_t$,

$$f(\cdot) \triangleq \begin{cases} \lim_{k\to\infty}\mathcal{R}_k & \text{if the limit exists} \\ R_{min}-1 & \text{otherwise} \end{cases},$$

$$f_k(\cdot) \triangleq \begin{cases} \mathcal{R}_k & \text{if } f(\cdot) \geq R_{min} \\ R_{min}-1 & \text{otherwise} \end{cases},$$

where R_{min} is the lowest achievable reward (recall that the reward is bounded). From the definition we have that

$$\{f_k\} \overset{k\to\infty}{\longrightarrow} f$$

pointwise everywhere. Since $\{f_k\}$ are uniformly bounded,

$$\lim_{k\to\infty} \mathrm{E}_\theta(f_k) = \mathrm{E}_\theta(f).$$

Finally, note that the set $\{f(\cdot) = y\}$ is invariant for any y because f is invariant: $f(\mathrm{T}x) = f(x)$. By the definition of ergodicity, such a set has either probability 1 or 0. Further we show that $\lim_{k\to\infty}\mathrm{E}_\theta(\mathcal{R}_k)$ is well defined, which ensures that

- $\mathrm{E}_\theta(f) = \mathrm{E}_\theta(\lim_{k\to\infty}\mathcal{R}_k) \geq R_{min}$ [3] ,
- $f = \mathrm{E}_\theta(f)$ almost surely.

Let $\{\mathbf{w}_{\theta,*}, \mathbf{W}_{\theta,ao}, \mathbf{p}_{\theta,0}\}$ be the minimal PSR without control for some policy θ, and let $\mathbf{W}_{\theta,*} \triangleq \sum_{ao}\mathbf{W}_{\theta,ao}$ be the PSR evolution matrix. Due to the properties of this matrix discussed earlier we have that

$$\mathbf{W}_\theta^\infty \triangleq \lim_{k\to\infty}\frac{1}{k}\sum_{t=0}^{k-1}\mathbf{W}_{\theta,*}^t \quad \text{exists.}$$

[3] Let B indicate the set for which $\lim_{k\to\infty}\mathcal{R}_k$ exists. Then $\lim_{k\to\infty}\mathrm{E}_\theta(\mathcal{R}_k) = \lim_{k\to\infty}\left[\int_B \mathcal{R}_k \mathrm{dP}_\theta + \int_{\bar{B}}\mathcal{R}_k \mathrm{dP}_\theta\right] = \lim_{k\to\infty}\int_B \mathcal{R}_k \mathrm{dP}_\theta + \lim_{k\to\infty}\int_{\bar{B}}\mathcal{R}_k \mathrm{dP}_\theta = \lim_{k\to\infty}\mathrm{E}_\theta(\mathcal{R}_k|B)\mathrm{P}_\theta(B) + \lim_{k\to\infty}\int_{\bar{B}}\mathcal{R}_k \mathrm{dP}_\theta$. Since both the left hand side limit and the first right hand side limit are well defined, then the second right hand side limit must exist as well, implying that $\mathrm{P}_\theta(\bar{B}) = 0$.

Therefore,

$$
\begin{align}
\lim_{k\to\infty} \mathrm{E}_\theta\left(\frac{1}{k}\sum_{t=0}^{k-1} R_t\right) &= \lim_{k\to\infty}\frac{1}{k}\sum_{t=0}^{k-1}\mathrm{E}_\theta(R_t) = \lim_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k-1}\mathrm{E}_\theta(R_t) \tag{1}\\
&= \lim_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k-1}\sum_{\bar{h}^{(t)}}\mathrm{E}(R_t|\bar{h}^{(t)})\mathrm{P}_\theta(\bar{h}^{(t)}) \tag{2}\\
&= \lim_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k-1}\sum_{\bar{h}^{(t-1)},a,o}\mathbf{r}_a^\top\mathbf{p}(\bar{h}^{(t-1)},a,o)\mathrm{P}_\theta(\bar{h}^{(t-1)},a,o) \tag{3}\\
&= \sum_a\lim_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k-1}\sum_{\bar{h}^{(t-1)},o}\mathbf{r}_a^\top\mathbf{V}_\theta\mathbf{p}_\theta(\bar{h}^{(t-1)},a,o)\mathrm{P}_\theta(\bar{h}^{(t-1)},a,o) \tag{4}\\
&= \sum_a\mathbf{r}_a^\top\mathbf{V}_\theta\left[\lim_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k-1}\sum_{\bar{h}^{(t-1)},o}\mathbf{p}_\theta(\bar{h}^{(t-1)},a,o)\mathrm{P}_\theta(\bar{h}^{(t-1)},a,o)\right] \notag\\
&= \sum_a\mathbf{r}_a^\top\mathbf{V}_\theta\left[\lim_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k-1}(\mathbf{p}_{\theta,0}^\top\mathbf{W}_{\theta,*}^t\mathbf{W}_{\theta,a*})^\top\right] \tag{5}\\
&= \sum_a\mathbf{r}_a^\top\mathbf{V}_\theta\left(\mathbf{p}_{\theta,0}^\top\left[\lim_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k-1}\mathbf{W}_{\theta,*}^t\right]\mathbf{W}_{\theta,a*}\right)^\top \notag\\
&= \sum_a\mathbf{r}_a^\top\mathbf{V}_\theta\left(\mathbf{p}_{\theta,0}^\top\mathbf{W}_\theta^\infty\mathbf{W}_{\theta,a*}\right)^\top = \sum_a\mathbf{r}_a^\top\mathbf{V}_\theta(\boldsymbol{\rho}_\theta^\top\mathbf{W}_{\theta,a*})^\top \tag{6}\\
&= \sum_a\mathbf{r}_a^\top\mathbf{V}_\theta\frac{\mathbf{W}_{\theta,a*}^\top\boldsymbol{\rho}_\theta}{\mathbf{w}_{\theta,*}^\top\mathbf{W}_{\theta,a*}^\top\boldsymbol{\rho}_\theta}\cdot\mathbf{w}_{\theta,*}^\top\mathbf{W}_{\theta,a*}^\top\boldsymbol{\rho}_\theta \notag\\
&= \sum_a\mathbf{r}_a^\top\mathbf{V}_\theta\boldsymbol{\rho}_\theta(a)\cdot\bar{\mathrm{P}}_\theta(a) \tag{7}\\
&= \sum_a\mathbf{r}_a^\top\mathbf{V}_\theta\boldsymbol{\rho}_\theta(a)\cdot\boldsymbol{\theta}_a^\top\boldsymbol{\rho}_\theta^{Pol} = \sum_a\mathbf{r}_a^\top\boldsymbol{\rho}_\theta^{PRP}(a)\cdot\boldsymbol{\theta}_a^\top\boldsymbol{\rho}_\theta^{Pol} \tag{8}
\end{align}
$$

where (3) follows from the definition of the linear PRP; (4) and (8) are due to Proposition 3 with $\mathbf{V}_\theta$ being the linear operator transforming the state of the PSR without control to the state of the PSR (with control); (5) is by definition of a linear PSR; (6) is due to the properties of the PSR evolution matrix $\mathbf{W}_{\theta,*}$, where $\boldsymbol{\rho}_\theta$ is the stationary distribution of the PSR without control induced by policy θ; and $\boldsymbol{\rho}_\theta(a)$ in (7) and (8) represents the state of the PSR without control obtained by starting from $\boldsymbol{\rho}_\theta$ and taking action a, while $\bar{\mathrm{P}}_\theta(a)$ is the average probability of taking action a.

Finally, note that $\boldsymbol{\rho}_\theta$ and $\boldsymbol{\rho}_\theta(a)$ are rational functions of $\theta \in \Theta$ by Theorem 3. Hence both $\boldsymbol{\rho}_\theta^{Pol}$ and $\boldsymbol{\rho}_\theta^{PRP}(a)$ are rational functions of θ as well since they can be immediately obtained from $\boldsymbol{\rho}_\theta$ and $\boldsymbol{\rho}_\theta(a)$ through summation and division (i.e., observe that the PSR state $\boldsymbol{\rho}_\theta^{PRP}$ is a vector of conditional predictions while the PSR without control state $\boldsymbol{\rho}_\theta$ is a vector of joint predictions).

□

Proposition 2. *Let $\{\mathbf{m}_*, \{\mathbf{M}_{ao}\}_{ao\in\mathcal{A}\times\mathcal{O}}, \mathbf{p}_0, \mathcal{Q} = \{q_1, ..., q_n\}\}$ be a n-dimensional linear PSR (with control) and $\{\mathbf{c}_0, \{\mathbf{T}_o\}_{o\in\mathcal{O}}, \{\mathbf{A}_a\}_{a\in\mathcal{A}}\}$ be the representation of a stochastic finite state policy π of size l. Let*

$$\mathbf{s}_0 \triangleq \mathbf{p}_0 \otimes \mathbf{c}_0 \in \mathbb{R}^{nl},$$

$$\mathbf{b}_* \triangleq \mathbf{m}_* \otimes \mathbf{1} \in \mathbb{R}^{nl},$$

$$\forall ao \in \mathcal{A}\times\mathcal{O} : \mathbf{B}_{ao} \triangleq \mathbf{M}_{ao} \otimes \mathbf{A}_a\mathbf{T}_o \in \mathbb{R}^{nl\times nl},$$

where $\otimes$ represents the Kronecker product. Then, the triple $\{\mathbf{b}_*, \{\mathbf{B}_{ao}\}_{ao\in\mathcal{A}\times\mathcal{O}}, \mathbf{s}_0\}$ satisfies the properties of a PSR without control induced by the policy π. In particular,

$$\mathrm{P}_\pi(a_1o_1, ..., a_ko_k) = \mathbf{s}_0^\top\mathbf{B}_{a_1o_1}\cdots\mathbf{B}_{a_ko_k}\mathbf{b}_*.$$

Proof.
For clarity, here is the description of the meaning of the stochastic finite state policy parameters:

- $\mathbf{c}_0$ - the initial distribution over the states of the policy
- $\mathbf{A}_a$ - the diagonal matrix with $[\mathbf{A}_a]_{ii}$ being equal to probability of taking action a in state i
- $\mathbf{T}_o$ - the transition matrix defining the state dynamics of the policy corresponding to observing percept o

Let $\mathbf{C}_{ao} \triangleq \mathbf{A}_a \mathbf{T}_o$. Recall that for any history h,

$$\begin{aligned}\mathrm{P}_\pi(\mathbf{o}_{1:k}|h, \mathbf{a}_{1:k}) &= \mathbf{p}(h)^\top \mathbf{M}_{a_1 o_1} \cdots \mathbf{M}_{a_k o_k} \mathbf{m}_* = \mathbf{p}(h)^\top \mathbf{M}_{\mathbf{ao}_{1:k}} \mathbf{m}_*, \\ \mathrm{P}_\pi(\mathbf{a}_{1:k}|h, \mathbf{o}_{1:k-1}) &= \mathbf{c}(h)^\top \mathbf{A}_{a_1} \mathbf{T}_{o_1} \cdots \mathbf{A}_{a_k} \mathbf{1} \triangleq \mathbf{c}(h)^\top \mathbf{C}_{\mathbf{ao}_{1:k-1}} \mathbf{A}_{a_k} \mathbf{1},\end{aligned}$$

where $\mathbf{1}$ is a column vector of ones of an appropriate size. Observe that

$$\mathrm{P}_\pi(ao) = \mathrm{P}_\pi(a)\mathrm{P}_\pi(o|a) = \mathbf{c}_0^\top \mathbf{C}_{ao}\mathbf{1} \cdot \mathbf{p}_0^\top \mathbf{M}_{ao}\mathbf{m}_*.$$

By induction (similarly to Wiewiora (2005)) we get,

$$\begin{aligned}\mathrm{P}_\pi(\mathbf{ao}_{1:k}, a_{k+1}, o_{k+1}) &= \mathrm{P}_\pi(\mathbf{ao}_{1:k}) \times \mathrm{P}_\pi(a_{k+1}|\mathbf{ao}_{1:k}) \times \mathrm{P}_\pi(o_{k+1}|\mathbf{ao}_{1:k}, a_{k+1}) \\ &= \mathbf{c}_0^\top \mathbf{C}_{\mathbf{ao}_{1:k}}\mathbf{1} \cdot \mathbf{p}_0^\top \mathbf{M}_{\mathbf{ao}_{1:k}}\mathbf{m}_* \\ &\times \mathbf{c}(\mathbf{ao}_{1:k})^\top \mathbf{A}_{a_{k+1}}\mathbf{1} \times \mathbf{p}(\mathbf{ao}_{1:k})^\top \mathbf{M}_{a_{k+1}o_{k+1}}\mathbf{m}_* \\ &= \mathbf{c}_0^\top \mathbf{C}_{\mathbf{ao}_{1:k}}\mathbf{1} \cdot \mathbf{p}_0^\top \mathbf{M}_{\mathbf{ao}_{1:k}}\mathbf{m}_* \\ &\times \frac{\mathbf{c}_0^\top \mathbf{C}_{\mathbf{ao}_{1:k}}}{\mathbf{c}_0^\top \mathbf{C}_{\mathbf{ao}_{1:k}}\mathbf{1}} \mathbf{A}_{a_{k+1}}\mathbf{1} \times \frac{\mathbf{p}_0^\top \mathbf{M}_{\mathbf{ao}_{1:k}}}{\mathbf{p}_0^\top \mathbf{M}_{\mathbf{ao}_{1:k}}\mathbf{m}_*} \mathbf{M}_{a_{k+1}o_{k+1}}\mathbf{m}_* \\ &= \mathbf{c}_0^\top \mathbf{C}_{\mathbf{ao}_{1:k}} \mathbf{A}_{a_{k+1}} \mathbf{T}_{o_{k+1}}\mathbf{1} \cdot \mathbf{p}_0^\top \mathbf{M}_{\mathbf{ao}_{1:k+1}}\mathbf{m}_* \\ &= \mathbf{p}_0^\top \mathbf{M}_{\mathbf{ao}_{1:k+1}}\mathbf{m}_* \cdot \mathbf{c}_0^\top \mathbf{C}_{\mathbf{ao}_{1:k+1}}\mathbf{1},\end{aligned}$$

where we used the fact that $\mathbf{C}_{\mathbf{ao}_{1:k}}\mathbf{1} = \mathbf{C}_{\mathbf{ao}_{1:k-1}}\mathbf{A}_{a_k}\mathbf{1}$ since every $\mathbf{T}_o$ is a transition matrix.
Now, the first property holds since

$$\mathbf{s}_0^\top \mathbf{b}_* = [\mathbf{p}_0 \otimes \mathbf{c}_0]^\top [\mathbf{m}_* \otimes \mathbf{1}] = \mathbf{p}_0^\top \mathbf{m}_* \cdot \mathbf{c}_0^\top \mathbf{1} = 1.$$

The second property also holds because

$$\begin{aligned}\left[\sum_{ao} \mathbf{B}_{ao}\right] \mathbf{b}_* &= \left[\sum_{ao} \mathbf{M}_{ao} \otimes \mathbf{C}_{ao}\right] [\mathbf{m}_* \otimes \mathbf{1}] = \sum_{ao} [\mathbf{M}_{ao}\mathbf{m}_* \otimes \mathbf{C}_{ao}\mathbf{1}] \\ &= \sum_{ao} [\mathbf{M}_{ao}\mathbf{m}_* \otimes \mathbf{A}_a\mathbf{1}] = \sum_a \left[\left(\sum_o \mathbf{M}_{ao}\right) \mathbf{m}_* \otimes \mathbf{A}_a\mathbf{1}\right] \\ &= \sum_a [\mathbf{m}_* \otimes \mathbf{A}_a\mathbf{1}] = \mathbf{m}_* \otimes \left[\sum_a \mathbf{A}_a\right] \mathbf{1} = \mathbf{m}_* \otimes \mathbf{1} = \mathbf{b}_*.\end{aligned}$$

Finally, the last property is due to

$$\begin{aligned}\mathrm{P}_\pi(\mathbf{ao}_{1:k}) &= \mathbf{p}_0^\top \mathbf{M}_{\mathbf{ao}_{1:k}}\mathbf{m}_* \cdot \mathbf{c}_0^\top \mathbf{C}_{\mathbf{ao}_{1:k}}\mathbf{1} = [\mathbf{p}_0 \otimes \mathbf{c}_0]^\top [\mathbf{M}_{\mathbf{ao}_{1:k}} \otimes \mathbf{C}_{\mathbf{ao}_{1:k}}][\mathbf{m}_* \otimes \mathbf{1}] \\ &= \mathbf{s}_0^\top \mathbf{B}_{\mathbf{ao}_{1:k}}\mathbf{b}_*.\end{aligned}$$

□

Proposition 3. *Let the action-observation stochastic process be generated from a linear n-dimensional PSR (with control),* $\mathbf{p}$*, and a stochastic finite state controller of size m parametrized by* $\theta \in \Theta$*. Let* $\mathbf{p}_\theta$ *be a* ***minimal*** *linear PSR without control representing this process. Then, the state of PSR (with control),* $\mathbf{p}$*, can be obtained from the state of PSR without control* $\mathbf{p}_\theta$ *by applying a linear operator.*

Proof. We will refer to $\mathbf{p}$ as both the PSR (with control) and the state of this PSR (which one is meant should be clear from the context). Equivalently, we will refer to $\mathbf{p}_\theta$ as both the PSR without control and the state of this PSR.

First, let $\{\mathbf{b}_*, \mathbf{B}_{ao}, \mathbf{s}_0\}$ be defined as in Proposition 2 from $\mathbf{p}$ and the parameters of the policy θ represented as $\{\mathbf{c}_0, \{\mathbf{T}_o\}_{o\in\mathcal{O}}, \{\mathbf{A}_a\}_{a\in\mathcal{A}}\}$. Let $\mathbf{U} = \mathbf{I} \otimes \mathbf{1}^\top \in \mathbb{R}^{n\times nm}$, where $\mathbf{I}$ is the n-dimensional identity matrix. Let

$$\mathbf{s}(h) \triangleq \mathbf{p}(h) \otimes \mathbf{c}(h)$$

be the state of this "PSR-like" construct after observing history h. Then,

$$\mathbf{U}\mathbf{s}(h) = [\mathbf{I} \otimes \mathbf{1}^\top]\mathbf{s}(h) = [\mathbf{I} \otimes \mathbf{1}^\top][\mathbf{p}(h) \otimes \mathbf{c}(h)] = \mathbf{p}(h) \otimes [\mathbf{1}^\top \mathbf{c}(h)] = \mathbf{p}(h).$$

Therefore we can recover the state of the environment $\mathbf{p}$ from the state of our construct by applying a linear operator. What is missing yet, is the connection between this construct and the PSR without control $\mathbf{p}_\theta$.
Recall that $\mathbf{p}$ is a vector of predictions of some sequences of observations given some sequences of actions. So it is clear that we can compute $\mathbf{p}$ from the joint distribution over action–observation pairs provided by the state of $\mathbf{p}_\theta$, which in general is a rational function of $\mathbf{p}_\theta$. Let $V_\theta(\cdot) : \mathbf{p}_\theta \to \mathbf{p}$ be such a mapping. Let the dimension of $\mathbf{p}_\theta$ be k, and $\mathbf{W}_\theta$ be $mn\times k$ matrix whose columns are coefficients of k core tests with respect to the construct $\{\mathbf{b}_*, \mathbf{B}_{ao}, \mathbf{s}_0\}$: $\mathbf{s}(h)^\top \mathbf{W}_\theta = \mathbf{p}_\theta(h)^\top$ for any h. Now, we have that for any h: $\mathbf{U}\mathbf{s}(h) = \mathbf{p}(h) = V(\mathbf{s}(h)^\top \mathbf{W}_\theta) = V(\mathbf{p}_\theta(h))$. From here the conclusion follows.
We can represent the operator V_θ with matrix $\mathbf{V}_\theta \in \mathbb{R}^{n\times k}$. $\mathbf{V}_\theta$ can be obtained by solving the following system of equations: $\mathbf{V}_\theta = \mathbf{P}\mathbf{P}_\theta^{-1}$, where $\mathbf{P}_\theta \in \mathbb{R}^{k\times k}$ represents a collection (matrix) of k linearly independent states (those corresponding to core histories) and $\mathbf{P} \in \mathbb{R}^{n\times k}$ represents a collection of environment states corresponding to the same histories. □

Proof of Corollary 5

In both representations the average reward function is linear in the stationary distribution as a function of the policy parameters, hence the complexity of the stationary distribution gives the upper bound on the complexity of the average reward. Let k be the size of the stochastic finite state controllers we consider. We analyze the complexity with respect to the POMDP representation first.
Observe that the HMM representing the combined system and the policy will have km hidden states. Since the Markov chain defined over the hidden states induced by any of the policies of interest is irreducible due to the ergodicity assumption, the corresponding transition matrix satisfies the conditions of Theorem 1 from Schweitzer (1968), which is a special case of Theorem 2. Each entry of this transition matrix is a polynomial of a fixed degree in the parameters of the policy. So are the entries of $\mathbf{H}_{1\to 2}^{-1}$ in Theorem 2 when we fix $\mathbf{E}_1, \mathbf{Z}_1, \boldsymbol{\rho}_1$ and let $\mathbf{E}_2$ be our parametrized transition matrix. One can analytically invert $\mathbf{H}_{1\to 2}^{-1}$ using Cramer's rule and observe that each entry of $\mathbf{H}_{1\to 2}$ is a rational function whose degree is $O(km)$. Therefore, by Theorem 2 the degree of the stationary belief state $\boldsymbol{\rho}_2$ is also $O(km)$.
Now we consider the reward process being represented by a linear PRP. Observe that the linear PSR without control representing the action–observation stochastic process induced by the policy above is at most kn dimensional. Each entry of the SDM describing this action–observation stochastic process is a rational function with the leading degree equal to the length of the history plus test. Therefore, the column vectors stored in $\mathbf{G}_\theta$ in the proof of Theorem 3 are of degree at most kn. The elements of the basis constructed from the columns of $\mathbf{G}_\theta$ are also of degree at most $O(kn)$ (see Proposition 1), as well as the entries of the bottom row of the evolution matrix $\mathbf{E}_\theta$ constructed in Theorem 3. Similarly to a POMDP case, we need to apply Theorem 2 for some fixed $\mathbf{E}_1, \mathbf{Z}_1, \boldsymbol{\rho}_1$, where $\mathbf{E}_2 \triangleq \mathbf{E}_\theta$ is the function of the policy parameters. Observe that from Theorem 2 we also have

$$\mathbf{H}_{1\to 2} = \mathbf{Z}_1^{-1}(\mathbf{Z}_1^{-1} + \mathbf{E}_1 - \mathbf{E}_2)^{-1},$$

where the resulting matrix $\mathbf{Z}_1^{-1} + \mathbf{E}_1 - \mathbf{E}_2$ has fixed entries in all the rows except for the bottom one, which is a rational function of degree at most $O(kn)$. Again, using Cramer's rule one can invert the matrix analytically, which results in $\mathbf{H}_{1\to 2}$ having entries being rational functions of degree at most $O(kn)$. Then, by Theorem 2 the degree of the stationary distribution of the PSR without control, $\boldsymbol{\rho}_\theta \triangleq \boldsymbol{\rho}_2$, is also at most $O(kn)$. Finally, one can verify that the quantities $\boldsymbol{\rho}_\theta^{PRP}(a)$ and $\boldsymbol{\rho}_\theta^{Pol}$ appearing in Theorem 4 will also have the leading degree of the

same order.
To conclude the proof, recall that the number of dimensions required to represent the same system in a linear PSR framework (n) is at most equal to the number of hidden states in the corresponding POMDP (m), and can be, at times significantly, smaller (Jaeger, 2000; Littman et al., 2001).

□

References

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

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.

Kemeny, J.G. and Snell, J.L. *Finite Markov chains*, volume 356. van Nostrand Princeton, NJ, 1960.

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

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.

Wiewiora, E. Learning predictive representations from a history. In *Proceedings of the 22nd international conference on machine learning*, pp. 964–971. ACM, 2005.

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