# Reconciling Rewards with Predictive State Representations

Andrea Baisero  Christopher Amato
{baisero.a,c.amato}@northeastern.edu
Khoury College of Computer Sciences, Northeastern University, Boston, USA

IJCAI 2021 MONTREAL

## Motivation and Contributions

PSRs are models of observation sequences which lack a latent state. While they are able to model arbitrary observation sequences, their ability to model non-observable rewards is limited.

**Contributions**
- Theory of non-observable PSR reward modeling.
- R-PSRs, an extension capable of modeling non-observable rewards.
- Value iteration for R-PSRs (R-PSR-VI).
- Evaluation on 63 domains from classical literature.

**Notation:** Superscript $^{(b)}$, $^{(p)}$, and $^{(r)}$ denote variables of POMDPs, PSRs, and R-PSRs.

## Partially Observable Markov Decision Processes (POMDPs)

A POMDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R \rangle$ is composed of:
- State, action and observation spaces $\mathcal{S}, \mathcal{A}$, and $\mathcal{O}$;
- State dynamics $T \colon \mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}$;
- Observation emissions $O \colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \Delta\mathcal{O}$;
- Reward function $R \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

We further denote the space of observable histories as $\mathcal{H} \doteq (\mathcal{A} \times \mathcal{O})^*$.
**Scope:** We primarily consider *finite* POMDPs (extensions to non-finite POMDPs also exist).

## Predictive State Representations (PSRs)

A PSR is a model of controlled observation sequences which:
- Does not have a *latent* state $s \in \mathcal{S}$ ($\implies$ easier to learn?).
- Has a *predictive* state $p(h)$ grounded in predictions of hypothetical *observable* futures.
- Can model the observation process of any finite POMDP (and more).
  (But has issues modeling rewards).

**Scope:** We primarily consider *linear* PSRs (extensions to non-linear PSRs also exist).

**Tests and their Outcomes**

A test $q \in \mathcal{Q} \doteq (\mathcal{A} \times \mathcal{O})^*$ is a hypothetical future sequence of actions and observations:
- Structurally equal to histories, but semantically distinct.
- Associated with an *outcome vector* $u(q) \in \mathbb{R}^{|\mathcal{S}|}$, where
$$[u(q)]_i \doteq \Pr(\bar{o}_q \mid s = i, \bar{a}_q) \,. \tag{1}$$
- A *core* test set $\mathcal{Q}^\dagger \subset \mathcal{Q}$ is any maximal set of lin. indep. tests; also satisfies $|\mathcal{Q}^\dagger| \leq |\mathcal{S}|$.
- The *outcome matrix* $U \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{Q}^\dagger|}$ is the col-stack of core outcomes,
$$[U]_{:i} = u(q_i), \quad q_i \in \mathcal{Q}^\dagger \,. \tag{2}$$

PSRs model *test probabilities* from given *histories*,
$$\begin{aligned} p(q \mid h) &\doteq \Pr(\bar{o}_q \mid h, \bar{a}_q) \\ &= b(h)^\top u(q) \\ &= p(h)^\top m_q \,, \end{aligned} \tag{3}$$
$$p(h) \doteq U^\top b(h) \,. \tag{4}$$

**Parameter Vectors:**
- $\{m_{ao} \mid a \in \mathcal{A}, o \in \mathcal{O}\}$ for the emission of observations, and
- $\{m_{aoq} \mid a \in \mathcal{A}, o \in \mathcal{O}, q \in \mathcal{Q}^\dagger\}$ for the dynamics.

**PSR Reward Functions and Observable Rewards**

**How to model PSR reward function?**
- Assume given reward function $R^{(p)}(h, a) = p(h)^\top \left[R^{(p)}\right]_{:a}$.
- Assume observable rewards.

**Non-Observable Rewards (Standard):**
- Reward available offline, at training time.
- Agent behavior *not conditioned* on rewards.
- Reward function $R \colon \mathcal{S} \to \mathbb{R}$.

**Observable Rewards (Non-Standard):**
- Reward available online, at execution time.
- Agent behavior *conditioned* on rewards.
- Reward function $R \colon \mathcal{O} \to \mathbb{R}$.

## Limitations of PSR Reward Models

**Open Questions:**
- Can POMDP rewards $R^{(b)}$ always be converted to PSR rewards $R^{(p)}$?
- Which $R^{(b)}$ can be converted to $R^{(p)}$?
- Can $R^{(b)}$ be approximated by $R^{(p)}$?
- Does approximate $R^{(p)}$ encode same task as $R^{(b)}$?

**Proposition**

For any finite POMDP and its respective PSR, a (linear or non-linear) function $f(p(h), a) \mapsto R^{(b)}(h, a)$ does not necessarily exist.

**Theorem (Accurate Linear PSR Rewards)**

A POMDP reward matrix $R^{(b)}$ can be accurately converted to a PSR reward matrix $R^{(p)}$ iff every column of $R^{(b)}$ is linearly dependent on the core outcome vectors (the columns of $U$). If this condition is satisfied, we say that the PSR is *accurate*, and $R^{(p)} = U^+ R^{(b)}$ accurately represents the POMDP rewards.

**Corollary**

Assuming that a PSR can be represented by a finite POMDP, then any PSR rewards $R^{(p)}$ are accurately represented by POMDP rewards $R^{(b)} = U R^{(p)}$.

**Theorem (Approximate Linear PSR Rewards)**

The linear approximation of POMDP rewards for non-accurate PSRs which results in the lowest reward approximation error is $R^{(p)} \doteq U^+ R^{(b)}$.

**Corollary**

$\tilde{R}^{(b)} \doteq U U^+ R^{(b)}$ is the reconstructed POMDP-form of the PSR approximation $R^{(p)}$ of the true POMDP rewards $R^{(b)}$. $\tilde{R}^{(b)}$ and $R^{(b)}$ are equal iff the accuracy condition holds.

## A Case Study of Approximate PSR Rewards

We use the *load/unload* domain as a case study to showcase the developed theory.

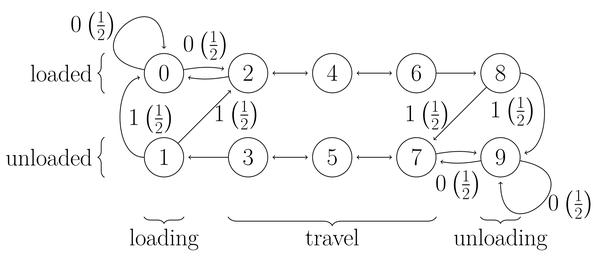


Figure: Load/unload domain. Rows indicate the agent's status (*loaded* or *unloaded*), and columns indicate the agent's position (observed as *loaded*, *travel*, and *unloaded*). Movements (*left* and *right*) are deterministic, and non-zero rewards are shown as "$x\,(y)$", where $x$ is the POMDP reward, and $y$ is the PSR approximation.

We can find a discrepancy between the POMDP rewards $R^{(b)}$ (note that only states 1 and 8 emit rewards),
$$R^{(b)} = \begin{pmatrix} 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \end{pmatrix}^\top , \tag{5}$$
the PSR approximation $R^{(p)} = U^+ R^{(b)}$ ,
$$R^{(p)} = \begin{pmatrix} 0.5 & -0.5 & -0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & -0.5 & 0.5 & 0.5 \end{pmatrix}^\top , \tag{6}$$
and the POMDP reconstruction $\tilde{R}^{(b)} = U R^{(p)}$ (note that states 0, 1, 8 and 9 emit rewards),
$$\tilde{R}^{(b)} = \begin{pmatrix} 0.5 & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.5 \end{pmatrix}^\top . \tag{7}$$
$\implies$ The approximate PSR rewards $R^{(p)}$ encode a catastrophically different task!

## Reward-Predictive State Representations (R-PSRs)

An R-PSR is a model of controlled observations and reward sequences which:
- Shares most properties (and derivation) with PSRs.
- **Can** model arbitrary non-observable rewards.

**Token Action $\zeta$**
- Defined to have unit reward $R(\cdot, \zeta) = 1$.
- Extended action space $\mathcal{Z} \doteq \mathcal{A} \cup \{\zeta\}$.
- Agent **cannot** choose token action $\zeta$, and $\zeta$ **cannot** be part of a history $h$ or test $q$.

**Intents and their Outcomes**

An intent $qz \in \mathcal{I} \doteq \mathcal{Q} \times \mathcal{Z}$ is a hypothetical future test $q$, ending in a final extended action $z$:
- Associated with *outcome vectors* $u(qz) \in \mathbb{R}^{|\mathcal{S}|}$, where
$$[u(qz)]_i \doteq \Pr(\bar{o}_q \mid s = i, \bar{a}_q) \, \mathbb{E}\left[R(s', z) \mid s = i, q\right] \,. \tag{8}$$
- A *core* intent set $\mathcal{I}^\dagger \subset \mathcal{I}$ is any maximal set of lin. indep. tests; also satisfies $|\mathcal{I}^\dagger| \leq |\mathcal{S}|$.
- The *outcome matrix* $U \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{I}^\dagger|}$ is the col-stack of core outcomes,
$$[U]_{:i} = u(qz_i), \quad qz_i \in \mathcal{I}^\dagger \,. \tag{9}$$

R-PSRs model *intent rewards* from given *histories*,
$$\begin{aligned} r(qz \mid h) &\doteq \Pr(\bar{o}_q \mid h, \bar{a}_q) R(hq, z) \\ &= b(h)^\top u(qz) \\ &= r(h)^\top m_{qz} \,, \end{aligned} \tag{10}$$
$$r(h) \doteq U^\top b(h) \,. \tag{11}$$
which generalizes both observation probabilities and reward values,
$$R(hq, \zeta) = 1 \implies r(q\zeta \mid h) = p(q \mid h) \,, \tag{12}$$
$$p(\varepsilon \mid h) = 1 \implies r(\varepsilon a \mid h) = R(h, a) \,. \tag{13}$$

**Parameter Vectors:**
- $\{m_{ao\zeta} \mid a \in \mathcal{A}, o \in \mathcal{O}\}$ for the emission of observations.
- $\{m_{\varepsilon a} \mid a \in \mathcal{A}\}$ for the emission of rewards.
- $\{m_{aoqz} \mid a \in \mathcal{A}, o \in \mathcal{O}, qz \in \mathcal{I}^\dagger\}$ for the dynamics.

## Evaluation

- Of 63 POMDPs from Cassandra's POMDP page, 8 ($\approx 13\%$) **cannot** be converted to PSRs.
- In at least 6 of those, the approximate PSR rewards change the task catastrophically.

| | *4x3* | *heaven/hell* | *iff* | *line4-2goals* | *load/unload* | *paint* | *parr* | *stand-tiger* |
|---|---|---|---|---|---|---|---|---|
| $d_\infty$ | 1.0 | 1.0 | 48.93 | 0.6 | 0.5 | 1.33 | 1.0 | 65.0 |
| rel-$d_\infty$ | 1.0 | 1.0 | 0.75 | 0.75 | 0.5 | 1.33 | 0.5 | 0.65 |

Table: PSR reward errors. Absolute $d_\infty \doteq \|R^{(b)} - \tilde{R}^{(b)}\|_\infty$, and relative rel-$d_\infty \doteq \frac{\|R^{(b)} - \tilde{R}^{(b)}\|_\infty}{\|R^{(b)}\|_\infty}$ error measures.

| Domain | Model | Random | POMDP-VI | PSR-VI | R-PSR-VI |
|---|---|---|---|---|---|
| *heaven/hell* | POMDP/R-PSR | $0.0 \pm 0.1$ | $\mathbf{1.4 \pm 0.0}$ | $0.0 \pm 0.0$ | $\mathbf{1.4 \pm 0.0}$ |
| | PSR | $\mathbf{-0.0 \pm 0.0}$ | $\mathbf{-0.0 \pm 0.0}$ | $\mathbf{-0.0 \pm 0.0}$ | $\mathbf{-0.0 \pm 0.0}$ |
| *line4*-2goals | POMDP/R-PSR | $\mathbf{0.4 \pm 0.0}$ | $\mathbf{0.4 \pm 0.0}$ | $\mathbf{0.4 \pm 0.0}$ | $\mathbf{0.4 \pm 0.0}$ |
| | PSR | $\mathbf{4.0 \pm 0.0}$ | $\mathbf{4.0 \pm 0.0}$ | $\mathbf{4.0 \pm 0.0}$ | $\mathbf{4.0 \pm 0.0}$ |
| *load/unload* | POMDP/R-PSR | $1.2 \pm 0.5$ | $\mathbf{4.5 \pm 0.1}$ | $0.6 \pm 0.2$ | $\mathbf{4.5 \pm 0.1}$ |
| | PSR | $4.0 \pm 1.0$ | $2.6 \pm 0.1$ | $\mathbf{9.1 \pm 0.5}$ | $2.6 \pm 0.1$ |
| *paint* | POMDP/R-PSR | $-4.2 \pm 1.4$ | $\mathbf{3.3 \pm 0.3}$ | $0.0 \pm 0.0$ | $\mathbf{3.3 \pm 0.3}$ |
| | PSR | $-3.2 \pm 1.0$ | $1.0 \pm 0.9$ | $\mathbf{3.3 \pm 0.0}$ | $1.0 \pm 1.0$ |
| *parr* | POMDP/R-PSR | $4.3 \pm 1.7$ | $\mathbf{7.1 \pm 0.0}$ | $6.5 \pm 1.8$ | $\mathbf{7.1 \pm 0.0}$ |
| | PSR | $4.3 \pm 0.8$ | $3.6 \pm 0.0$ | $\mathbf{6.3 \pm 0.0}$ | $3.6 \pm 0.0$ |
| *stand-tiger* | POMDP/R-PSR | $-122.3 \pm 43.1$ | $\mathbf{49.2 \pm 23.4}$ | $0.0 \pm 0.0$ | $\mathbf{49.8 \pm 23.2}$ |
| | PSR | $-122.7 \pm 26.4$ | $-151.1 \pm 17.6$ | $\mathbf{0.0 \pm 0.0}$ | $-150.2 \pm 18.0$ |

Table: Policy return estimates for each policy (columns) by each model (rows).