Reconciling Rewards with Predictive State Representations

IJCAI 2021, Montréal

Andrea Baisero    Christopher Amato
{baisero.a,c.amato}@northeastern.edu
Northeastern University, Boston, USA
Motivation and Contributions

Predictive State Representations (PSRs)
- **Stateless** models of non-Markov observation sequences.
- Same observation process as any finite POMDP (*).
- Issues modeling *non-observable* rewards.

Contributions
- Theory of PSR reward modeling accuracy, i.e., which POMDP rewards can be modeled by PSRs?
- Reward-Predictive State Representations (R-PSRs), capable of modeling *non-observable* rewards.
- Value Iteration (VI) for R-PSRs.
- Evaluation on 63 classic domains from literature.
Scope and Notation

Scope:

- **Finite** POMDPs.
- **Linear** PSRs.

Overloaded Notation:

- In POMDPs, as function of history-belief state
  \[ R^{(b)}(h, a) = b(h)^\top \left[ R^{(b)} \right] : a \]
- In PSRs, as function of history-predictive state
  \[ R^{(p)}(h, a) = p(h)^\top \left[ R^{(p)} \right] : a \]
- In R-PSRs, as function of history-reward-predictive state
  \[ R^{(r)}(h, a) = r(h)^\top \left[ R^{(r)} \right] : a \]
Background

Predictive State Representations (PSRs)

PSRs:
- Models of controlled *observation* sequences.
- Same generative process as (finite) POMDPs.
- No latent state $\implies$ easier to learn from experience.
- *Predictive state* $p(h) \in \mathbb{R}^D$ grounded in prediction.

Tests:
- Hypothetical future $q \in Q \triangleq (A \times O)^*$.
- Outcome $u(q) \in \mathbb{R}^{|S|}$, s.t.
  $$[u(q)]_i \triangleq \Pr(\bar{o}_q \mid s = i, \bar{a}_q).$$
- Core tests $Q^\dagger \subset Q$, maximal lin. indep. set, $|Q^\dagger| \leq |S|$.
- Outcome matrix $[U]_{i} = u(q_i), q_i \in Q^\dagger$. 

A. Baisero, C. Amato — “Reconciling Rewards with Predictive State Representations” — IJCAI ’21
Background

Predictive State Representations (PSRs)

Test Probabilities:

\[ p(q \mid h) \doteq \Pr(\bar{o}_q \mid h, \bar{a}_q) \]
\[ = b(h) \top u(q) \]
\[ = p(h) \top m_q \]
\[ p(h) \doteq U \top b(h) \]

Predictive State \( p(h) \):

- Predicts test probabilities \( \Rightarrow \) generates observations.
- Grounded in test probabilities,

\[ [p(h)]_i = p(q_i \mid h), \quad q_i \in Q^\dagger. \]
How to model PSR reward function?

- Assume given reward function $[5, 6, 1]$
  \[ R^{(p)}(h, a) = p(h)^\top \left[ R^{(p)} \right]_{:a}, \]
- Assume observable rewards $[7, 8]$.

Non-Observable Rewards:

- Reward available offline, at training time.
- Agent behavior *not conditioned* on past rewards.
- Reward function $R : S \rightarrow \mathbb{R}$.

Observable Rewards:

- Reward available online, at execution time.
- Agent behavior *conditioned* on past rewards.
- Reward function $R : O \rightarrow \mathbb{R}$. 
Limitations of PSR Reward Models

Open Questions:

• Can $R^{(b)}$ be converted to $R^{(p)}$?
• 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 $R^{(b)}$?
Limitations of PSR Reward Models

Can \( R^{(b)} \) be converted to \( R^{(p)} \)?

**Proposition**

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

**Proof by Example.**

(Degenerate) POMDP with \(|S| \gg 1, |O| = 1\).

\[ |S| \gg 1 \implies |\{ R^{(b)}(h, a) \mid h, a \}| \gg |A| . \]

On the other hand,

\[ |O| = 1 \implies p(h) = (1) \implies |\{ R^{(p)}(h, a) \mid h, a \}| \leq |A| . \]
Limitations of PSR Reward Models

Which $R^{(b)}$ can be converted to $R^{(p)}$?

**Theorem (Accurate Linear PSR Rewards)**

$R^{(b)}$ can be accurately converted to $R^{(p)}$ iff every column of $R^{(b)}$ is lin. dep. on the core outcome vectors (the columns of $U$).

If this accuracy condition is satisfied, $R^{(p)} = U + R^{(b)}$.

**Corollary**

$R^{(p)}$ can be accurately converted to $R^{(b)} = U R^{(p)}$. 
Limitations of PSR Reward Models

Can $R^{(b)}$ be approximated to $R^{(p)}$?

Theorem (Approximate Linear PSR Rewards)

$R^{(b)}$ can be approximated by $R^{(p)} = U + R^{(p)}$, which results in the lowest reward approximation error.

Corollary

$\tilde{R}^{(b)} = U U^T R^{(b)}$ is the reconstructed POMDP-form of the PSR approximation $R^{(p)}$ of the true POMDP rewards $R^{(b)}$.

$\tilde{R}^{(b)} = R^{(b)}$ iff the accuracy condition is satisfied.
Limitations of PSR Reward Models

Does approximate $R^{(p)}$ encode same task as $R^{(b)}$?

Figure: Load/unload domain, with $R^{(b)}$ and $\tilde{R}^{(b)}$ (in parentheses).

Approximate rewards catastrophically change the task.
Token action $\zeta$:

- Unit reward $R(\cdot, \zeta) = 1$.
- Extended action space $\mathcal{Z} = \mathcal{A} \cup \{\zeta\}$.
- Not available to the agent:
  - Agent cannot choose $\zeta$.
  - Environment cannot accept $\zeta$.
  - $\zeta$ cannot be part of a history $h$ or test $q$. 
Reward-Predictive State Representations

R-PSRs:
- Models of controlled observation and reward sequences.
- Same decision process as (finite) POMDPs (this time for real)
- Reward-predictive state $r(h) \in \mathbb{R}^D$ grounded in hypothetical rewards.

Intents and their Rewards:
- Hypothetical future with extended action $qz \in \mathcal{I} \doteq Q \times \mathcal{Z}$.
- Outcome $u(qz) \in \mathbb{R}^{|S|}$, s.t.

$$[u(qz)]_i \doteq \Pr(\bar{o}_q | s = i, \bar{a}_q) \mathbb{E} \left[ R(s', z) | s = i, q \right]. \quad (1)$$

- Core intents $\mathcal{I}^\dagger \subset \mathcal{I}$, maximal lin. indep. set, $|\mathcal{I}^\dagger| \leq |S|$.
- Outcome matrix $[U]_{i} = u(qz_i), qz_i \in \mathcal{I}^\dagger$. 

A. Baisero, C. Amato — “Reconciling Rewards with Predictive State Representations” — IJCAI ’21
Reward-Predictive State Representations

Intent Rewards:

\[ r(qz \mid h) \doteq p(q \mid h)R(hq, z) = b(h)^\top u(qz) = r(h)^\top m_{qz} \]
\[ r(h) \doteq U^\top b(h) \]

Reward-Predictive State \( r(h) \):

- Predicts intent rewards
  \[ \implies \text{generates observations and rewards.} \]
  \[ R(hq, \zeta) = 1 \implies r(q\zeta \mid h) = p(q \mid h) \]
  \[ p(\epsilon \mid h) = 1 \implies r(\epsilon a \mid h) = R(h, a) \]

- Grounded in intent rewards,
  \[ [r(h)]_i = r(qz_i \mid h), \quad qz_i \in \mathcal{I}^\dagger. \]
Evaluation

Value Iteration for R-PSRs (R-PSR-VI)

R-PSR-VI:

- Dynamic programming exact solution method.
- Builds PWLC values $V^*(p(h))$ for increasing horizons. Derives optimal policy tree $\pi^*$.
- Similar derivation to POMDP-VI [3, 4] and PSR-VI [7, 1]. PSR methods can be adapted to R-PSRs.
Evaluation

63 classic domains from Cassandra’s POMDP page [2]:

1. Converted to PSR and R-PSR, check reward accuracy.
2. Run POMDP-VI, PSR-VI, and R-PSR-VI.
3. Let each model evaluate each policy (including Random).

Results:

- 8/63 PSRs ($\approx 13\%$) are not accurate.
- All relative errors are significant.

|                | 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         |

\[
d_\infty \triangleq \| R^{(b)} - \tilde{R}^{(b)} \|_{\infty}
\]

\[
\text{rel-}d_\infty \triangleq \frac{\| R^{(b)} - \tilde{R}^{(b)} \|_{\infty}}{\| R^{(b)} \|_{\infty}}
\]
## Evaluation

### Results:

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

Contributions

- Theory of PSR reward modeling accuracy.
- Reward-Predictive State Representations (R-PSRs).
- Value Iteration (VI) for R-PSRs.
- Evaluation on 63 classic domains from literature.

Evaluation confirms:

- \( \approx 13\% \) (8/63) POMDPs not convertible to PSRs.
- PSR-VI with non-accurate approximate PSRs \( \rightarrow \) Catastrophically sub-optimal policies.
- R-PSRs are accurate reward models.
- R-PSR-VI results in the same optimal policies as POMDP-VI.
References I

B. Boots, S. M. Siddiqi, and G. J. Gordon.
Closing the learning-planning loop with predictive state representations.
The International Journal of Robotics Research, 30(7):954–966, 2011.

A. R. Cassandra.
Tony’s POMDP file repository page, 1999.

A. R. Cassandra, L. P. Kaelbling, and M. L. Littman.
Acting optimally in partially observable stochastic domains.
In Proceedings of the 12th AAAI National Conference on Artificial Intelligence, volume 94, pages 1023–1028, 1994.

A. R. Cassandra, M. L. Littman, and N. L. Zhang.
Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes.
In Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence, pages 54–61, 1997.
References II

M. T. Izadi and D. Precup.
A planning algorithm for predictive state representations.
In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 1520–1521, 2003.

M. T. Izadi and D. Precup.
Point-based planning for predictive state representations.
In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 126–137. Springer, 2008.

M. R. James, S. Singh, and M. L. Littman.
Planning with predictive state representations.
In *Proceedings of the International Conference on Machine Learning and Applications*, pages 304–311. IEEE, 2004.

M. R. James, T. Wessling, and N. Vlassis.
Improving approximate value iteration using memories and predictive state representations.
In *Proceedings of the 21st National Conference on Artificial Intelligence*, volume 1, pages 375–380, 2006.