# Learning Predictive State Representations Using Non-Blind Policies


Michael Bowling Peter McCracken Michael James
James Neufeld Dana Wilkinson

University of Alberta
Toyota Technical Center
University of Waterloo


ICML 2006

UNIVERSITY OF ALBERTA

# Outline

1. What is a PSR?
2. Extracting PSRs from Data.
3. Prediction Estimators:
   Problem and Solution
4. Non-Blind Exploration

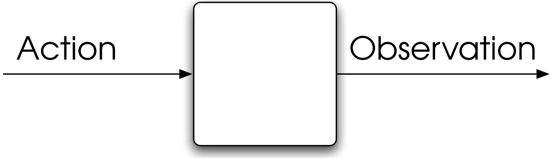
Very Brief Tutorial

Short Punchline

Bonus

UNIVERSITY OF ALBERTA

# Decision Process

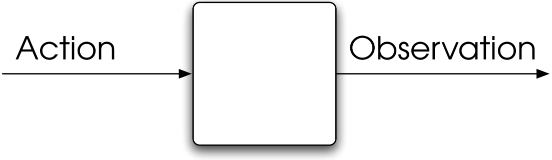



$$a_1, o_1, a_2, o_2, \ldots, a_n, o_n$$

## General Form

$$\Pr(o_{n+1} | a_1, o_1, \ldots, a_n, o_n, a_{n+1})$$

# Decision Process

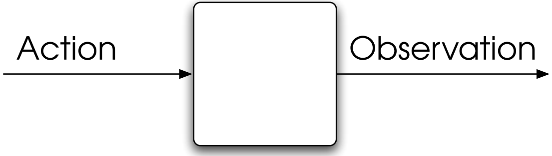



$$a_1, o_1, a_2, o_2, \ldots, a_n, o_n$$

## Markov Decision Process

$$\Pr(o_{n+1} | a_1, o_1, \ldots, a_n, o_n, a_{n+1}) = \Pr(o_{n+1} | o_n, a_{n+1})$$

# Decision Process




$$a_1, o_1, a_2, o_2, \ldots, a_n, o_n$$

## General Form

$$\Pr(o_{n+1} | a_1, o_1, \ldots, a_n, o_n, a_{n+1})$$

# Histories, Tests, and Predictions

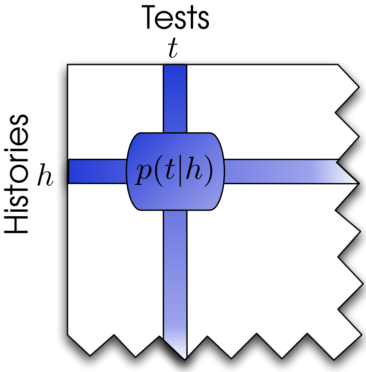
## Notation

| | | |
|---|---|---|
| History($h$) | $a_1, o_1, a_2, o_2, \ldots, a_n, o_n$ | |
| Test($t$) | $a_1, o_1, a_2, o_2, \ldots, a_n, o_n$ | (but in the future) |
| Prediction | $p(t\|h)$ | |

$$
\begin{aligned}
p(a_1, o_1, \ldots, a_n, o_n | h) &\equiv \prod_{i=1}^{n} \Pr(o_i | h a_1, o_1, \ldots, a_i) \\
\pi(a_1, o_1, \ldots, a_n, o_n | h) &\equiv \prod_{i=1}^{n} \Pr(a_i | h a_1, o_1, \ldots, a_{i-1}, o_{i-1}) \\
\Pr(t|h) &= p(t|h)\pi(t|h)
\end{aligned}
$$

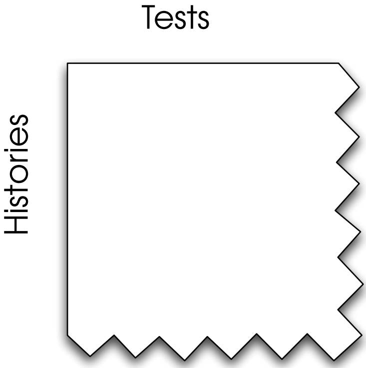
# System Dynamics Matrix

- Countable number of tests and histories.
- Infinite matrix of all predictions.



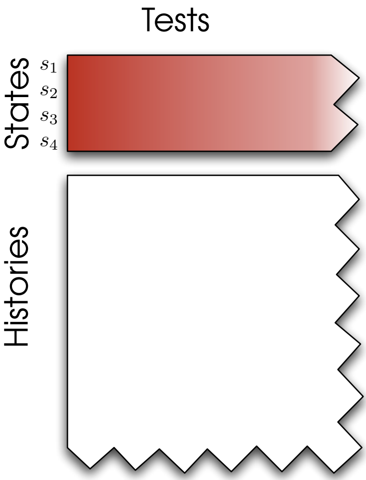
# POMDPs

- Underlying states.



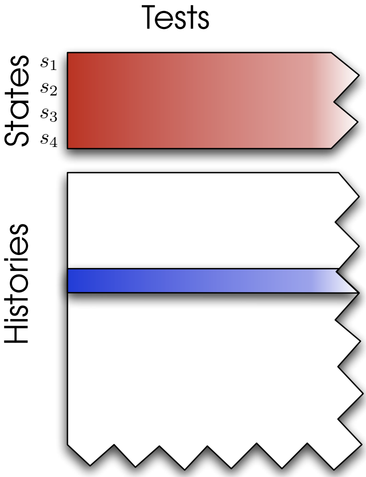
# POMDPs

- Underlying states.



# POMDPs

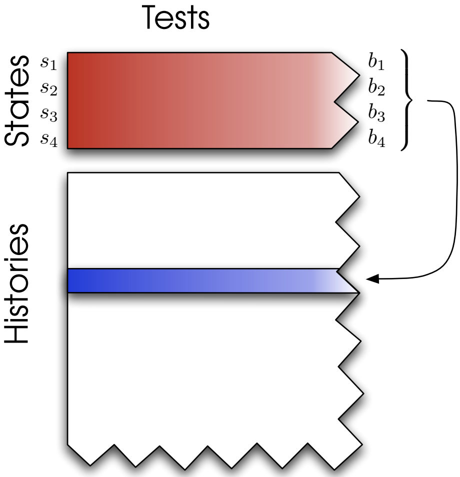
- Underlying states.
- Histories correspond to belief states.




UNIVERSITY OF ALBERTA

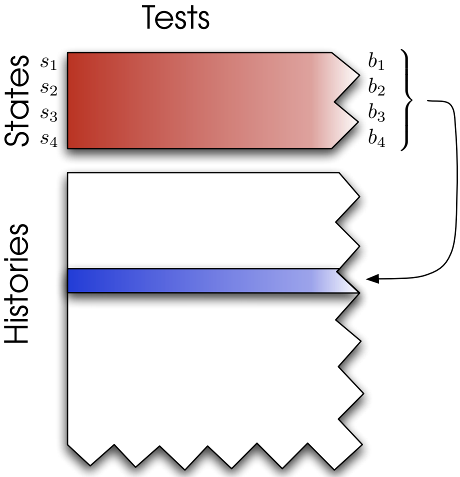
# POMDPs

- Underlying states.
- Histories correspond to belief states.
- History row is a linear combination of state rows.



# POMDPs

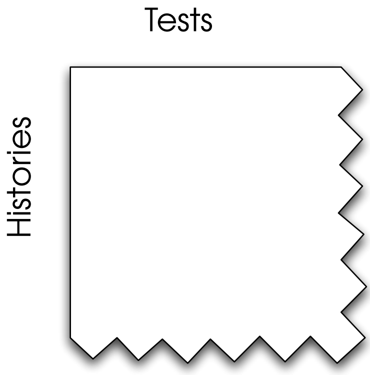
- Underlying states.
- Histories correspond to belief states.
- History row is a linear combination of state rows.
- $\therefore$ rank(SDM) $\leq |\mathcal{S}|$




UNIVERSITY OF ALBERTA

# Predictive State Representations

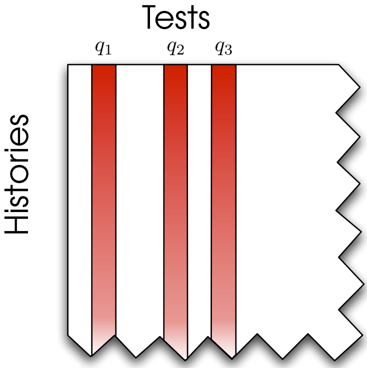
- Find linearly independent tests.



# Predictive State Representations

- Find linearly independent tests.

"Core Tests"
$Q$



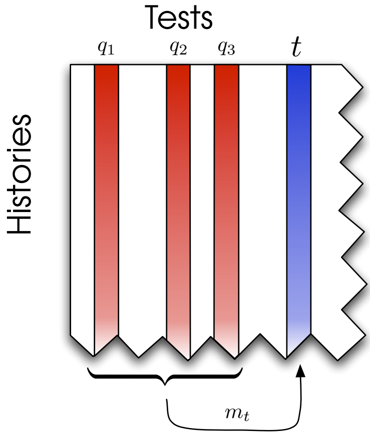
# Predictive State Representations

- Find linearly independent tests.

> "Core Tests"
> $Q$

- Any test is a linear combination of core tests.

$$p(t|h) \quad = \quad p(Q|h)m_t$$



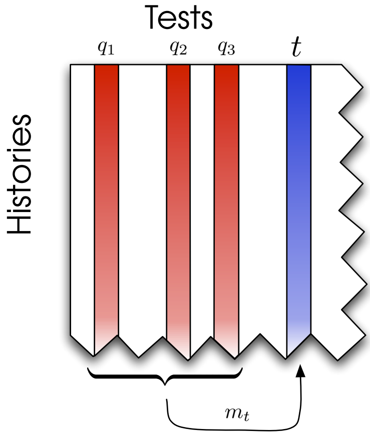
# Predictive State Representations

- Find linearly independent tests.

"Core Tests"
$Q$

- Update predictions:

$$
\begin{aligned}
p(Q|hao) &= \frac{p(aoQ|h)}{p(ao|h)} \\
&= \frac{p(Q|h)M_{aoQ}}{p(Q|h)m_{ao}}
\end{aligned}
$$




UNIVERSITY OF ALBERTA

# Extracting PSRs from Data

# What Data?

$$a_1, o_1, a_2, o_2, \ldots, a_n, o_n$$

# What Data?

$$a_1, o_1, a_2, o_2, \ldots, a_n, o_n$$

- How are actions chosen?
  - Unknown policy.
  - Known policy.
  - Controlled policy.

# What Data?

$$a_1, o_1, a_2, o_2, \ldots, a_n, o_n$$

- How are actions chosen?
    - Unknown policy.
    - Known policy.
    - Controlled policy.

## Note

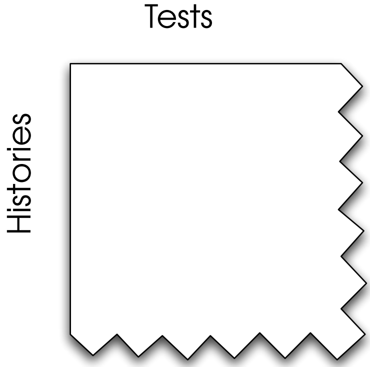
Existing algorithms require a particular control policy. Either:

- Exhaustively trying history-test pairs, or
- Random actions.

# Extracting PSRs from Data

(James & Singh, 2004) (Rosencrantz et al., 2004)
(Wolfe et al., 2005) (Wiewiora, 2005)
(McCracken & Bowling, 2006)

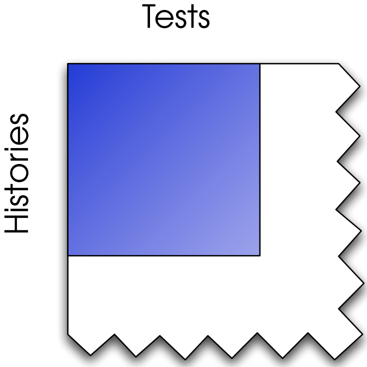
- The common formula:
  - Find core tests.
  - Find update parameters.



# Extracting PSRs from Data

(James & Singh, 2004) (Rosencrantz et al., 2004)
(Wolfe et al., 2005) (Wiewiora, 2005)
(McCracken & Bowling, 2006)

- The common formula:
  - Find core tests.
  - Find update parameters.
  - Estimate part of the system dynamics matrix.

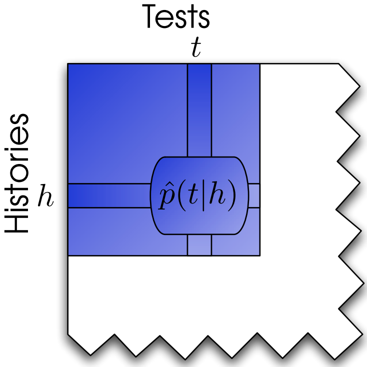



UNIVERSITY OF ALBERTA

# Extracting PSRs from Data

(James & Singh, 2004) (Rosencrantz et al., 2004)
(Wolfe et al., 2005) (Wiewiora, 2005)
(McCracken & Bowling, 2006)

- The common formula:
  - Find core tests.
  - Find update parameters.
  - Estimate part of the system dynamics matrix.
  - Estimate a subset of predictions.

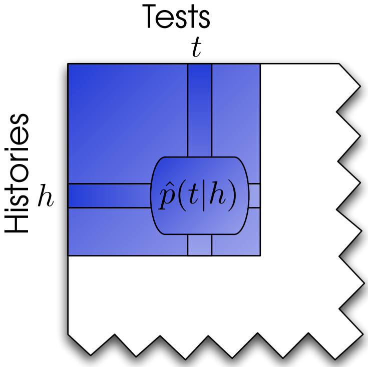


# Extracting PSRs from Data

(James & Singh, 2004) (Rosencrantz et al., 2004)
(Wolfe et al., 2005) (Wiewiora, 2005)
(McCracken & Bowling, 2006)

- The common formula:
  - Find core tests.
  - Find update parameters.
  - Estimate part of the system dynamics matrix.
  - Estimate a subset of predictions.

$$\hat{p}_\bullet(t|h) = \frac{\#ha_1o_1 \ldots a_no_n}{\#ha_1 \ldots a_n}$$



# Problem

$$E\left[\hat{p}_\bullet(t|h)\right] = p(t|h)\frac{\prod_{i=1}^{n}\Pr(a_i|ha_1o_1\ldots a_{i-1}o_{i-1})}{\prod_{i=1}^{n}\Pr(a_i|ha_1\ldots a_{i-1})}$$

## Definition

A policy is **blind** if actions are selected independent of preceeding observations. *I.e.,*

$$\Pr(a_n|a_1, o_1 \ldots a_{n-1}, o_{n-1}) = \Pr(a_n|a_1, \ldots, a_n)$$

## Observation

$\hat{p}_\bullet(t|h)$ is only an unbiased estimator of $p(t|h)$ if $\pi$ is blind.

# What Data?

$$a_1, o_1, a_2, o_2, \ldots, a_n, o_n$$

- How are actions chosen?
  - Unknown policy.
  - Known policy.
  - Controlled policy.

# Prediction Estimators

## Policy is Known

$$\hat{p}_\pi(t|h) = \frac{\#ht}{\#h} \frac{1}{\pi(t|h)}$$

## Policy is Not Known

$$\hat{p}_{\cancel{\pi}}(t|h) = \prod_{i=1}^{n} \frac{\#ha_1o_1 \ldots a_io_i}{\#ha_1o_1 \ldots a_i}$$

## Theorem

$\hat{p}_\pi(t|h)$ *and* $\hat{p}_{\cancel{\pi}}(t|h)$ *are unbiased estimators of* $p(t|h)$.

UNIVERSITY OF ALBERTA

# Exploration

## Goal

Choose actions to reduce error in the estimated system dynamics matrix.

## Approach

- Add intelligent exploration to James & Singh's "reset" algorithm.
- Since $\hat{p}_\pi(t|h)$ is an unbiased estimator, we want to take actions to reduce the variance.
- Solve as an optimization problem.

UNIVERSITY OF ALBERTA

# Estimator Variance

$$
\begin{aligned}
V\left[\hat{p}_\pi(t|h)\middle|\#h = n\right] &= \frac{p(t|h)}{n\pi(t|h)} - \frac{p(t|h)^2}{n} \\
&\leq \frac{1}{4n\pi(t|h)^2}
\end{aligned}
$$

$$
E\left[V\left[\hat{p}_\pi(t|h)\middle|\#h = n\right]\middle|k \text{ trajectories}\right] \leq \frac{1}{4k\, p(h)\pi(h)\pi(t|h)^2}
$$

# Exploration

## Intuition

Find the policy that maximizes the worst-case (over all predictions) bound on the root expected inverse variance.

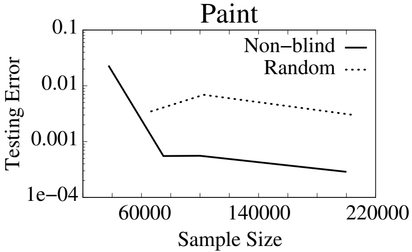
## Optimization Problem

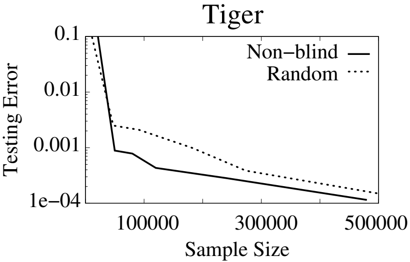
Maximize: $\min_{h,t}\left(\sqrt{v_{i-1}(h,t)^{-1}} + 2\sqrt{k_i\, p(h)}\pi(ht)\right)$

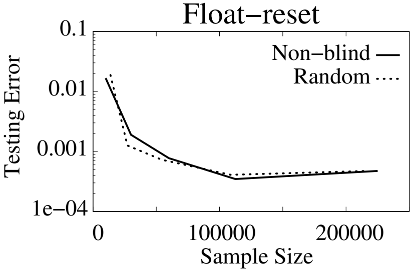
Subject to: **Sequence form** constraints on $\pi(ht)$:

1. $\pi(\phi) = 1$,
2. $\forall h, o \in \mathcal{O} \quad \pi(h) = \sum_a \pi(hao)$, and
3. $\forall h, a \in \mathcal{A}, \{o, o'\} \subseteq \mathcal{O} \quad \pi(hao) = \pi(hao')$.

# Results




UNIVERSITY OF ALBERTA

# Summary

- Contributions
  - Unbiased prediction estimators for non-blind policies.
  - Variance analysis in the case of a known policy.
  - Estimators used in "intelligent" exploration, which was shown can speed learning.
- Future Work
  - Better objective functions for exploration.
  - Investigate when non-blind exploration proves helpful.

Questions?