MAFTEC

Post-Hoc Interpretation of POMDP policies

Geoffrey Laforest, Olivier Buffet,
Alexandre Niveau, Bruno Zanuttini

Caen University

March 2025

Summary

Main goal and ideas

Main goal: redescribe pre-computed POMDP policies to make them more compact and interpretable

Idea: use symbolic features of the form $\mathbf{K}(x)$, $\mathbf{K}(\neg x)$, $\mathbf{K}(x \vee y)$.

Hopes

- Compress history and policies thanks to epistemic states representation
- Explainability
- KBP synthesis (ultimately going back to RL setting)

Partially Observable Markov Decision Process (POMDP)

Partially Observable Markov Decision Process (POMDP)

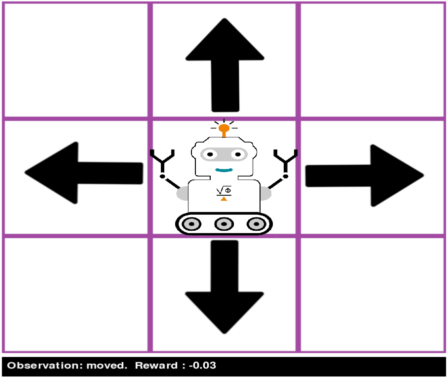
A POMDP is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \boldsymbol{P}, \mathcal{R}, O, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states
- $\mathcal{A}$ is a finite set of actions a
- $\mathcal{O}$ is a finite set of observations
- $\boldsymbol{P}$ is a state transition matrix, s.t. $P_{ss'}^{a} = P\left(S_{t+1} = s' \mid S_t = s, A_t = a\right)$
- $\mathcal{R}$ is a reward function, s.t. $\mathcal{R}_s^a = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$
- O is an observation function
- γ is a discount factor, $\gamma \in [0, 1]$

POMDP - Wumpus example

The **Wumpus** example as a guiding thread

- Deterministic action/observation/reward model
- $A = \{$*move-left, move-right, move-top, move-down, smell*$\}$
- $\mathcal{O} = \{$*wumpus-hit, goal-reached, moved, wumpus, no-wumpus*$\}$



Dealing with POMDPs - History-based approaches

Defintion of History

A history H_t is a sequence of actions and observations

$$H_t = A_0 O_1, \dots, A_{t-1} O_t$$

Defintion of a Belief state

A belief state b_h is a distribution over states conditioned on the history h

$$b_h = \left[P\left(S_t = s_1 \mid H_t = h\right), \dots, P\left(S_t = s_n \mid H_t = h\right)\right]$$

The Belief update

$$b'_h\left(s'\right) = \eta O\left(o \mid s', a\right) \sum_{s \in S} P\left(s' \mid s, a\right) b_h(s)$$

Wumpus trajectory example (1)

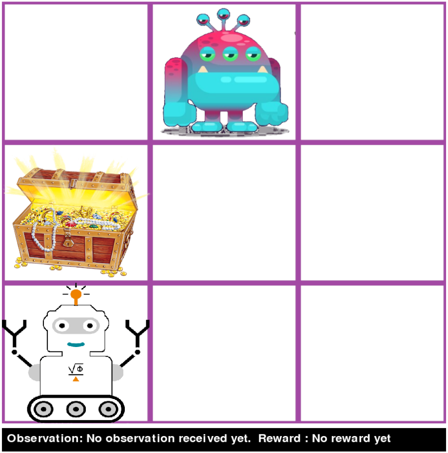


Figure: Real state of the world (unknown to the agent)

Wumpus trajectory example (2)

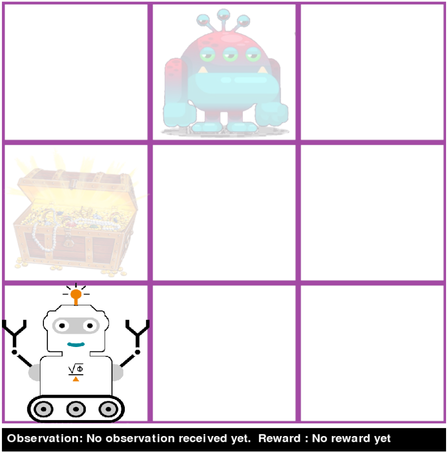


Figure: Initial State. Agent only sees his position

Wumpus trajectory example (3)

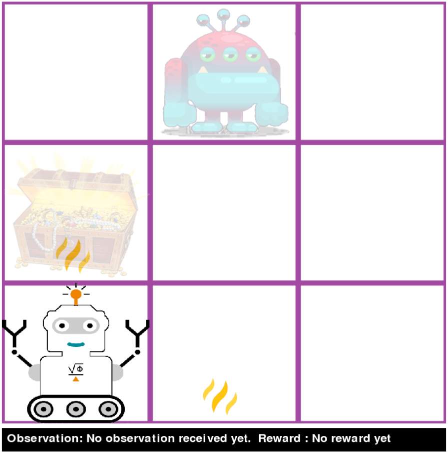


Figure: First action: smell

Wumpus trajectory example (4)

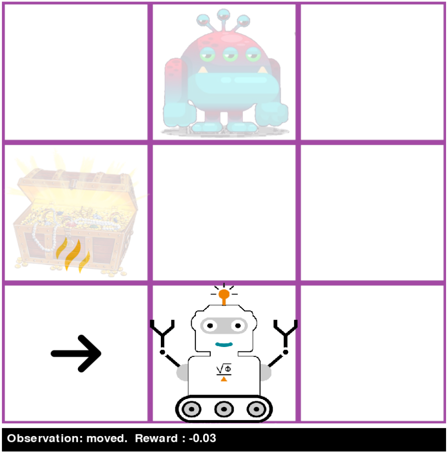


Figure: Move

Wumpus trajectory example (5)

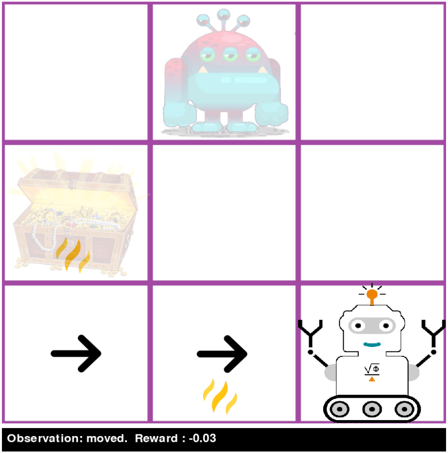


Figure: Move

Wumpus trajectory example (6)

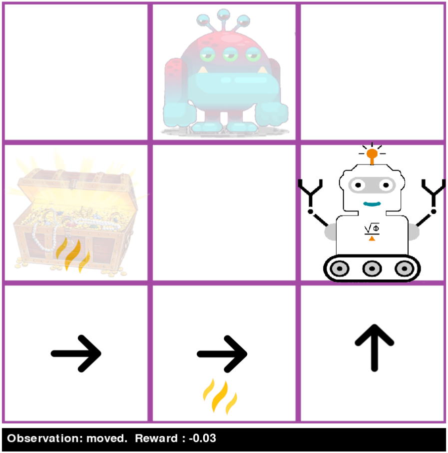


Figure: Move

Wumpus trajectory example (7)

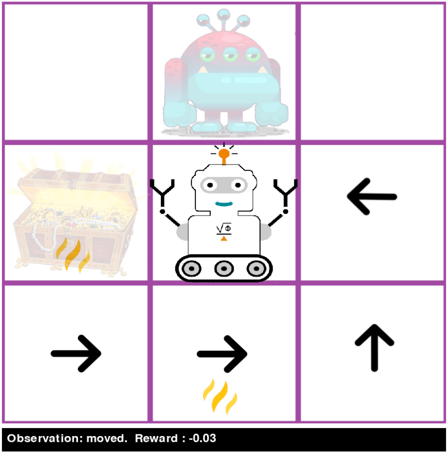


Figure: Move

Wumpus trajectory example (8)

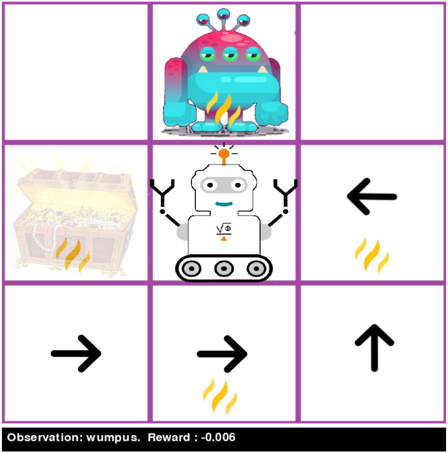


Figure: Smell. Deduction of the Wumpus position

POMDP Policies

Policy: mapping from histories to sets of actions

Typical representations:

- Tree over actions/observations
- Automaton (finite-state controller)

Limitations:

- Huge size: $(|A| \times |O|)^t$
- Poor readability (abstract states)

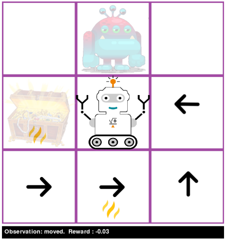
Post-Hoc Interpretation of policies

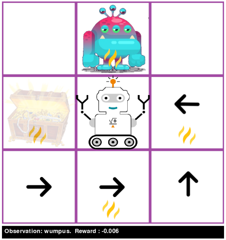
- Idea
  - Transform **belief-based** policies into more interpretable policies defined on the **epistemic state** space
    $\longrightarrow$ post-hoc interpretation of policies
- Post-hoc interpretation of policies
  - Obtain a (near-)optimal policy using an *off-the-shelf* solver
  - Compute epistemic representations of this policy
  - Compare them with an FSC-based representation of the same policy

Epistemic states and features

- Definition of epistemic features
  - Propositional variables / state features $f_i \in F$, f_i predicate on the states of the MDP.
  - Literal ℓ_i: f_i or $\neg f_i$
  - A propositional clause of width w is a disjunction of w literals: $(f_1 \vee \cdots \vee f_p \vee \neg f_{p+1} \vee \cdots \vee \neg f_w)$
  - An epistemic feature of width w is an epistemic atom of the form $\mathbf{K}(\ell_1 \vee \cdots \vee \ell_w)$
- Interpretation:
  - Value of a feature = probability that it is true
  - Epistemic state = value of each feature (embedding)

Feature values and update - width $= 1$





$$a_t = \text{"smell"}$$
$$o_{t+1} = \text{"wumpus-odor"}$$

Update: $\mathbf{v}_t \leftarrow \mathbf{v}_{t+1}$

Feature	Value
$\mathbf{K}(A_{(0,1)})$	0
$\mathbf{K}(A_{(1,1)})$	1
$\mathbf{K}(G_{(1,0)})$	1/4
$\mathbf{K}(W_{(0,1)})$	1/3
$\mathbf{K}(\neg W_{(0,1)})$	2/3

Feature	Value
$\mathbf{K}(A_{(0,1)})$	0
$\mathbf{K}(A_{(1,1)})$	1
$\mathbf{K}(G_{(1,0)})$	1/3
$\mathbf{K}(W_{(0,1)})$	1
$\mathbf{K}(\neg W_{(0,1)})$	0

Post-hoc interpretation of policies - Method

- Method for post-hoc interpretation of policies
  - Obtain a (near-)**optimal policy** using an *off-the-shelf* solver. Solver used for experiments: SARSOP
  - **Project** the policy onto epistemic features
  - Compute **epistemic representations** of this policy
    - Linear representation through MILP solving
    - Decision tree

Projection of policy (1)

- **Setting:** nondeterministic policy $\tilde{p}$
- **Epistemic features:** ordered tuple $\Phi = (\varphi_1, \ldots, \varphi_n)$
- **Projection of belief state:**
  - $\Phi(b) := (\varphi_1(b), \ldots, \varphi_n(b)) \in \mathbb{R}^n$
  - Each component = value (probability) that the corresponding feature is true
- **Projectable policy:**
  - $\tilde{p}$ is projectable onto Φ if there exists a function $\tilde{\pi} \colon \mathbb{R}^n \to \mathcal{P}(\mathcal{A})$
  - such that for all b:

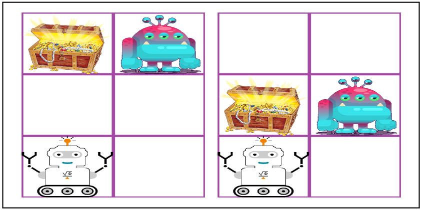
$$\emptyset \subset \tilde{\pi}\,(\Phi(b)) \subseteq \tilde{p}(b)$$

Projection of policy (2)

Example with $w = 1$ and only positive literals.

$\mathbf{K}(A_{(2,0)}) = 1$, $\mathbf{K}(G_{(0,0)}) = \mathbf{K}(G_{(1,0)}) = 0.5$, $\mathbf{K}(W_{(0,1)}) = \mathbf{K}(W_{(1,1)}) = 0.5$

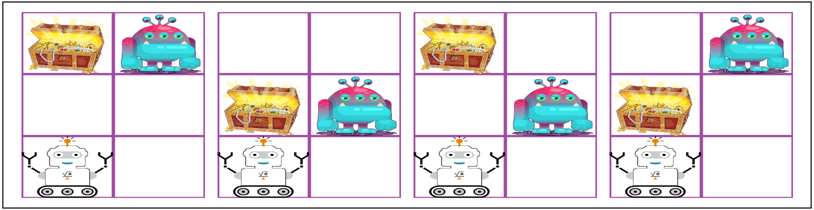
Belief state b_1



- Projection onto epistemic features $\Phi = (\varphi_1, \ldots, \varphi_5)$
- $\Phi(b_1) = (1,\ 0.5,\ 0.5,\ 0.5,\ 0.5) \in \mathbb{R}^5$

Projection of policy (3)

Belief state b_2



- $\Phi(b_2) = (1,\ 0.5,\ 0.5,\ 0.5,\ 0.5) = \Phi(b_1)$
- Same epistemic vector $\Rightarrow$ same input to $\tilde{\pi}$
- If $\tilde{p}(b_1) \cap \tilde{p}(b_2) = \emptyset$, then:

$$\emptyset \subset \tilde{\pi}(\Phi(b_1)) \subseteq \tilde{p}(b_1) \cap \tilde{p}(b_2) = \emptyset$$

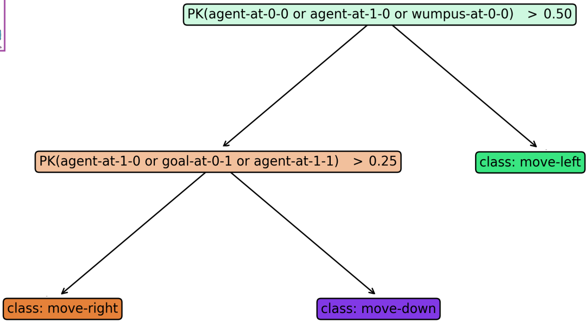
- $\Rightarrow$ $\tilde{p}$ is **not projectable** onto Φ

Learning epistemic representations

- A **supervised learning** framework
  - Set of labeled examples $\{(\Phi(b), \tilde{\pi}^*(\Phi(b))) \mid b \in \mathcal{R}^*(b_0)\}$
  - Learn a classifier, e.g.
    - Linear (MILP).
    - Decision Tree
    - And many others! Logistic regression, neural networks, XGBoost...

Goal: learn to fit the data perfectly (i.e. zero classification error).
Full representation of the policy $\neq$ minimizing generalization error.

Policy – Decision Tree Representation



Non-Deterministic Finite-State Controller (FSC)

Non-Deterministic Finite-State Controller (FSC)

- N is a set of Nodes
- N_0 is a set of distinguisehed initial nodes, $N_0 \subseteq N$
- $act : N \to \mathcal{A}$ maps a node to an action
- $\delta : N \times \mathcal{O} \to \mathcal{P}(N)$ is a transition function

- One can represent an arbitrary policy as a NFSC using a direct generalization of [Grześ et al., 2015]

Results overview

- **small** epistemic **width** is enough.
- Epistemic representations *vs* FSCs: performances **on par**
- Comparing epistemic representations to one another
  - Impact of epistemic width $>$ Positive *vs* negative features *vs* both
  - Larger epistemic width

    $\longrightarrow$ Bigger linear representants $\neq$ sparser trees
  - **Trees** very often **better** than linear models
- Models - especially trees - can help with **features selection**.

Results on Mastermind - Size

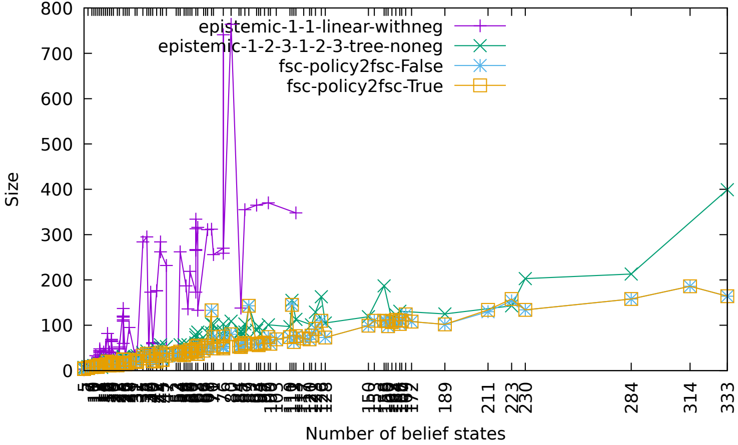


Figure: Mastermind: size results

Results on Minesweeper - Size

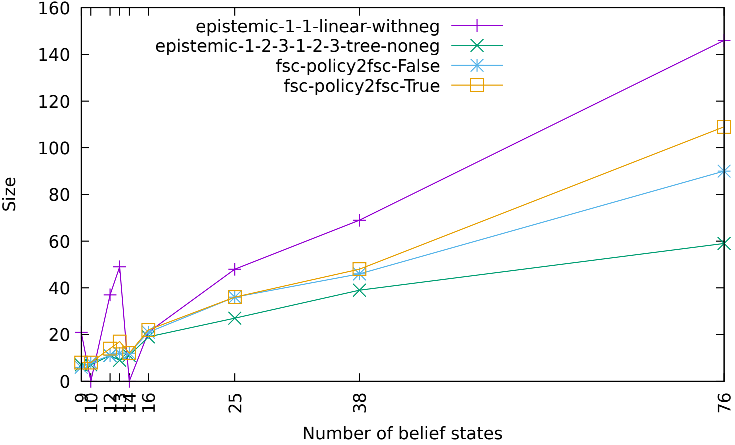


Figure: Minesweeper: size results

Comparisons of Epistemic Representations (3)

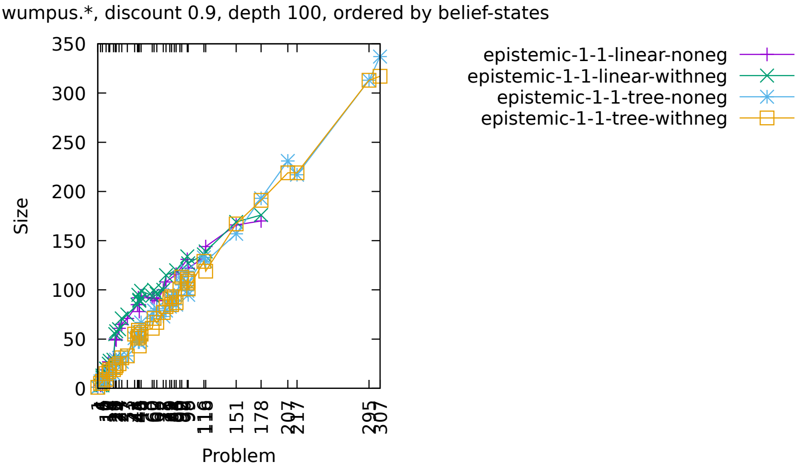


Figure: Size results on wumpus for width=1

Comparisons of Epistemic Representations (4)

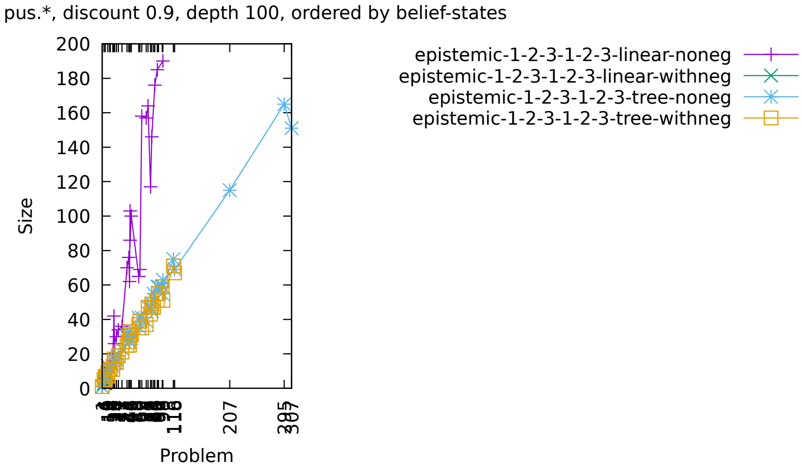


Figure: Size results on wumpus for widths=1, 2, 3

Limitations and future work

- Limitations
  - Number of features **grows quickly** $\Rightarrow$ learning high-dimensional manifolds is **cursed!**
  - Results are only for **small instances** of our benchmarks.
  - Features are built in a systematic way but may not be the most informative nor interpretable.
- Future directions
  - **Factored models** to scale our experiments.
  - Feature selection or synthesis ?
  - Producing **factual** / **counterfactual** local explanations.

    Use decision tree splitting rules to make important state explanations.

The End

Thank you

Grześ, M., Poupart, P., Yang, X., and Hoey, J. (2015).
Energy efficient execution of pomdp policies.
IEEE Transactions on Cybernetics, 45(11):2484–2497.