Efficient Solutions to Factored MDP with Imprecise Transition Probabilities

Karina Valdivia Delgado [1,2]
Scott Sanner [2]
Leliane Nunes de Barros [1]
Fabio G. Cozman [1]

1. University of Sao Paulo
2. NICTA & ANU

Outline

1. Markov Decision Processes (MDPs)

Outline

1. Markov Decision Processes (MDPs)

2. MDPs with **Imprecise Probabilities** (MDP**IPs**)

Outline

1. Markov Decision Processes (MDPs)
2. MDPs with **Imprecise Probabilities** (MDP**IPs**)
3. Representing **Factored** MDPIPs

Outline

1. Markov Decision Processes (MDPs)
2. MDPs with **Imprecise Probabilities** (MDP**IPs**)
3. Representing **Factored** MDPIPs
4. Efficient **Solutions** for Factored MDPIPs

Outline

1. Markov Decision Processes (MDPs)
2. MDPs with **Imprecise Probabilities** (MDP**IPs**)
3. Representing **Factored** MDPIPs
4. Efficient **Solutions** for Factored MDPIPs
5. Summary

MDP - Formal model

MDP is defined by $\mathcal{M} = (S, A, R, P, \gamma)$:

- S is a set of states.
- A is a set of actions.
- $R(s, a)$ is a reward function.
- $P(s'|s, a)$ are transition probabilities $\forall s, s' \in S$ and $\forall a \in A$
- γ discount factor

MDP - Formal model

MDP is defined by $\mathcal{M} = (S, A, R, P, \gamma)$:

- S is a set of states.
- A is a set of actions.
- $R(s, a)$ is a reward function.
- $P(s'|s, a)$ are transition probabilities $\forall s, s' \in S$ and $\forall a \in A$
- γ discount factor

How to act in an MDP?

- Policy $\pi : S \to A$
- But what criteria to optimize?

MDP - Value Function

- Define value of a policy π:

$$V_\pi(s) = E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t | s = s_0\right]$$

MDP - Value Function

- Define value of a policy π:

$$V_\pi(s) = E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t | s = s_0\right]$$

- MDP optimal policy π^*:

$$V_{\pi^*}(s) \geq V_{\pi'}(s) \;\; \forall \pi', s$$

MDP - Value Iteration

- Given optimal $t-1$-stage-to-go value function
  - How to act optimally with t decisions?

MDP - Value Iteration

- Given optimal $t-1$-stage-to-go value function
  - How to act optimally with t decisions?

$$Q^t(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{t-1}(s')$$

MDP - Value Iteration

- Given optimal $t-1$-stage-to-go value function
  - How to act optimally with t decisions?

$$Q^t(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{t-1}(s')$$

$$V^t(s) = \max_{a \in A} Q^t(s,a)$$

MDP - Value Iteration

- Given optimal $t-1$-stage-to-go value function
  - How to act optimally with t decisions?

$$Q^t(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{t-1}(s')$$

$$V^t(s) = \max_{a \in A} Q^t(s,a)$$

  - At $t = \infty$, convergence:

$$\lim_{t \to \infty} \max_{s} |V^t(s) - V^{t-1}(s)| = 0$$

MDPIP - Introduction

Why imprecision in transition probabilities?

MDPIP - Introduction

Why imprecision in transition probabilities?

- Imprecise or conflicting **expert elicitations**

MDPIP - Introduction

Why imprecision in transition probabilities?

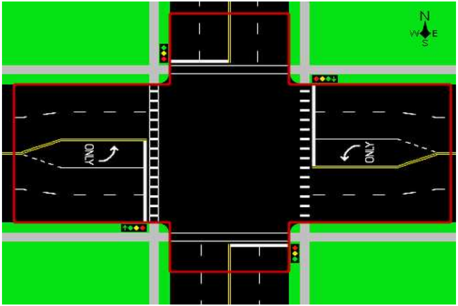
- Imprecise or conflicting **expert elicitations**
- **Insufficient data** to estimate precise transition models

MDPIP - Introduction

Why imprecision in transition probabilities?

- Imprecise or conflicting **expert elicitations**
- **Insufficient data** to estimate precise transition models
- **Non-stationary** (but bounded) transition probabilities

MDPIP - Introduction



Example: **non-stationarity** in **traffic arrival & turn probabilities**

- fluctuate each hour of the day
- drift over time (probabilities measured every 2-3 years)

MDPIP - Introduction

- MDPIP is defined by $\mathcal{M} = (S, A, R, K, \gamma)$

MDPIP - Introduction

- MDPIP is defined by $\mathcal{M} = (S, A, R, K, \gamma)$
- What's new?
  - **Credal set** $K = \{P\}$ of possible transition probabilities

MDPIP Value Iteration

- Be as robust as possible, given uncertainty:

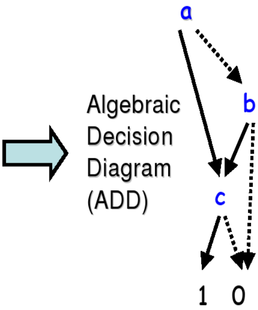
$$V^t(s) = \max_{a \in A} \min_{P \in K} \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{t-1}(s') \right\}$$

Representing large MDPIPs

- **Compact representation:**
  - Factored state and action variables
  - Decision diagrams (DDs) for reward and transition

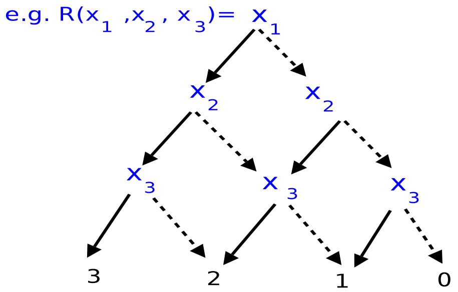
Algebraic Decision Diagrams (ADDs)

a	b	c	F(a,b,c)
0	0	0	0.00
0	0	1	0.00
0	1	0	0.00
0	1	1	1.00
1	0	0	0.00
1	0	1	1.00
1	1	0	0.00
1	1	1	1.00



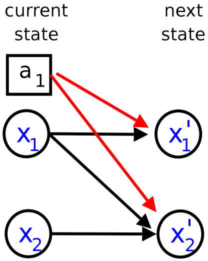
Compact representation for R: ADD

Reward R as an ADD:



Compact representation for K: DCN

Dynamic Credal Networks (DCN) [Cozman00] (DBN extension)



a)

CPT action a_1

x_1	$x_1' = 0$
F	p_{11}
T	p_{12}

$$0 \leqslant p_{11} \leqslant 0.6$$

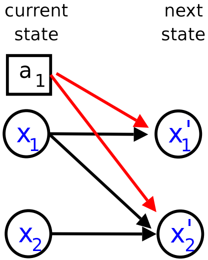
$$0 \leqslant p_{12} \leqslant 1$$

$$0.3 \leqslant p_{11} + p_{12} \leqslant 0.8$$

b)

Compact representation for K: DCN

Dynamic Credal Networks (DCN) [Cozman00] (DBN extension)



a)

CPT action a_1

x_1	$x_1' = 0$
F	p_{11}
T	p_{12}

$0 \leq p_{11} \leq 0.6$

$0 \leq p_{12} \leq 1$

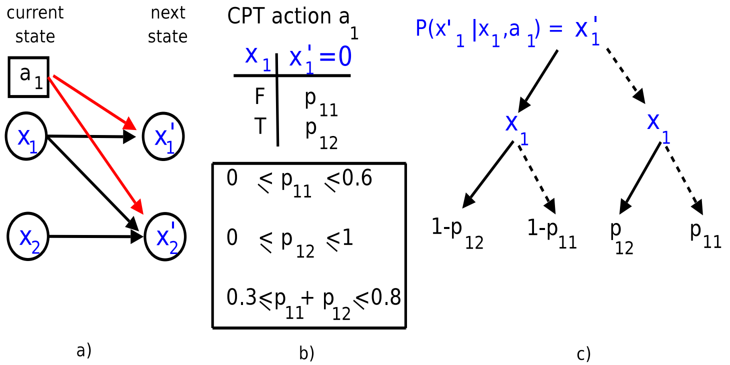
$0.3 \leq p_{11} + p_{12} \leq 0.8$

b)

Note: **product of CPTs** yields **polynomial** expressions ($p_1^2 p_2$)
$\Longrightarrow$ restricted to **multilinear** ($p_1 p_2 p_3$) if CPTs do not share p_i

Compact representation for K: DCN + PADD

CPTs represented as **Parameterized ADDs (PADDs)**



Parameterized ADD

Extending ADDs to PADDs:

- **Equality testing** on leaf expressions

Parameterized ADD

Extending ADDs to PADDs:

- **Equality testing** on leaf expressions
- **Binary operations** on leaves:
  - Easy to do +, *, - on algebraic expressions

Parameterized ADD

Extending ADDs to PADDs:

- **Equality testing** on leaf expressions
- **Binary operations** on leaves:
  - Easy to do +, *, - on algebraic expressions
  - Division does not yield a PADD (fractional leaves)

Parameterized ADD

Extending ADDs to PADDs:

- **Equality testing** on leaf expressions
- **Binary operations** on leaves:
  - Easy to do +, *, - on algebraic expressions
  - Division does not yield a PADD (fractional leaves)
  - max and min do not yield PADDs (inequality decision nodes)
    - but no need to perform these for factored MDPIPs!

Solving large MDPIPs

- **Compact representation:**
  - Factored state and action variables
  - Decision diagrams (DDs) for reward and transition

Solving large MDPIPs

- **Compact representation:**
  - Factored state and action variables
  - Decision diagrams (DDs) for reward and transition
- **Compact and efficient solutions:**
  - *SPUDD-IP*: Factored value iteration with DDs

Solving large MDPIPs

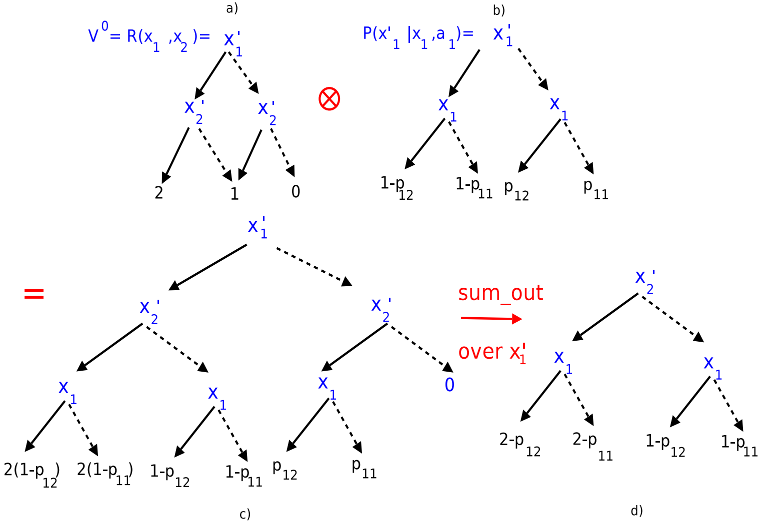
- **Compact representation:**
  - Factored state and action variables
  - Decision diagrams (DDs) for reward and transition
- **Compact and efficient solutions:**
  - *SPUDD-IP*: Factored value iteration with DDs
- **Bounded error approximations:**
  - *APRICODD-IP*: Naive pruning approach
  - *OBJECTIVE-IP*: Pruning where it counts

Factored MDPIP Value Iteration: SPUDD-IP

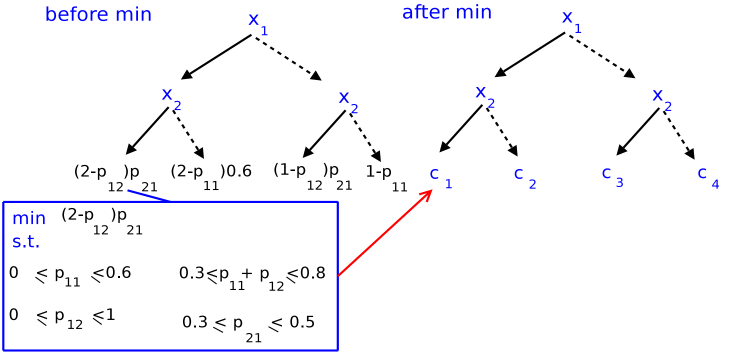
***SPUDD-IP*: Extend SPUDD [HoeyStHuBout99] to MDPIPs**

$$V^t(\vec{x}) = \max_{a \in A} \left\{ R(\vec{x}, a) \oplus \gamma \min_{\vec{p}} \boxed{\sum_{\vec{x}'} \bigotimes_{i=1}^{n} P(x_i' | pa_a(x_i'), a, \vec{p}) V^{t-1}(\vec{x}')} \right\}$$

First iteration:

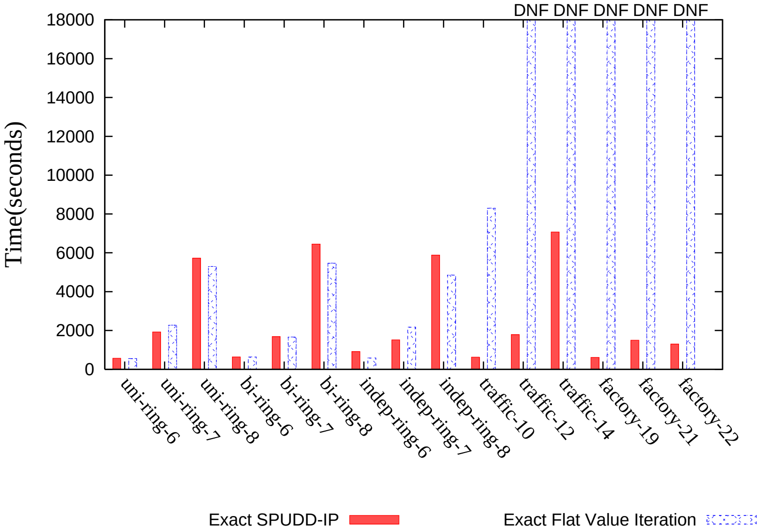


First iteration (continued):



$$V^t(\vec{x}) = \max_{a\in A}\left\{ R(\vec{x},a) \oplus \gamma \min_{\vec{p}} \boxed{\sum_{x'_i(i\neq 1)} \bigotimes_{i=1(i\neq 1)}^{n} P(x'_i|pa_a(X'_i),a,\vec{p}) \sum_{x'_1} P(x'_1|pa_a(X'_1),a,\vec{p}) V^{t-1}(\vec{x}')} \right\}$$

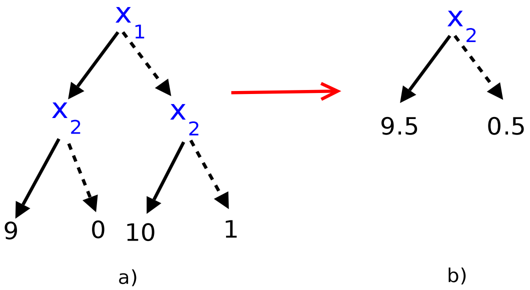
SPUDD-IP vs. Flat MDPIP Value Iteration



Approximate solution for MDPIPs: APRICODD-IP

***APRICODD-IP*: APRICODD [StHoeyBout00] for MDPIPs**

- After each iteration, prune the values that are similar
- Achieves a **bounded** approximate solution



Approximate solution for MDPIPs: OBJECTIVE-IP

APRICODD-IP **sucks** (results in a moment)

Approximate solution for MDPIPs: OBJECTIVE-IP

APRICODD-IP **sucks** (results in a moment)

- Why? Because nonlinear solution calls dominate time

Approximate solution for MDPIPs: OBJECTIVE-IP

APRICODD-IP **sucks** (results in a moment)

- Why? Because nonlinear solution calls dominate time
- APRICODD prunes value ADD

Approximate solution for MDPIPs: OBJECTIVE-IP

APRICODD-IP **sucks** (results in a moment)

- Why? Because nonlinear solution calls dominate time
- APRICODD prunes value ADD
  - Need to prune PADDs *before* nonlinear solver call

Approximate solution for MDPIPs: OBJECTIVE-IP

***APRICODD-IP* sucks** (results in a moment)

- Why? Because nonlinear solution calls dominate time
- APRICODD prunes value ADD
  - Need to prune PADDs *before* nonlinear solver call

***Objective-IP*: approximate PADD objective instead**

Approximate solution for MDPIPs: OBJECTIVE-IP

***APRICODD-IP* sucks** (results in a moment)

- Why? Because nonlinear solution calls dominate time
- APRICODD prunes value ADD
  - Need to prune PADDs *before* nonlinear solver call

***Objective-IP*: approximate PADD objective instead**

- PADD approximation techniques (see paper for alg/theorem)

Approximate solution for MDPIPs: OBJECTIVE-IP

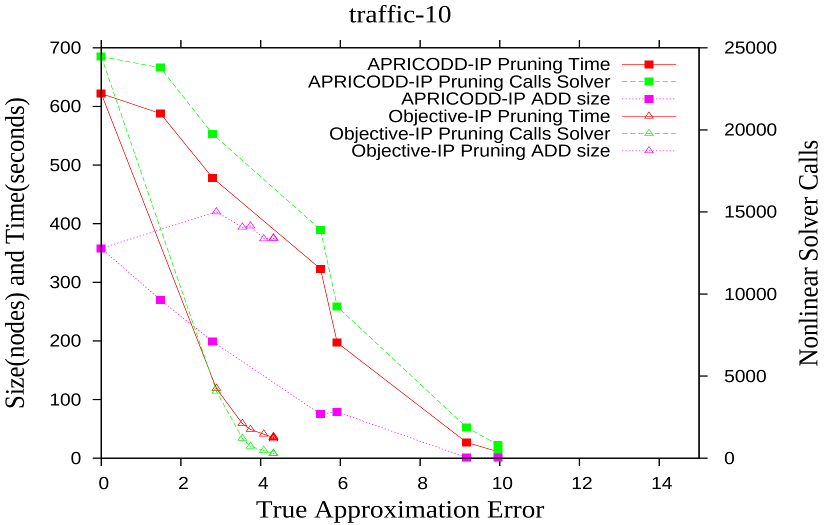
APRICODD-IP **sucks** (results in a moment)

- Why? Because nonlinear solution calls dominate time
- APRICODD prunes value ADD
    - Need to prune PADDs *before* nonlinear solver call

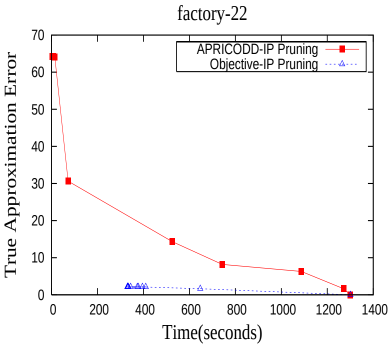
***Objective-IP*: approximate PADD objective instead**

- PADD approximation techniques (see paper for alg/theorem)
- Produces **bounded** approximately optimal solution

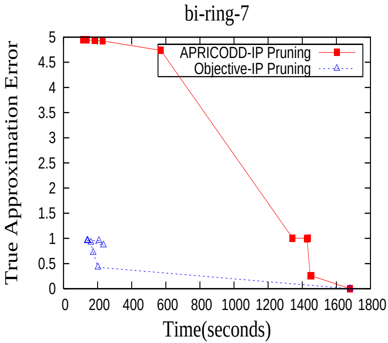
Objective-IP vs. APRICODD-IP

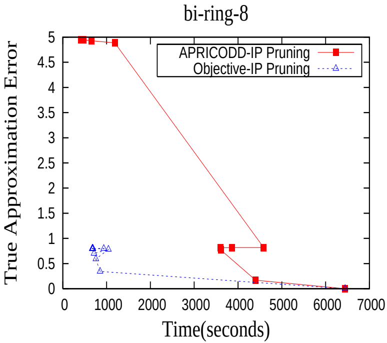


Objective-IP vs. APRICODD-IP



Objective-IP vs. APRICODD-IP





Related Work

- **Bounded Parameter MDPs** (Givan, Leach, & Dean, 2000)
  - Flat transition probabilities are interval bounded

Related Work

- **Bounded Parameter MDPs** (Givan, Leach, & Dean, 2000)
  - Flat transition probabilities are interval bounded
- **Markov Decision Process with Set-valued Transitions** (Trevizan, Cozman, & de Barros, 2007)
  - Flat finite belief set of transition probabilities

Related Work

- **Bounded Parameter MDPs** (Givan, Leach, & Dean, 2000)
  - Flat transition probabilities are interval bounded
- **Markov Decision Process with Set-valued Transitions** (Trevizan, Cozman, & de Barros, 2007)
  - Flat finite belief set of transition probabilities
- **Zero-sum Alternating Markov Games** (Littman, 1994)
  - A subset of flat MDPIPs
  - More computationally feasible when Nature's effects can be modeled as explicit action

Related Work

- **Bounded Parameter MDPs** (Givan, Leach, & Dean, 2000)
  - Flat transition probabilities are interval bounded
- **Markov Decision Process with Set-valued Transitions** (Trevizan, Cozman, & de Barros, 2007)
  - Flat finite belief set of transition probabilities
- **Zero-sum Alternating Markov Games** (Littman, 1994)
  - A subset of flat MDPIPs
  - More computationally feasible when Nature's effects can be modeled as explicit action

$\Longrightarrow$ **Above are strict subsets of Factored MDPIPs**

- Transition probabilities are polynomial expressions of linearly constrained variables (i.e., from multiplying CPTs in DCN)

Summary and Future Work

- **Contributions**
  - Parameterized ADDs

Summary and Future Work

- **Contributions**
  - Parameterized ADDs
  - Factored MDPIP Value Iteration:
    - *SPUDD-IP*

Summary and Future Work

- **Contributions**
  - Parameterized ADDs
  - Factored MDPIP Value Iteration:
    - *SPUDD-IP*
    - Up to 2 orders improvement over flat VI!

Summary and Future Work

- **Contributions**
  - Parameterized ADDs
  - Factored MDPIP Value Iteration:
    - *SPUDD-IP*
    - Up to 2 orders improvement over flat VI!
  - Factored MDPIP Approximate Value Iteration
    - *APRICODD-IP*: extension of previous ideas

Summary and Future Work

- **Contributions**
  - Parameterized ADDs
  - Factored MDPIP Value Iteration:
    - *SPUDD-IP*
    - Up to 2 orders improvement over flat VI!
  - Factored MDPIP Approximate Value Iteration
    - *APRICODD-IP*: extension of previous ideas
    - *OBJECTIVE-IP*: pruning where it counts, lower error & faster!

Summary and Future Work

- **Contributions**
  - Parameterized ADDs
  - Factored MDPIP Value Iteration:
    - *SPUDD-IP*
    - Up to 2 orders improvement over flat VI!
  - Factored MDPIP Approximate Value Iteration
    - *APRICODD-IP*: extension of previous ideas
    - *OBJECTIVE-IP*: pruning where it counts, lower error & faster!
- **Future Work**

Summary and Future Work

- **Contributions**
  - Parameterized ADDs
  - Factored MDPIP Value Iteration:
    - *SPUDD-IP*
    - Up to 2 orders improvement over flat VI!
  - Factored MDPIP Approximate Value Iteration
    - *APRICODD-IP*: extension of previous ideas
    - *OBJECTIVE-IP*: pruning where it counts, lower error & faster!
- **Future Work**
  - More targeted approximations, more cache reuse

Summary and Future Work

- **Contributions**
  - Parameterized ADDs
  - Factored MDPIP Value Iteration:
    - *SPUDD-IP*
    - Up to 2 orders improvement over flat VI!
  - Factored MDPIP Approximate Value Iteration
    - *APRICODD-IP*: extension of previous ideas
    - *OBJECTIVE-IP*: pruning where it counts, lower error & faster!
- **Future Work**
  - More targeted approximations, more cache reuse
  - Trading off robustness with average-case