The Complexity of Graph-Based Reductions for Reachability in MDPs

Stéphane Le Roux and Guillermo A. Pérez

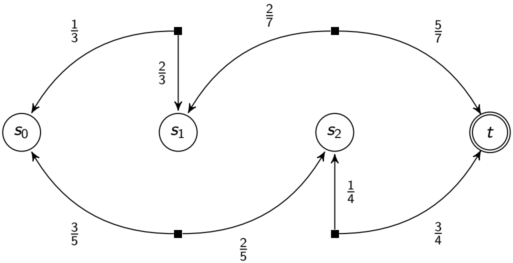
Technische Universität Darmstadt
Université libre de Bruxelles

FoSSaCS 2018

Markov decision processes

Markov decision processes

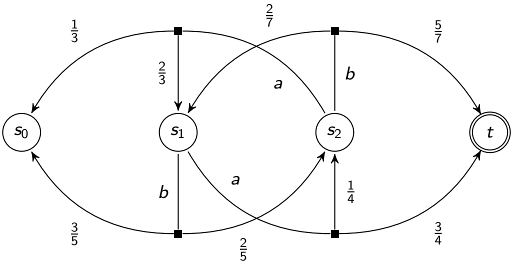
An MDP is a tuple $\mathcal{M} = (S, A, \delta, T)$ with $\delta : S \times A \to \mathbb{D}(S)$.



Markov decision processes

Markov decision processes

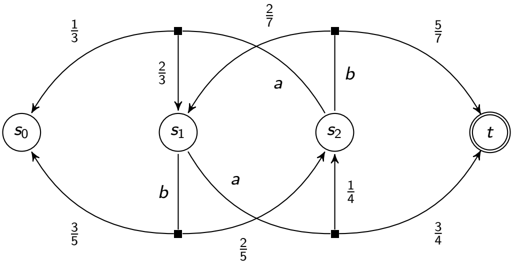
An MDP is a tuple $\mathcal{M} = (S, A, \delta, T)$ with $\delta : S \times A \to \mathbb{D}(S)$.



Markov decision processes

Markov decision processes

An MDP is a tuple $\mathcal{M} = (S, A, \delta, T)$ with $\delta : S \times A \to \mathbb{D}(S)$.



- A (memoryless deterministic) strategy $\sigma : S \to A$, is a way to choose actions from every state.
- An MDP restricted to transitions consistent with a given strategy is a Markov chain.

Reachability in MDPs

Consider an MDP $\mathcal{M} = (S, A, \delta, T)$.

Reachability probability value

For $s \in S$, we denote by $\mathbb{P}^{s}_{\mathcal{M}^\sigma}[\Diamond T]$ the probability of eventually reaching T in $\mathcal{M}$ from s under σ.

Reachability in MDPs

Consider an MDP $\mathcal{M} = (S, A, \delta, T)$.

Reachability probability value

For $s \in S$, we denote by $\mathbb{P}^s_{\mathcal{M}^\sigma}[\Diamond T]$ the probability of eventually reaching T in $\mathcal{M}$ from s under σ.

Maximal reachability probability value

We are interested in maximizing the probability of eventually reaching T (with a memoryless deterministic strategy)

$$\mathbf{Val}_\delta(s) := \max_\sigma \mathbb{P}^s_{\mathcal{M}^\sigma}[\Diamond T].$$

Reachability in MDPs

Consider an MDP $\mathcal{M} = (S, A, \delta, T)$.

Reachability probability value

For $s \in S$, we denote by $\mathbb{P}^s_{\mathcal{M}^\sigma}[\Diamond T]$ the probability of eventually reaching T in $\mathcal{M}$ from s under σ.

Maximal reachability probability value

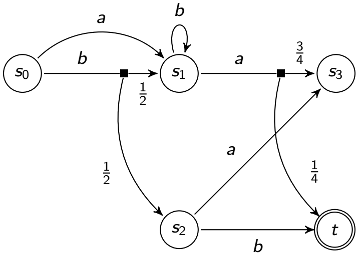
We are interested in maximizing the probability of eventually reaching T (with a memoryless deterministic strategy)

$$\mathbf{Val}_\delta(s) := \max_\sigma \mathbb{P}^s_{\mathcal{M}^\sigma}[\Diamond T].$$

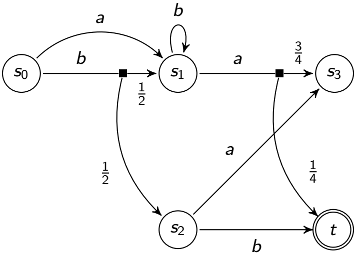
Theorem (Filar, Vrieze 97; Puterman 94)

Given $\mathcal{M}$, a state s, and $\tau \in \mathbb{Q}$, determining whether $\mathbf{Val}_\delta(s) \geq \tau$ is decidable in polynomial time (via an encoding into a linear program).

Example 1

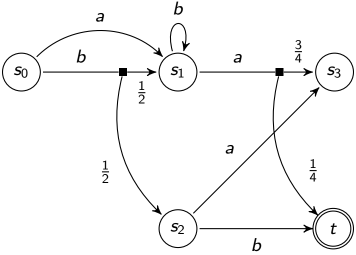


Example 1



Should play $\sigma : s_0 \mapsto b, s_1 \mapsto a, s_2 \mapsto b$

Example 1



Should play $\sigma : s_0 \mapsto b, s_1 \mapsto a, s_2 \mapsto b$
Since $\mathbf{Val}(s_2) = 1$ and $\mathbf{Val}(s_1) = \frac{1}{4}$,

$$\mathbf{Val}(s_0) \geq \frac{1}{8} + \frac{1}{2} = \frac{5}{8}$$

Motivation: why study (redux for) reachability in MDPs?

Verification

Markov decision processes are perfect models for systems with stochastic and non-deterministic components. Verifying safety and liveness properties in MDPs reduces to reachability analysis.

Motivation: why study (redux for) reachability in MDPs?

Verification

Markov decision processes are perfect models for systems with stochastic and non-deterministic components. Verifying safety and liveness properties in MDPs reduces to reachability analysis.

- The running-time of value iteration is inversely proportional to the smallest transition probability value.

Motivation: why study (redux for) reachability in MDPs?

Verification

Markov decision processes are perfect models for systems with stochastic and non-deterministic components. Verifying safety and liveness properties in MDPs reduces to reachability analysis.

- The running-time of value iteration is inversely proportional to the smallest transition probability value.

Artificial intelligence

In reinforcement learning, MDPs are not known a priori: transition probability values are learned within a desired confidence interval.

Motivation: why study (redux for) reachability in MDPs?

Verification

Markov decision processes are perfect models for systems with stochastic and non-deterministic components. Verifying safety and liveness properties in MDPs reduces to reachability analysis.

- The running-time of value iteration is inversely proportional to the smallest transition probability value.

Artificial intelligence

In reinforcement learning, MDPs are not known a priori: transition probability values are learned within a desired confidence interval.

- More unknown transitions probabilities translates into longer learning times.

de Alfaro's end components

Consider an MDP $\mathcal{M} = (S, A, \delta, T)$.

End components

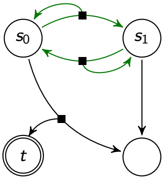
$Q \subseteq S$ and $\alpha : S \to \mathcal{P}(A)$ are an end component if playing actions allowed by α ensures staying in Q and the induced digraph in $\mathcal{M}$ is strongly connected

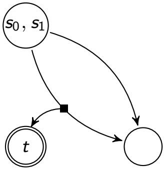
de Alfaro's end components

Consider an MDP $\mathcal{M} = (S, A, \delta, T)$.

End components

$Q \subseteq S$ and $\alpha : S \to \mathcal{P}(A)$ are an end component if playing actions allowed by α ensures staying in Q and the induced digraph in $\mathcal{M}$ is strongly connected





de Alfaro's end components

Consider an MDP $\mathcal{M} = (S, A, \delta, T)$.

End components

$Q \subseteq S$ and $\alpha : S \to \mathcal{P}(A)$ are an end component if playing actions allowed by α ensures staying in Q and the induced digraph in $\mathcal{M}$ is strongly connected

They are awesome!

All states in an end component have the same value (for all same-support distributions); and they can be "collapsed". Maximal end components are computable in polynomial time!

More graph-based reductions

Efficient reductions [Ciesinski, Baier, Größer, Klein 08]

Before value iteration, one can compute in polynomial time

- extremal-probability states,
- essential states [D'Argenio, Jeannet, Jensen, Larsen 02],
- maximal end components.

More graph-based reductions

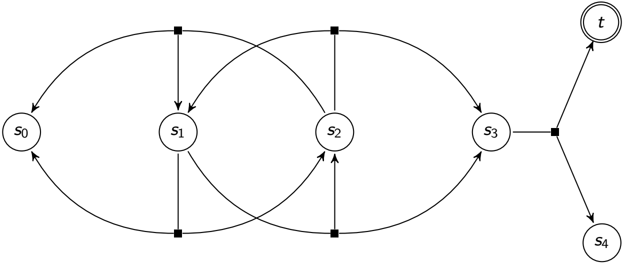
Efficient reductions [Ciesinski, Baier, Größer, Klein 08]

Before value iteration, one can compute in polynomial time

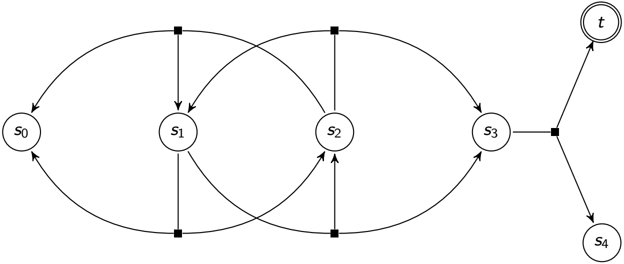
- extremal-probability states,
- essential states [D'Argenio, Jeannet, Jensen, Larsen 02],
- maximal end components.

Can we do better/more?

Other components

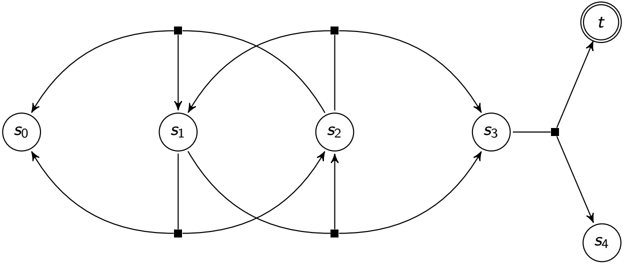


Other components



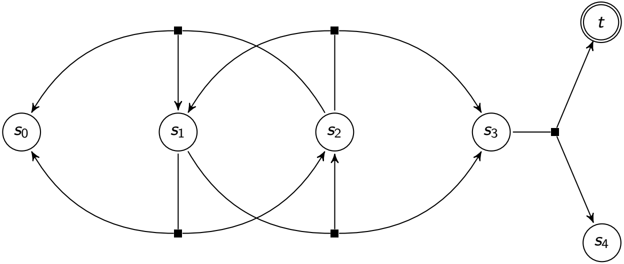
- Extremal-probability states: $\mathbf{Val}_\delta(s_0) = \mathbf{Val}_\delta(s_4) = 0$

Other components



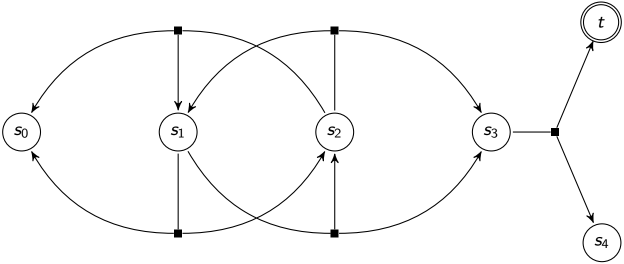
- Extremal-probability states: $\mathbf{Val}_\delta(s_0) = \mathbf{Val}_\delta(s_4) = 0$
- End components, essential states: $\emptyset$

Other components



- Extremal-probability states: $\mathbf{Val}_\delta(s_0) = \mathbf{Val}_\delta(s_4) = 0$
- End components, essential states: $\emptyset$
- Other: $\mathbf{Val}_\delta(s_1) = \mathbf{Val}_\delta(s_2)$

Other components



- Extremal-probability states: $\mathbf{Val}_\delta(s_0) = \mathbf{Val}_\delta(s_4) = 0$
- End components, essential states: $\emptyset$
- Other: $\mathbf{Val}_\delta(s_1) = \mathbf{Val}_\delta(s_2)$

(the above analysis holds for all same-support δ!)

The never-worse relation

Consider an MDP $\mathcal{M} = (S, A, \delta, T)$.

Never worse

For states $Q \subseteq S$ and a state s, we say Q is never worse than s if

$$\mathbf{Val}_\mu(s) \leq \max_{q \in Q} \mathbf{Val}_\mu(q)$$

for all $\mu : S \times A \to \mathbb{D}(S)$ with the same support as δ.

The never-worse relation

Consider an MDP $\mathcal{M} = (S, A, \delta, T)$.

Never worse

For states $Q \subseteq S$ and a state s, we say Q is never worse than s if

$$\mathbf{Val}_{\mu}(s) \leq \max_{q \in Q} \mathbf{Val}_{\mu}(q)$$

for all $\mu : S \times A \to \mathbb{D}(S)$ with the same support as δ.

Theorem (Collapsing NWR-equivalent states)

If s is never worse than q and vice versa, then they can be "collapsed".

The never-worse relation

Consider an MDP $\mathcal{M} = (S, A, \delta, T)$.

Never worse

For states $Q \subseteq S$ and a state s, we say Q is never worse than s if

$$\mathbf{Val}_\mu(s) \leq \max_{q \in Q} \mathbf{Val}_\mu(q)$$

for all $\mu : S \times A \to \mathbb{D}(S)$ with the same support as δ.

Theorem (Collapsing NWR-equivalent states)

If s is never worse than q and vice versa, then they can be "collapsed".

Theorem (Removing sub-optimal actions)

If $A \setminus \{a\}$ is never worse than a from s, then playing a from s can be ruled out.

First check: captures known reductions

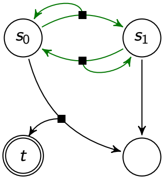
Proposition (Known reductions are special cases)

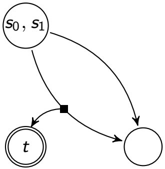
Known polynomial-time computable reduction heuristics (end components, extremal-probability states, essential states, ...) are all special cases of the NWR.

First check: captures known reductions

Proposition (Known reductions are special cases)

Known polynomial-time computable reduction heuristics (end components, extremal-probability states, essential states, ...) are all special cases of the NWR.





s_0 and s_1 are NWR-equivalent

Second check: captures other gadgets

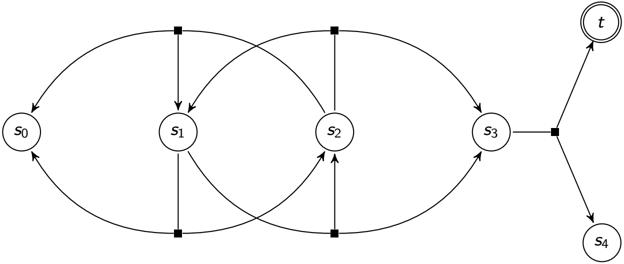
Proposition

Other reduction heuristics (patterns), again special cases of the NWR, are computable in polynomial time.

Second check: captures other gadgets

Proposition

Other reduction heuristics (patterns), again special cases of the NWR, are computable in polynomial time.



s_1 and s_2 are NWR-equivalent

Third check: works in practice?

PRISM: Randomized consensus shared coin protocol

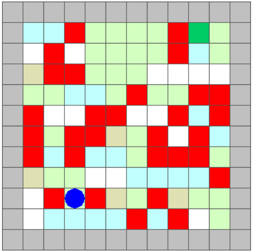
Formula	No reds.	Known reds.	New reds.
φ_1	400	392	76
φ_2	400	392	92

$$\varphi_1 = \Diamond\,(\text{“finished”} \wedge \text{“all coins equal 1”})$$

$$\varphi_2 = \Diamond\,(\text{“finished”} \wedge \neg\text{“all coins equal 1”})$$

Third check: works in practice?

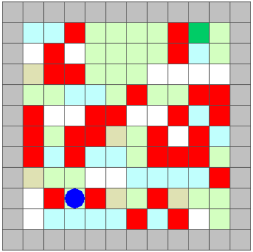
PAC learning a gridworld



The objective is to maximize the probability of reaching the green state while avoiding the red ones. The success probability of moves is unknown.

Third check: works in practice?

PAC learning a gridworld



The objective is to maximize the probability of reaching the green state while avoiding the red ones. The success probability of moves is unknown.

	No reds.	Known reds.	New reds.
Distributions	400	102	8
Episodes	1,133,243	948,882	83,564
Total steps	11,683,438	7,848,560	734,465

Graph-based characterization

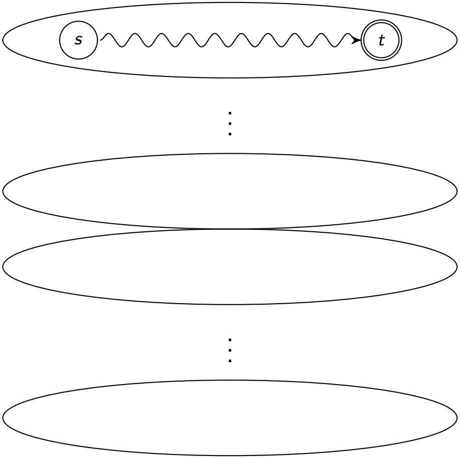
Theorem

Q is "sometimes worse" than s iff there exists a (Q, s)-drift partition.

Graph-based characterization

Theorem

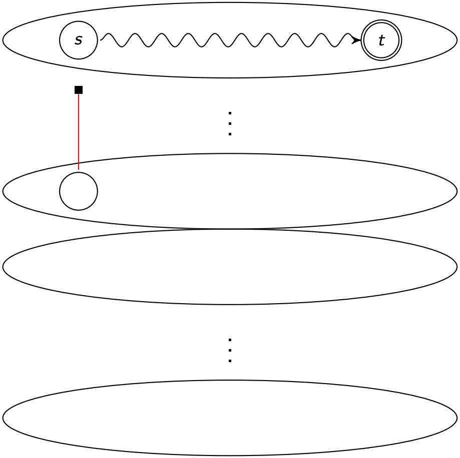
Q is "sometimes worse" than s iff there exists a (Q, s)-drift partition.



Graph-based characterization

Theorem

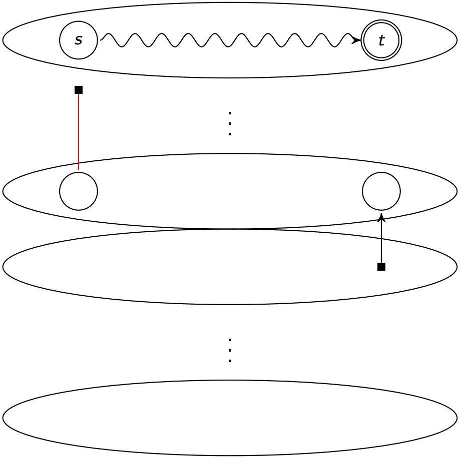
Q is "sometimes worse" than s iff there exists a (Q, s)-drift partition.



Graph-based characterization

Theorem

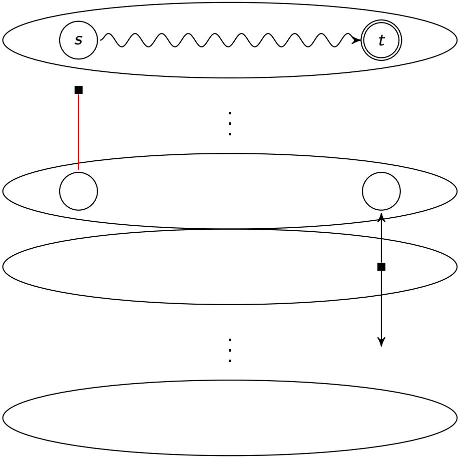
Q is "sometimes worse" than s iff there exists a (Q, s)-drift partition.



Graph-based characterization

Theorem

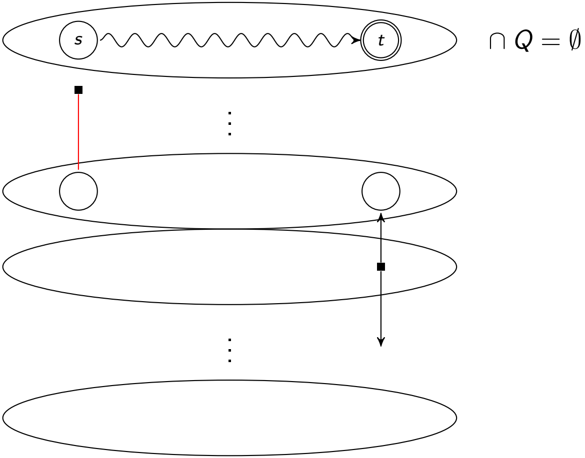
Q is "sometimes worse" than s iff there exists a (Q, s)-drift partition.



Graph-based characterization

Theorem

Q is "sometimes worse" than s iff there exists a (Q, s)-drift partition.



If: from the partition to sometimes worse

Fix some $0 < \varepsilon < 1$ and let δ be such that

If: from the partition to sometimes worse

Fix some $0 < \varepsilon < 1$ and let δ be such that

1. all transitions in the *s*–*t* path have probability $1 - \varepsilon$ (at least),
2. all transitions to states below have probability $1 - \varepsilon$ (at least),
3. all transitions to states above have probability ε at most.

If: from the partition to sometimes worse

Fix some $0 < \varepsilon < 1$ and let δ be such that

1. all transitions in the s–t path have probability $1 - \varepsilon$ (at least),
2. all transitions to states below have probability $1 - \varepsilon$ (at least),
3. all transitions to states above have probability ε at most.

One can then prove that

- $\mathbf{Val}_\delta(s) \geq (1 - \varepsilon)^{|S|}$ and
- $\mathbf{Val}_\delta(q) \leq 1 - (1 - \varepsilon)^{|S|}$ for all $q \in Q$.

If: from the partition to sometimes worse

Fix some $0 < \varepsilon < 1$ and let δ be such that

1. all transitions in the *s*–*t* path have probability $1 - \varepsilon$ (at least),
2. all transitions to states below have probability $1 - \varepsilon$ (at least),
3. all transitions to states above have probability ε at most.

One can then prove that

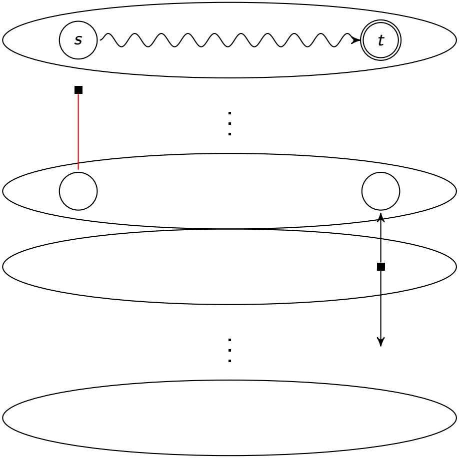
- $\mathbf{Val}_\delta(s) \geq (1 - \varepsilon)^{|S|}$ and
- $\mathbf{Val}_\delta(q) \leq 1 - (1 - \varepsilon)^{|S|}$ for all $q \in Q$.

For sufficiently small ε, we get

$$\mathbf{Val}_\delta(s) > \max_{q \in Q} \mathbf{Val}_\delta(q)$$

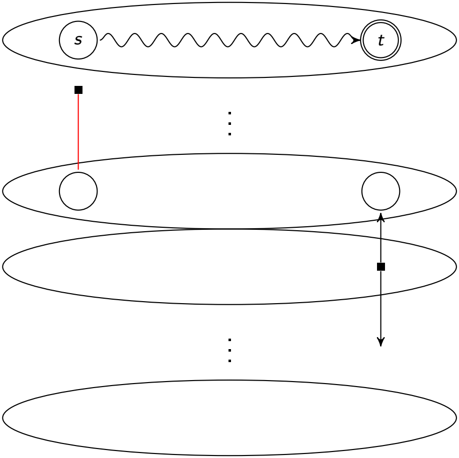
Only-if: from sometimes worse to a partition

Assuming that Q is sometimes worse than s, let $x_0 < x_1 < \cdots$ be the values of all the states (and distributions), with $x_i = \mathbf{Val}(s)$...



Only-if: from sometimes worse to a partition

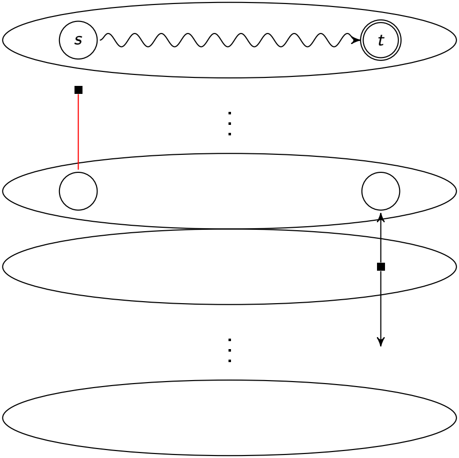
Assuming that Q is sometimes worse than s, let $x_0 < x_1 < \cdots$ be the values of all the states (and distributions), with $x_i = \mathbf{Val}(s)$...



$\{r : \mathbf{Val}(r) \geq \mathbf{Val}(s)\}$

Only-if: from sometimes worse to a partition

Assuming that Q is sometimes worse than s, let $x_0 < x_1 < \cdots$ be the values of all the states (and distributions), with $x_i = \mathbf{Val}(s)$...



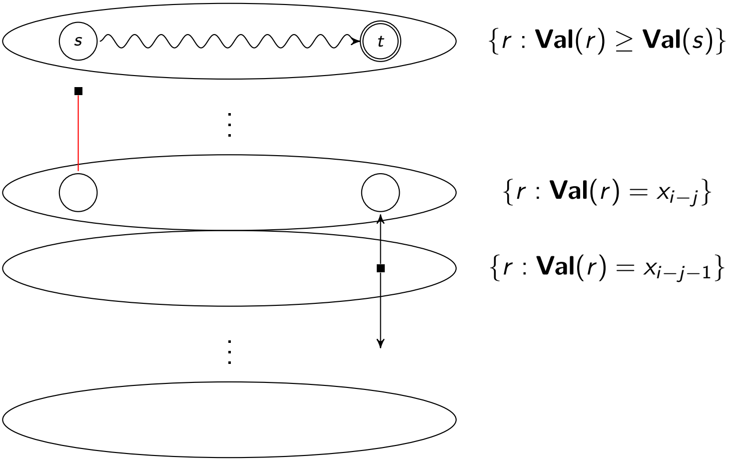
$\{r : \mathbf{Val}(r) \geq \mathbf{Val}(s)\}$

$\{r : \mathbf{Val}(r) = x_{i-j}\}$

$\{r : \mathbf{Val}(r) = x_{i-j-1}\}$

Only-if: from sometimes worse to a partition

Assuming that Q is sometimes worse than s, let $x_0 < x_1 < \cdots$ be the values of all the states (and distributions), with $x_i = \mathbf{Val}(s)$...



One can then show this is indeed a (Q, s)-drift partition.

The complexity of the NWR

Theorem (NWR-membership)

Given an MDP $\mathcal{M}$, Q and s, determining if Q is never worse than s is coNP*-complete.*

The complexity of the NWR

Theorem (NWR-membership)

Given an MDP $\mathcal{M}$, Q and s, determining if Q is never worse than s is coNP*-complete.*

- Membership follows from the graph-based characterization
- Hardness is proved via a reduction from 2-disjoint-paths to the existence of a drift partition

The complexity of the NWR

Theorem (NWR-membership)

Given an MDP $\mathcal{M}$, Q and s, determining if Q is never worse than s is coNP-*complete.*

- Membership follows from the graph-based characterization
- Hardness is proved via a reduction from 2-disjoint-paths to the existence of a drift partition

Wait what!?

- Did you just try to sell me a coNP pre-processing procedure for a polynomial-time problem? [Fijalkow 18]

The complexity of the NWR

Theorem (NWR-membership)

Given an MDP $\mathcal{M}$, Q and s, determining if Q is never worse than s is coNP-*complete.*

- Membership follows from the graph-based characterization
- Hardness is proved via a reduction from 2-disjoint-paths to the existence of a drift partition

Wait what!?

- Did you just try to sell me a coNP pre-processing procedure for a polynomial-time problem? [Fijalkow 18]
- Yes! but value iteration is exponential in the worst case.
- Also, learning the probabilities takes exponentially many experiments.

The complexity of the NWR

Theorem (NWR-membership)

Given an MDP $\mathcal{M}$, Q and s, determining if Q is never worse than s is coNP-*complete.*

- Membership follows from the graph-based characterization
- Hardness is proved via a reduction from 2-disjoint-paths to the existence of a drift partition

Wait what!?

- Did you just try to sell me a coNP pre-processing procedure for a polynomial-time problem? [Fijalkow 18]
- Yes! but value iteration is exponential in the worst case.
- Also, learning the probabilities takes exponentially many experiments.
- The relation can be queried using a SAT solver.
- Non-tractability further motivates under-approximating the relation.

Efficient under-approximations of the NWR

Iterative algorithm

Let $\hat{R}$ be the relation containing all NWR-pairs one gets from

- extremal-probability states,
- essential states,
- maximal end components.

Repeat until convergence: "grow" $\hat{R}$ using efficiently-computable rules that imply more NW-pairs.

Efficient under-approximations of the NWR

Iterative algorithm

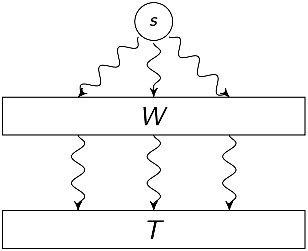
Let $\hat{R}$ be the relation containing all NWR-pairs one gets from

- extremal-probability states,
- essential states,
- maximal end components.

Repeat until convergence: "grow" $\hat{R}$ using efficiently-computable rules that imply more NW-pairs.

$$\hat{R} \subseteq \mathrm{NWR}$$

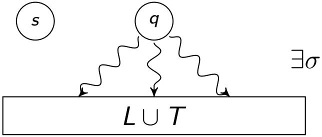
Rule 1



Proposition (Rule 1)

Given s and $Q \subseteq S$, if we find the above pattern with $W = \{r : Q \text{ is NW than } r\}$ then Q is never worse than s.

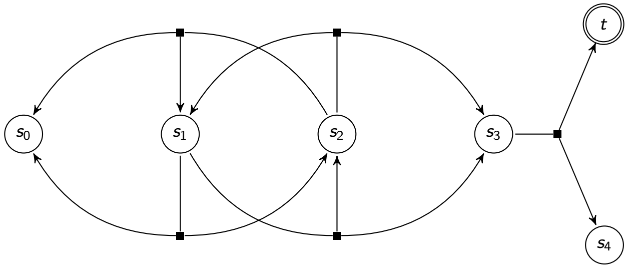
Rule 2



Proposition (Rule 2)

Given s and q, if we find the above pattern with $L = \{r : r \text{ is NW than } s\}$ then q is never worse than s.

Back to those other components



- Rule 1: s_3 is never worse than $s_1, s2$
- Rule 2: s_1, s_2 are never worse than $s3$

Fin

Conclusions

- Nice relation giving a sufficient condition for MDP reductions
- Seems to work in practice (in terms of reduction efficiency) [Bharadwaj, Le Roux, P., Topcu IJCAI'17]
- Exact complexity of the full relation [Le Roux, P. FoSSaCS'18]

Fin

Conclusions

- Nice relation giving a sufficient condition for MDP reductions
- Seems to work in practice (in terms of reduction efficiency) [Bharadwaj, Le Roux, P., Topcu IJCAI'17]
- Exact complexity of the full relation [Le Roux, P. FoSSaCS'18]

Future work

- Relation to "value-preserving sets"?
- More experiments (SAT-solvers for full relation; impact on MC running time)
- Extensions
  - on-the-fly algorithms
  - finite-horizon reachability
  - reward MDPs (expected mean payoff, etc.)