

Synthesis in MDP with strong guarantees

Jean-François Raskin (ULB)

IFIP Working group 2.2
Bordeaux
September 2017

Mix MDP and Games

- **Markov decision processes = env. behaves stochastically**

+ optimal strategies

- outliers

- **Zero-sum games = env. is the adversary**

+ robustness of winning strategies

- may be over conservative

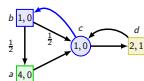
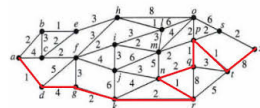
- **Can we have the best of both ?**

- **“Variations on the stochastic shortest path problem”**

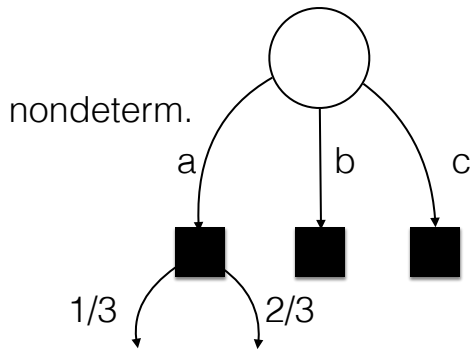
(see VMCA15 paper for details and other pointers)

- **“Threshold Constraints with Guarantees for Parity Objectives in Markov Decision Processes”**

(see ICALP17 paper for details and other pointers)



MDP=nonterm.+stoch.



Stochastic choice

MDP $D=(S, s_{init}, A, \delta, w)$:

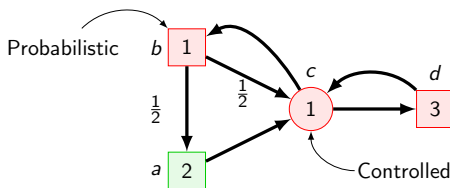
- S : finite set of states
- s_{init} : initial state
- A : finite set of actions
- $\delta : S \times A \rightarrow \text{Dist}(S)$

if $|A|=1$,

Markov Chain

(purely stochastic)

MDP=nondeterm.+stoch.



ω -regular - parity condition

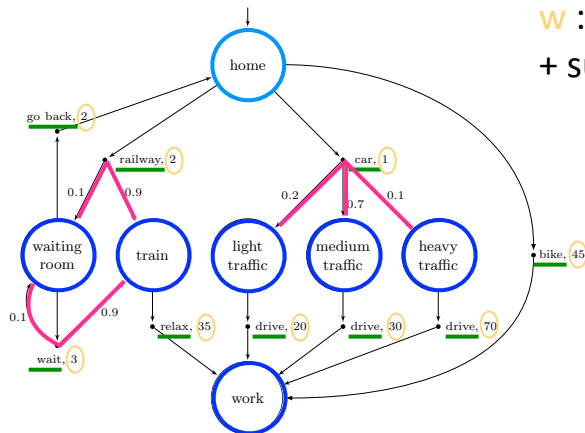
$p : S \rightarrow \{0, 1, \dots, d\}$

MDP=nondeterm.+stoch.

Quantitative

$w : S \times A \rightarrow \mathbb{Z}$

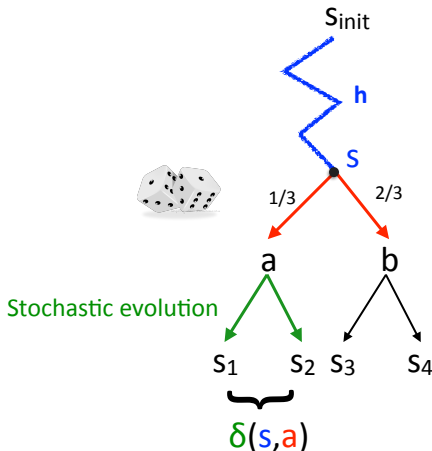
+ sum, mean-payoff, etc.



Strategies

MDP $D=(S,s_{init},A,\delta,w)$

strategy $\sigma : (S \times A)^* \cdot S \rightarrow \text{Dist}(A)$



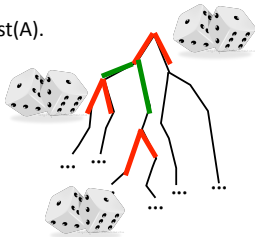
for each history h
=seq. of pairs of states and actions,
 σ prescribes a **possibly randomized**
choice of action to play

... can be represented as
a **Stochastic Moore Machine**

Strategies

(General) Strategy:

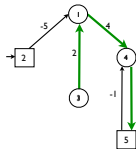
$\sigma: (S.A)^*. S \rightarrow \text{Dist}(A)$.



Memoryless strategy:

$\sigma_m: S \rightarrow \text{Dist}(A)$.

$\Sigma_{1,m}$ = set of memoryless strategies

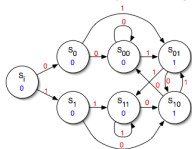


Finite-memory strategy:

$\sigma_f: (S.A)^*. S \rightarrow \text{Dist}(A)$

but **regular** (finite Moore machine)

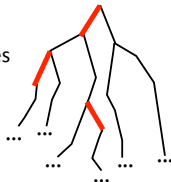
Σ_f = set of finite memory strategies



Pure strategy:

$\sigma_p: (S.A)^*. S \rightarrow A$.

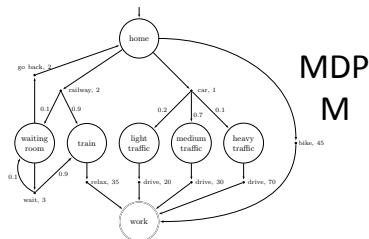
Σ_p = set of pure strategies



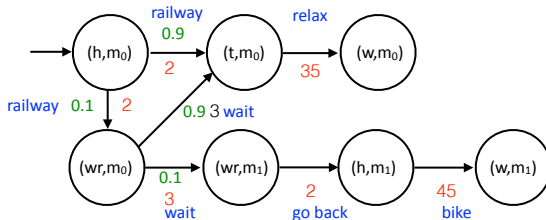
+memoryless and pure

MDP+Strategy=Markov Chain

Strategy : σ



Resulting Markov Chain : $M \otimes \sigma$



Prob. of event - Expected value

Event=measurable set of runs E of a MC,

Every event has a **uniquely defined probability** in a MC M .

- $\mathbb{P}(M)(E)$ =**probability that a run belongs to E**
when starting from dinit, and M is executed for ∞ -many steps.
- $\mathbb{E}(M)(f)$ =**the expected value or expectation of f** over initial runs in M . Where f is a measurable function $f : \text{Runs}(M) \rightarrow \mathbb{R} \cup \{\infty\}$.

Outcome of a strategy

Game view

Given a MDP M and a strategy σ , the outcomes of σ in M , noted **Out**(M, σ), are all the (initial) infinite paths that can occur under the strategy σ , that is all the

$$s_0 s_1 \dots s_n \dots$$

such that for all $i \geq 0$:

$$\delta(s_i, \sigma(s_0 s_1 \dots s_i))(s_{i+1}) > 0.$$

Plan

1. Variation on the Stochastic Shortest Path Problem

Given a \mathbb{N}_0 -weighted MDP M , a target set T , and two thresholds c_1 and c_2 , decide if there exists a strategy σ such that:

- A. All outcomes compatible with σ reach T within time c_1
- B. Under σ , the expected time to reach T is less than c_2

2. Variation on the ω -regular verification problem

Given a MDP M and two ω -regular objectives defined by parity functions p_1 and p_2 , and a threshold $c \in [0,1]$, decide if there exists a strategy σ such that:

- C. All outcomes compatible with σ satisfy p_1
- D. Under σ , the probability that p_2 is satisfied is larger than c

Plan

1. Variation on the Stochastic Shortest Path Problem

Given a \mathbb{N}_0 -weighted MDP M , a target set T , and two thresholds c_1 and c_2 , decide if there exists a strategy σ such that:

- A. All outcomes compatible with σ reach T within time c_1
- B. Under σ , the expected time to reach T is less than c_2

2. Variation on the ω -regular verification problem

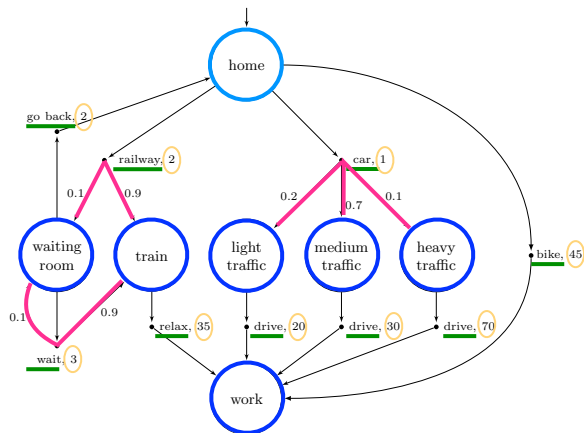
Given a MDP M and two ω -regular objectives defined by parity functions p_1 and p_2 , and a threshold $c \in [0,1]$, decide if there exists a strategy σ such that:

- C. All outcomes compatible with σ satisfy p_1
- D. Under σ , the probability that p_2 is satisfied is larger than c

The Stochastic Shortest Path Problem

Weighted MDP with weight in \mathbb{N}_0

Running example

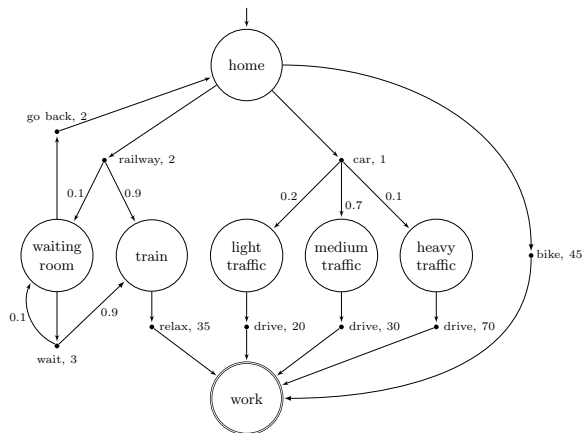


Expectation

Min. expected length to target

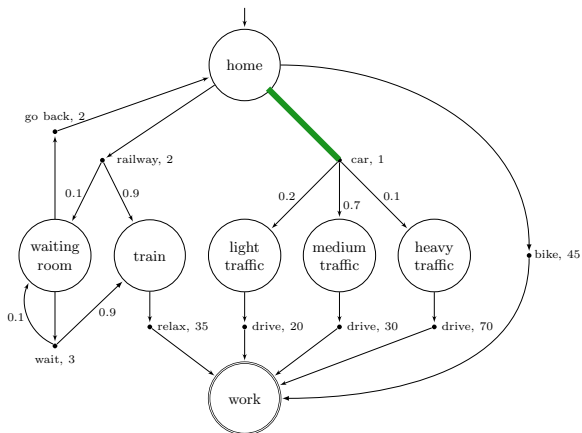
Problem (SSP-E): Given a single dimensional weighted MDP $D=(S, s_{init}, A, \delta, w)$, a set of target states $T \subseteq S$, and a value $v \in \mathbb{N}$, decide if there exists a strategy σ such that $\mathbb{E}(D, \sigma)(T \mid S) \leq v$?

Back to the example



How to minimize the **expected length** ?

Back to the example



Take the **car**:

$$1 + 0.2 \times 20 + 0.7 \times 30 + 0.1 \times 70 = \mathbf{33} \text{ minutes}$$

Min. expected length to target

Algorithms:

- value iteration
- reduction to LP

LP for min. expected length

- Remove states that cannot reach T (their expectation is $+\infty$)

- For all $s \in T$, the expectation is 0



- for other states, let x_s =expectation from state s be a solution of the following LP:

$$\text{Max } \sum_{s \in S \setminus T} x_s$$

under the constraints

$$x_s \leq w(a) + \delta(s, a, s') \cdot x_{s'} \text{ for all } s, s' \in S \setminus T, \text{ for all } a \in A(s)$$

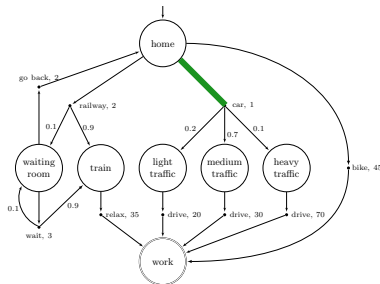
Min. expected length to target

Theorem: The SSP-E problem can be decided in **polynomial** time. Optimal **pure memoryless** strategies always exist and can be constructed in polynomial time.



"Angry? No I'm not angry
you took a job elsewhere."

Outliers



With car, the prob. of **long** runs (e.g. 71 minutes) is not negligible (10%).

What if the employee is **risk-averse** ?

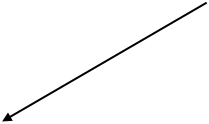
Forcing short paths with
high probability

Forcing short paths with high probability

Problem (SSP-P): Given a single dimensional weighted MDP $D=(S, s_{init}, A, \delta, w)$, a set of target states $T \subseteq S$, and a value $v \in \mathbb{N}$, and probability threshold $\alpha \in (0, 1]$, decide if there exists a strategy σ such that $\mathbb{P}(D, \sigma)(TS^T \leq v) \geq \alpha$?

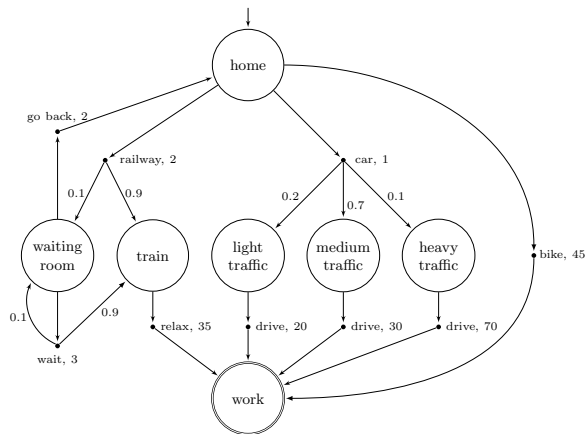
Forcing short paths with high probability

Problem (SSP-P): Given a single dimensional weighted MDP $D=(S,s_{init},A,\delta,w)$, a set of target states $T\subseteq S$, and a value $v\in\mathbb{N}$, and probability threshold $\alpha\in(0,1]$, decide if there exists a strategy σ such that $\mathbb{P}(D, \sigma)(TS^T\leq v)\geq\alpha$?



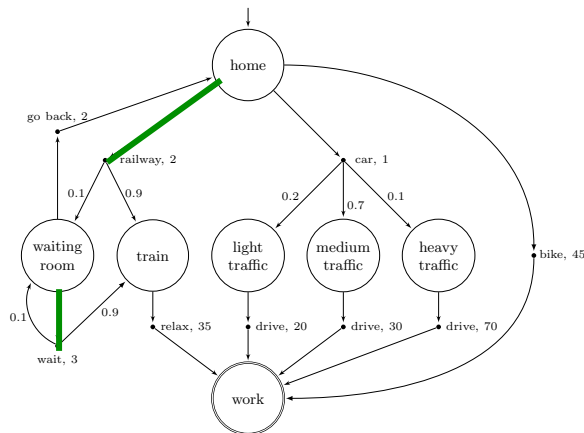
Percentile constraint

Back to the example



Is it possible to reach work
within 40' with prob. $\geq 95\%$?

Back to the example



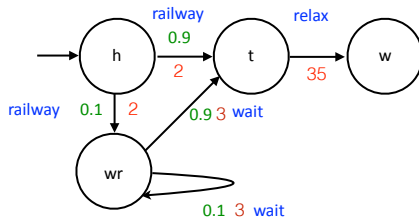
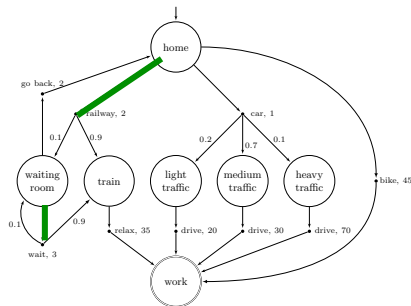
Is it possible to reach work within 40' with prob. $\geq 95\%$?

Yes !

Solution: Take the **train** !

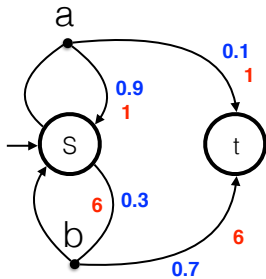
Back to the example

Take the **train** !



$$\mathbb{P}(D, \sigma)(TST \leq 40) \geq 0.9 + 0.1 \times 0.9 = 0.99$$

Memory is necessary



Reach u with $TS^{(t)} \leq 10$ with **prob.** $\geq 75\%$?

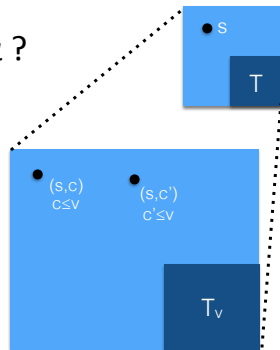
Always “a”: after 10 times, the prob. to be in s is $\geq 0.34 \succ$ **KO**

Always “b”, can only be played one time: prob. to reach t is $0.7 \succ$ **KO**

Play 4 times “a” then 1 time “b”: t is reached with prob. $\geq 81\% \succ$ **OK**

Algorithm

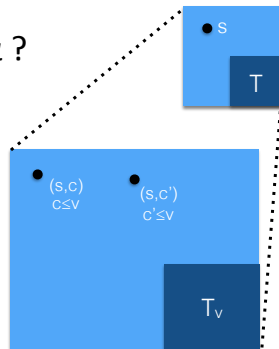
- Transform the MDP: **explicit accumulated cost** up to upper bound v
- Solve a **reachability query**: can we reach $T_v = \{ (s,c) \mid s \in T \wedge c \leq v \}$ with probability $\geq \alpha$?
- Such query can be solved a.o. using LP



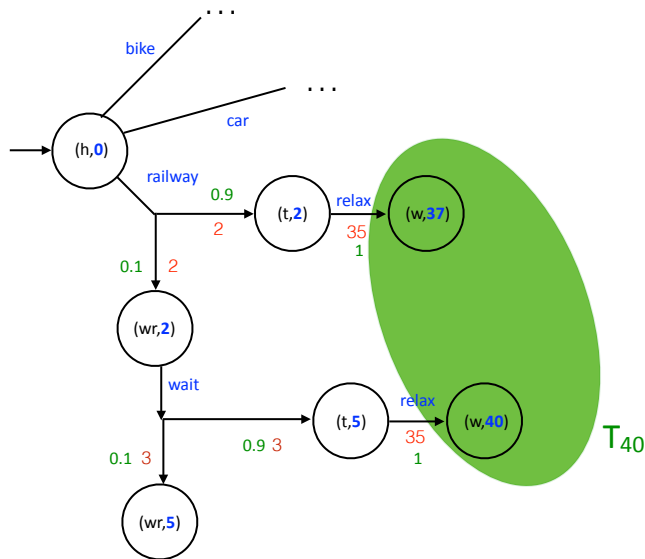
Algorithm

- Transform the MDP: **explicit accumulated cost** up to upper bound v
- Solve a **reachability query**: can we reach $T_v = \{ (s, c) \mid s \in T \wedge c \leq v \}$ with probability $\geq \alpha$?
- Such query can be solved a.o. using LP

Memory=accumulated cost



Back to the example



$$T_{40} = \{ (w,c) \mid c \leq 40 \}$$

if $\sigma = \text{take train}$ then
 $\mathbb{P}(D, \sigma)(\Diamond T_{40}) \geq 0.99$

Classical
**reachability
query**
(LP or Value Iter.)

Forcing short paths with high probability

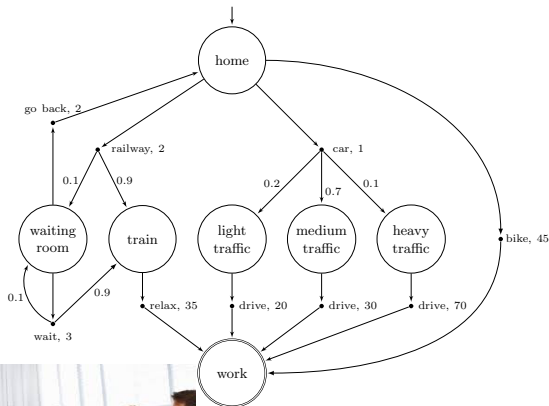
Theorem: The SSP-P problem can be decided in **pseudo-polynomial** time and it is **PSpace-Hard***. Optimal pure strategies with pseudo-polynomial **memory** always exist and can be constructed in **pseudo-polynomial time**.

To the best of my knowledge, the exact complexity of the decision problem is open.

*C. Haase and S. Kiefer. **The odds of staying on budget**. CoRR, abs/1409.8228, 2014.

Good expectation under
acceptable worst-case

What if you **ought** to be at work within one hour ?



- **Train** option leaves a small probability of **not** reaching work within 1 hour (i.e. 1%)
- What if this is **unacceptable** ?
- Take your **bike** !
- **But can we do better ?**
i.e. be sure to be at work within one hour with a better expectation than 45 min. ?



Worst-case guarantees with good expectation

Problem (SSP-WE): Given a single-dimensional weighted MDP $D=(S, s_{\text{init}}, A, \delta, w)$, a set of target states $T \subseteq S$, and two values v_1 and $v_2 \in \mathbb{N}$, decide if there exists a **unique strategy** σ such that:

1. [**Worst-case**] **for all** $\rho \in \text{Out}(D, \sigma): TS^T(\rho) \leq v_1$
2. [**Expectation**] $\mathbb{E}(D, \sigma)(TS^T) \leq v_2$

It is a **natural problem**: **avoid unacceptable outliers at all cost!**

Two views

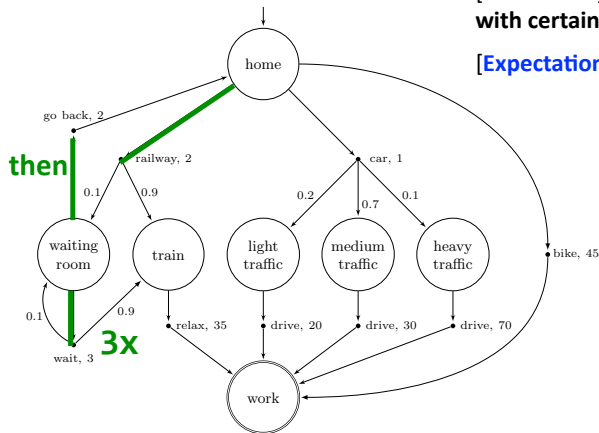
- A **game** + an **expected behavior** of the env./adversary given as a stochastic Moore machine: you want a winning strategy (worst-case) that **behaves well/better** against the **expected behavior** of the env./adversary
- A **MDP** (expectation) + you want to **avoid outliers** at all cost (worst-case)

Back to the example

- Wait for the train
- After three delays, goes back home and bike

[Worst-case] **safe**: at work within 58 minutes
with certainty

[Expectation] $\approx 37, 34 \dots$ minutes (< 45 —Bike)



Algorithm

- **Explicit accumulated cost** up to $v+1$
- Solve a **reachability game**: removes actions and states that cannot avoid cost $v+1$ before reaching T
 - then remove actions that leads to states where no action can be played (**unsafe states**)... up to a **fixed point**.
 - We obtain a MDP in which **all strategies are safe** (reach T within v)
- **Optimize** on the resulting MDP the **expected length** to T

Worst-case guarantees with good expectation

Theorem (SSP-WE): The SSP-WE problem can be decided in pseudo-polynomial time and is **PP-Hard*** (and so NP-Hard). Pseudo-polynomial **memory** is always sufficient and in general necessary, and satisfying strategies can be constructed in pseudo-polynomial time.

*by a reduction to K^{th} largest subset problem which was shown PP-complete recently. C. Haase and S. Kiefer.

Plan

1. Variation on the Stochastic Shortest Path Problem

Given a \mathbb{N}_0 -weighted MDP M , a target set T , and two thresholds c_1 and c_2 , decide if there exists a strategy σ such that:

- A. All outcomes compatible with σ reach T within time c_1
- B. Under σ , the expected time to reach T is less than c_2

2. Variation on the ω -regular verification problem

Given a MDP M and two ω -regular objectives defined by parity functions p_1 and p_2 , and a threshold $c \in [0,1]$, decide if there exists a strategy σ such that:

- C. All outcomes compatible with σ satisfy p_1
- D. Under σ , the probability that p_2 is satisfied is larger than c

MDPs with
two parity objectives

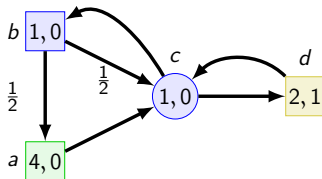
threshold with guarantees

What if we have two parity objectives, p_1 and p_2 , an initial state s , and want a strategy λ ensuring:

$$1 : \text{Out}_{\lambda,s} \subseteq \llbracket p_1 \rrbracket$$

$$2 : \mathbb{P}_{\lambda,s}(p_2) \geq 1$$

Here $c \models_{\mathcal{M}} S(p_1) \wedge AS(p_2)$



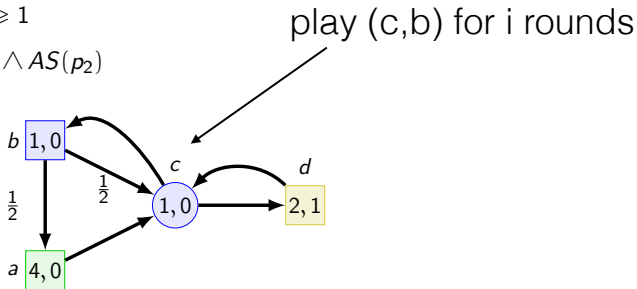
threshold with guarantees

What if we have two parity objectives, p_1 and p_2 , an initial state s , and want a strategy λ ensuring:

$$1 : \text{Out}_{\lambda,s} \subseteq \llbracket p_1 \rrbracket$$

$$2 : \mathbb{P}_{\lambda,s}(p_2) \geq 1$$

Here $c \models_{\mathcal{M}} S(p_1) \wedge AS(p_2)$



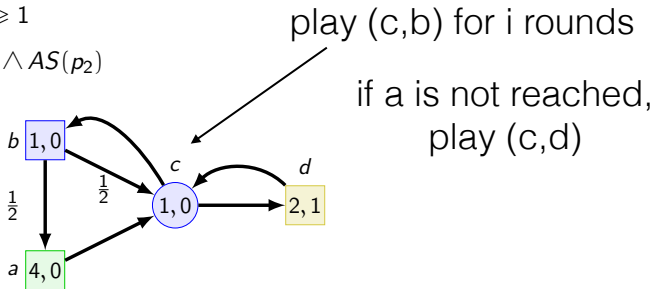
threshold with guarantees

What if we have two parity objectives, p_1 and p_2 , an initial state s , and want a strategy λ ensuring:

$$1 : \text{Out}_{\lambda,s} \subseteq \llbracket p_1 \rrbracket$$

$$2 : \mathbb{P}_{\lambda,s}(p_2) \geq 1$$

Here $c \models_{\mathcal{M}} S(p_1) \wedge AS(p_2)$



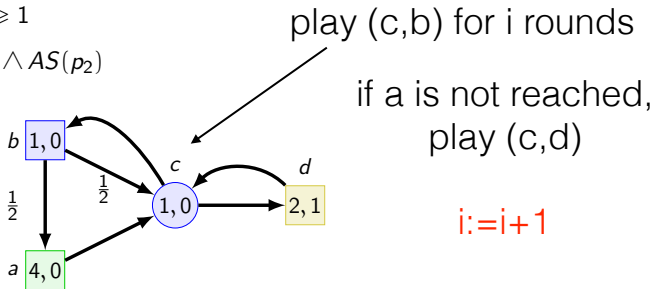
threshold with guarantees

What if we have two parity objectives, p_1 and p_2 , an initial state s , and want a strategy λ ensuring:

$$1 : \text{Out}_{\lambda,s} \subseteq \llbracket p_1 \rrbracket$$

$$2 : \mathbb{P}_{\lambda,s}(p_2) \geq 1$$

Here $c \models_{\mathcal{M}} S(p_1) \wedge AS(p_2)$



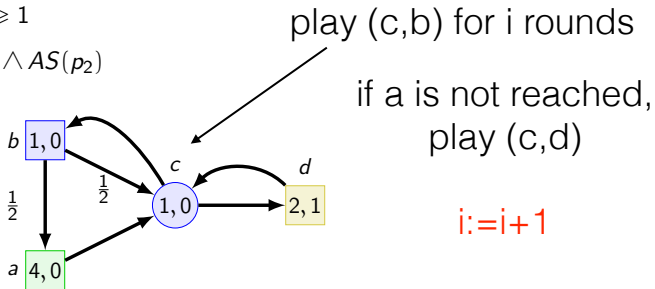
threshold with guarantees

What if we have two parity objectives, p_1 and p_2 , an initial state s , and want a strategy λ ensuring:

$$1 : \text{Out}_{\lambda,s} \subseteq \llbracket p_1 \rrbracket$$

$$2 : \mathbb{P}_{\lambda,s}(p_2) \geq 1$$

Here $c \models_{\mathcal{M}} S(p_1) \wedge AS(p_2)$



(c,d) is taken ∞ many times with Proba 0.

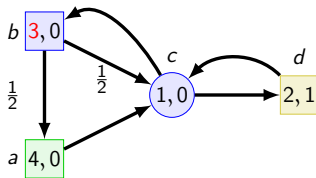
Winning each condition is not sufficient

Remark

Having a strategy λ_1 for $c \models_{\mathcal{M}} S(p_1)$ and a strategy λ_2 for $c \models_{\mathcal{M}} P_{\geq 1}(p_2)$ is not sufficient to have a strategy ensuring $c \models_{\mathcal{M}} S(p_1) \wedge P_{\geq 1}(p_2)$.

The following example has no winning strategy.

Problem: 2 does not cancel 3.



Ingredients of a solution:

Safe reachability

+

(Ultra Good) **End-Components**

Safe reachability

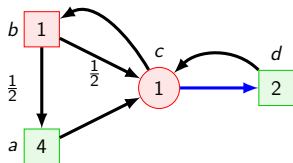
Definition

A set of states T can be reached safely from a state s with respect to a parity condition p if $s \models S(p) \wedge AS(\Diamond T)$.

This problem can be decided in $NP \cap \text{co-NP}$ [1].

On this example, a can be reached safely from c with respect to p .

We alternate between the two possible actions:

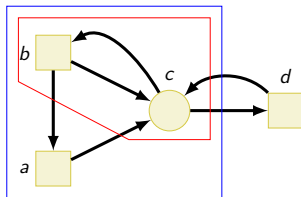


End-component

Definition

A subgraph C is an end-component if:

- C is strongly connected
- $\text{Post}_{\square}(C) \subseteq C$



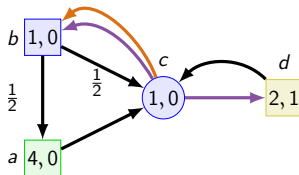
Theorem

For all strategy λ , $\mathbb{P}(\text{inf}(\lambda) = \text{EC}) = 1$ [3].

Condition for sure and almost-sure

An end-component C is *ultra-good* (UGEC) if we have:

- from all state, a strategy λ_1 reaching safely the maximum of p_1 with respect to p_1
- from all state, a strategy λ_2 having probability 1 of satisfying both p_1 and p_2



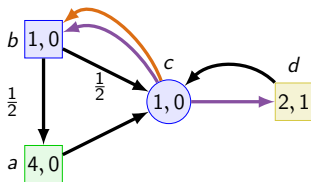
Lemma

The following holds: $\forall s \in \text{UGEC} : s \models_{\mathcal{M}} S(p_1) \wedge AS(p_2)$

Strategy for sure and almost-sure

Strategy at round i :

- Play i times λ_2 .
- If the current maximum of the round is odd, play λ_1 until reaching the maximum of p_1 .
- begin the next round $i + 1$



Lemma

The following holds: $\forall s \in UGEC : s \models_{\mathcal{M}} S(p_1) \wedge AS(p_2)$

General result

Theorem

Given an MDP \mathcal{M} , a state $s_0 \in S$, and two priority functions p_1, p_2 , it can be decided in $\text{NP} \cap \text{coNP}$ if $s_0 \models S(p_1) \wedge P_{\sim k}(p_2)$ for $\sim \in \{>, \geq\}$ and $k \in \mathbb{Q} \cap [0, 1]$.

If the answer is Yes, then there exists an infinite-memory witness strategy, and infinite memory is in general necessary. This decision problem is at least as hard as solving parity games.

→ Proof of this result relies on the notion of UGEC and safe reachability.

More results

- Variations on the **stochastic shortest path** problem (VMCAI'15).
 - MDP with **two parity conditions**: p_1 surely and p_2 above some threshold (ICALP'17).
 - MDP with **mean-payoff objective(s)**: ensure minimal performance and good expectation (STACS'14 and LICS'15).
 - MDP with **multi-objective percentile constraints** (CAV'15).
 - MDP with **several environments** (FSTTCS'15).
 - **POMDP** with **discounted sum objectives**: ensure minimal performance and good expectation (IAAA'17)
- ... and by others:
- MDP with **mean-payoff** (expectation) and **parity** (surely) - by Kupferman et al. (CONCUR'16).
 - MDP with **mean-payoff** (expectation) and **energy** (surely) - by Brádzil et al. (ATVA'16).
 - ...

Conclusion

- **Zero-sum games = env. is the adversary**
 - + robustness of winning strategies
 - solutions may be over conservative
- **Markov decision processes = env. behaves stochastically**
 - + allow for optimal strategies
 - not robust against outliers
- We have **algorithms** to analyse a **mix of MDP and games**



Thanks !
Questions ?