M1 Internship: Backward Responsibility in Counterexamples of Model Checkers

Roxane Van den Bossche

Tutors: *Christel Baier, Sascha Klüppelholz, Jakob Piribauer*
Coworker: *Johannes Lehmann*

Institut für Informatik, Technische Universität Dresden

école normale supérieure paris—saclay

CHAIR OF ALGEBRAIC AND LOGICAL FOUNDATIONS OF COMPUTER SCIENCE

TECHNISCHE UNIVERSITÄT DRESDEN

February - June 2023

Outline

1. The problem
2. Semivalues
3. Optimistic and pessimistic responsibilities
4. Characterisation of the optimistic responsibility
5. Complexity results

What is model checking? (at least in our case)

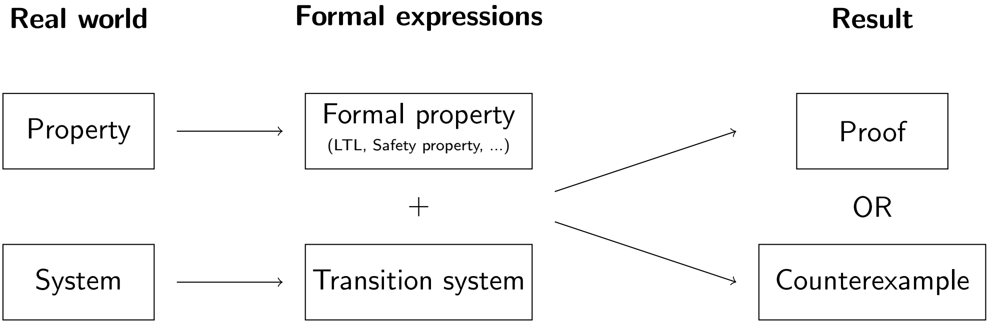


Figure: Model checking in a diagram

The problem: Intuition

"Backward Responsibility in Counterexamples of Model Checkers"

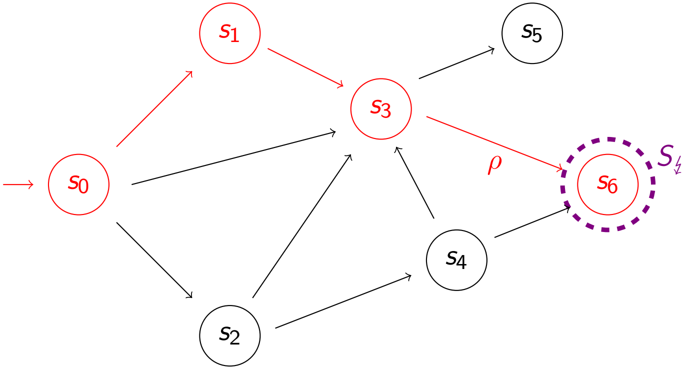


Figure: A transition system with a counterexample in red

The problem: Transition systems

Definitions:

- **Transition systems:** $TS = (S, \rightarrow, s_0)$, deterministic
- **Runs:** infinite sequence of states $\rho = \rho_0\rho_1 \ldots \in S^\omega$ where $\rho_0 = s_0$ and $\forall i \in \mathbb{N}\ \rho_i \rightarrow \rho_{i+1}$
- **Set of bad states:** $S_\lightning \subseteq S$
- **Counterexamples:** $\rho = \rho_0 \ldots \rho_k \in S^*$ such that it is the prefix of a run, $\rho_k \in S_\lightning$, $\rho_i \notin S_\lightning$ for $i \in \{0, \ldots, k-1\}$ and $\rho_i \neq \rho_j$ for all $i \neq j$ *ie.* they are loop-free.

Semivalues

- Finite set of *players* X
- *Coalitions*: $C \subseteq X$
- *Cooperative games*: $v\colon 2^X \to \mathbb{R}$
- Set of cooperative games on X: G^X

Definition (Semivalue)

Let X be a finite set of players with $n := |X|$. Then $\mathcal{R}\colon G^X \to X \to \mathbb{R}$ is a *semivalue* if there exists a weight vector $p = (p_0, \dots, p_{n-1})$ such that, for any game $v \in G^X$ and player $i \in X$, we have

$$\mathcal{R}(v, i) = \sum_{C \subset X \setminus \{i\}} p_{|C|}[v(C \cup \{i\}) - v(C)]$$

Semivalues

Definition

We call $p = (p_0, \ldots, p_{n-1})$ a *weight vector* if

$$\sum_{k=0}^{n-1} \binom{n-1}{k} p_k = 1.$$

Classical semivalues:

- The *Shapley value*: $p_k^{\mathcal{S}} := \frac{(n-k-1)!k!}{n!}$
- The *Banzhaf value*: $p_k^{\mathcal{B}} := \frac{1}{2^{n-1}}$

Intuition

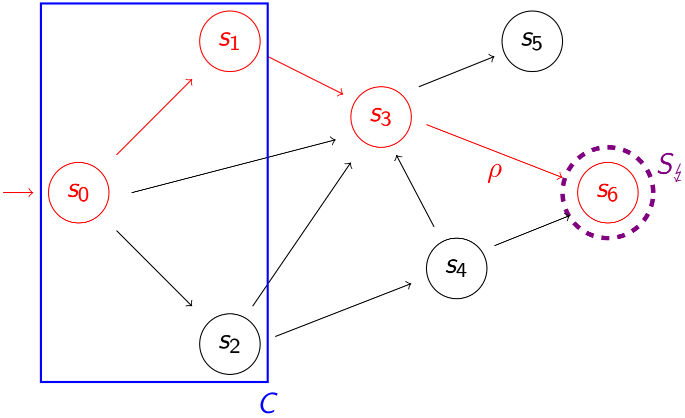


Figure: A transition system with a counterexample in red

Intuition

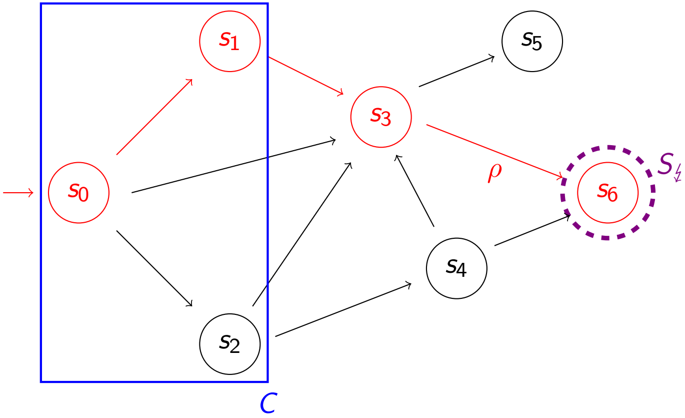


Figure: A transition system with a counterexample in red

➔ How to quantify the actions of $S \setminus (C \cup \rho)$?

Intuition

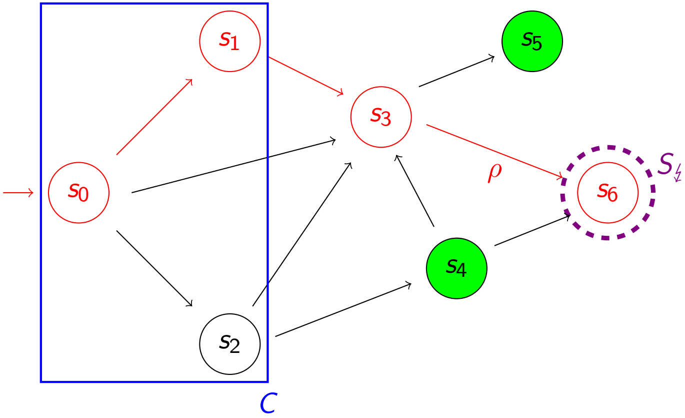


Figure: A transition system with a counterexample in red

➔ How to quantify the actions of $S \setminus (C \cup \rho)$?

Optimistic and pessimistic responsibilities

Safety games: $(S_{Safe}, S_{Reach}, \to, s_0, S_{\lightning})$

- Transition system: $(S, \to, s_0)$
- $S := S_{Safe} \uplus S_{Reach}$ and $S_{\lightning} \subseteq S$
- Winning condition of the form $\Omega_{S_{\lightning}} = \{\rho \mid \forall i \in \mathbb{N}\colon \rho_i \notin S_{\lightning}\}$
- A strategy for *Safe* is a function $\sigma\colon S_{Safe} \to S$ with $s \to \sigma(s)$ for all $s \in S_{Safe}$ (same for *Reach*)
- A strategy for *Safe* is winning if, for all strategies of *Reach*, the induced play is winning for *Safe*, *ie.* $\rho \in \Omega_{S_{\lightning}}$.

Optimistic and pessimistic responsibilities

$\mathcal{G}^{TS}_{\rho,S_{\lightning}}(C)$: safety game defined as $(C, S \setminus C, \rightarrow', s_0, S_{\lightning})$ where

- $\rightarrow'$ is $\rightarrow$ in which actions from ρ are "engraved" for $\rho \setminus C$.
- *Safe* controls C
- *Reach* controls $S \setminus C$

Definition (Optimistic and pessimistic cooperative games)

Let $C \subseteq S$.

Optimistic cooperative game: $$v_\top(C) = \begin{cases} 1 & \text{if player } \textit{Safe} \text{ wins } \mathcal{G}^{TS}_{\rho,S_{\lightning}}(C \cup (S \setminus \rho)) \\ 0 & \text{otherwise} \end{cases}$$

Pessimistic cooperative game: $$v_\perp(C) = \begin{cases} 1 & \text{if player } \textit{Safe} \text{ wins } \mathcal{G}^{TS}_{\rho,S_{\lightning}}(C) \\ 0 & \text{otherwise.} \end{cases}$$

Optimistic and pessimistic responsibilities

Now we can apply semivalues :-)

Remember how semivalues look like:

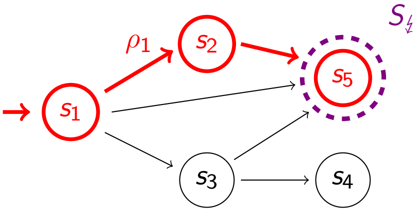
$$\mathcal{R}(v, i) = \sum_{C \subset X \setminus \{i\}} p_{|C|}[v(C \cup \{i\}) - v(C)]$$

Definition (Responsibility)

Let ρ be a counterexample, let $\mathcal{R}$ be a semivalue on G^S.

1. The *optimistic responsibility of s with respect to* $\mathcal{R}$ is $\mathcal{R}(v_\top, s)$.
2. The *pessimistic responsibility of s with respect to* $\mathcal{R}$ is $\mathcal{R}(v_\bot, s)$.

Example



s	$\mathcal{S}(v_\top, s)$	$\mathcal{S}(v_\bot, s)$	$\mathcal{B}(v_\top, s)$	$\mathcal{B}(v_\bot, s)$
s_1	1	0.5	1	0.5
s_2	0	0	0	0
s_3	0	0.5	0	0.5
s_4	0	0	0	0
s_5	0	0	0	0

Figure: Working example 3, run 1

Characterisation of the optimistic responsibility

Set of winning states: $\mathtt{WS}_\top := \{s \in S \mid v_\top(\{s\}) = 1\}$

Set of responsible states: $\mathtt{RS}_\top(\mathcal{R}) := \{s \in S \mid \mathcal{R}(v_\top, s) > 0\}$

Proposition

Let $\mathcal{R}$ be a semivalue with $\mathtt{Weights}_0(\mathcal{R}) > 0$. Then we have

$$\mathcal{R}(v_\top, s) > 0 \iff v_\top(\{s\}) = 1, \text{i.e. } \mathtt{RS}_\top(\mathcal{R}) = \mathtt{WS}_\top.$$

Characterisation of the optimistic responsibility

Theorem (Characterisation)

Let $\mathcal{R}$ be a semivalue, then there exists $K \in \mathbb{R}$ such that

- $\forall s \notin \mathtt{WS}_\top, \mathcal{R}(v_\top, s) = 0$
- $\forall s \in \mathtt{WS}_\top, \mathcal{R}(v_\top, s) = K$

and $K = \displaystyle\sum_{k=0}^{n-w} \binom{n-w}{k} p_k$ where $w := |\mathtt{WS}_\top|$.

Additionally, we have $\mathtt{WS}_\top \subseteq \rho$.

Complexity results

Optimistic case:

- Positivity problem: Linear time
- Threshold problem and computation problem: Quadratic time

Pessimistic case:

- Positivity problem: in NP (actually NP-complete)
- Threshold problem: in PSPACE
- Computation problem: in #P

Conclusion

Summary:

- Two notions of responsibility
- Both intuitive and effective (automatic repair)
- Simple characterisation for the optimistic responsibility
- Linear complexity for the optimistic responsibility
- Pessimistic responsibility is more complex

Other contributions:

- Quick implementation (coalition trees, attractor algorithm)
- Article submitted at AAAI
- Recursive responsibility, an inspiring fail
- Generalisation to LTL properties
- A conjecture tested for $n \leq 5$: Banzhaf and Shapley values give equivalent results
- And next...

References

[1] C. Mascle, C. Baier, F. Funke, S. Jantsch, and S. Kiefer, "Responsibility and verification: Importance value in temporal logics," in *LICS*, pp. 1–14, IEEE, 2021.

[2] C. Baier, F. Funke, and R. Majumdar, "A game-theoretic account of responsibility allocation," in *IJCAI*, pp. 1773–1779, ijcai.org, 2021.

[3] C. Baier, C. Dubslaff, F. Funke, S. Jantsch, R. Majumdar, J. Piribauer, and R. Ziemek, "From verification to causality-based explications (invited talk)," in *ICALP*, vol. 198 of *LIPIcs*, pp. 1:1–1:20, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.

[4] P. Dubey, A. Neyman, and R. J. Weber, "Value theory without efficiency," *Mathematics of Operations Research*, vol. 6, no. 1, pp. 122–128, 1981.

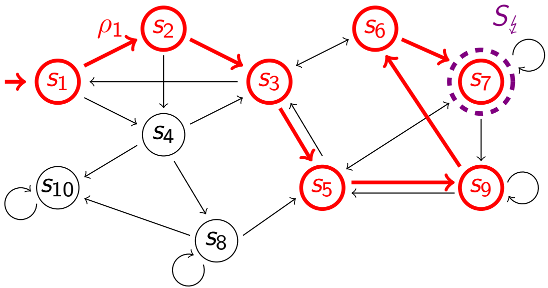
[5] A. Laruelle and F. Valenciano, "Shapley-shubik and banzhaf indices revisited," *Mathematics of Operations Research*, vol. 26, no. 1, pp. 89–104, 2001.

[6] C. Baier and J. Katoen, *Principles of model checking*. MIT Press, 2008.

[7] E. Grädel, W. Thomas, and T. Wilke, eds., *Automata, Logics, and Infinite Games: A Guide to Current Research [outcome of a Dagstuhl seminar, February 2001]*, vol. 2500 of *Lecture Notes in Computer Science*, Springer, 2002.

Thank you for your attention.

Example 2



s	$\mathcal{S}(v_\top, s)$	$\mathcal{S}(v_\bot, s)$	$\mathcal{B}(v_\top, s)$	$\mathcal{B}(v_\bot, s)$
s_1	0.1667	0.0238	0.1667	0.04
s_2	0.1667	0.0238	0.1667	0.04
s_3	0.1667	0.2238	0.1667	0.2
s_4	0	0.0571	0	0.12
s_5	0.1667	0.2238	0.1667	0.2
s_6	0.1667	0.2238	0.1667	0.2
s_7	0	0	0	0
s_8	0	0	0	0
s_9	0.1667	0.2238	0.1667	0.2
s_{10}	0	0	0	0

Figure: Working example 10, run 1