Maurice H. ter Beek
Dejan Ničković (Eds.)

Formal Methods for Industrial Critical Systems

LNCS 12327

25th International Conference, FMICS 2020
Vienna, Austria, September 2–3, 2020
Proceedings

Springer

Lecture Notes in Computer Science 12327

More information about this series at http://www.springer.com/series/7408

Maurice H. ter Beek · Dejan Ničković (Eds.)

Formal Methods for Industrial Critical Systems

25th International Conference, FMICS 2020
Vienna, Austria, September 2–3, 2020
Proceedings

Springer

Editors
Maurice H. ter Beek
ISTI, Consiglio Nazionale delle Ricerche
Pisa, Italy

Dejan Ničković
AIT Austrian Institute of Technology GmbH
Vienna, Austria

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-58297-5 ISBN 978-3-030-58298-2 (eBook)
https://doi.org/10.1007/978-3-030-58298-2

LNCS Sublibrary: SL2 – Programming and Software Engineering

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume contains the papers presented at the 25th International Conference on Formal Methods in Industrial Critical Systems (FMICS 2020), virtually held during September 2–3, 2020. This year the conference was organized under the umbrella of QONFEST, alongside with the 31st International Conference on Concurrency Theory (CONCUR 2020), the 17th International Conference on Quantitative Evaluation of Systems (QEST 2020), and the 18th International Conference on Formal Modeling and Analysis of Timed Systems (FORMATS 2020).

FMICS this year reached its 25th edition, providing for a quarter of a century a forum for researchers who are interested in the development and application of formal methods in industry. FMICS celebrated its 25th birthday with a panel in which the founders and previous chairpersons of the FMICS working group of ERCIM acted as panelists. They recalled the original motivation and beginning of FMICS, shared some success stories, and presented an extensive survey on the past, present, and future of formal methods in research, education, and industry. The detailed report of this study, included in these proceedings, presents an analysis of the opinions of 130 renowned experts in formal methods, as well as thought-provoking position statements on formal methods of 111 of them.

This year we received 26 submissions. Each of these submissions went through a rigorous review process, and as a result each paper received at least three reports. We selected 11 papers for presentation during the conference and inclusion in these proceedings, an acceptance rate of 42%. The conference featured invited talks by Roderick Bloem (Graz University of Technology, Austria, joint keynote with CONCUR and FORMATS), Thomas Henzinger (IST, Austria, joint keynote with CONCUR and QEST), and Stefan Resch (Thales, Austria).

Following a tradition established over the years, Springer provided an award for the best FMICS paper. This year, the reviewers selected the contribution "Verifiable and Scalable Mission-Plan Synthesis for Multiple Autonomous Agents" by Rong Gu, Eduard Enoiu, Cristina Seceleanu, and Kristina Lundqvist for the FMICS 2020 Best Paper Award.

We are grateful to all involved in FMICS 2020. In particular the Program Committee members and subreviewers for their accurate and timely reviewing, all authors for their submissions, and all attendees of the conference for their participation. We thank the general chair of QONFEST, Ezio Bartocci, for providing the logistics that enabled and facilitated the organization of FMICS. We are very grateful to QONFEST platinum sponsor Interchain Foundation, gold sponsors Vienna Center for Logics and Algorithms (VCLA) and Vienna University of Technology, and bronze sponsors ERCIM and Springer.

August 2020

Maurice H. ter Beek
Dejan Ničković

Organization

FMICS Program Chairs

Maurice H. ter Beek	ISTI-CNR, Italy
Dejan Ničković	AIT Austrian Institute of Technology, Austria

FMICS Program Committee

Bernhard Aichernig	TU Graz, Austria
Jiri Barnat	Masaryk University, Czech Republic
Davide Basile	ISTI-CNR, Italy
Carlos Budde	University of Twente, The Netherlands
Rance Cleaveland	University of Maryland, USA
Thao Dang	VERIMAG, France
Michael Dierkes	Rockwell Collins, France
Georgios Fainekos	Arizona State University, USA
Alessandro Fantechi	University of Florence, Italy
Wan Fokkink	Vrije Universiteit Amsterdam, The Netherlands
Maria Del Mar Gallardo	University of Malaga, Spain
Ichiro Hasuo	University of Tokyo, Japan
Klaus Havelund	NASA, USA
Thierry Lecomte	CLEARSY, France
Alberto Lluch Lafuente	Technical University of Denmark, Denmark
Axel Legay	UC Louvain, Belgium
Gabriele Lenzini	University of Luxembourg, Luxembourg
Florian Lorber	Aalborg University, Denmark
Tiziana Margaria	University of Limerick and LERO, Ireland
Radu Mateescu	Inria, France
Franco Mazzanti	ISTI-CNR, Italy
Stefan Mitsch	Carnegie Mellon University, USA
José N. Oliveira	University of Minho, Portugal
Jaco van de Pol	Aarhus University, Denmark
Adam Rogaliewicz	Brno University of Technology, Czech Republic
Markus Roggenbach	Swansea University, UK
Matteo Rossi	Polytechnic University of Milan, Italy
Bernhard Steffen	TU Dortmund University, Germany
Stefano Tonetta	FBK, Italy
Jan Tretmans	TNO, The Netherlands
Andrea Vandin	Sant'Anna School of Advanced Studies, Italy
Tim Willemse	Eindhoven University of Technology, The Netherlands
Kirsten Winter	University of Queensland, Australia
Lijun Zhang	Chinese Academy of Sciences, China

Additional Reviewers

Lukas Armborst	Vojtěch Havlena
Tomáš Fiedor	Iraklis Symeonidis

FMICS Steering Committee

Alessandro Fantechi	University of Florence, Italy
Hubert Garavel	Inria, France
Stefania Gnesi	ISTI-CNR, Italy
Diego Latella	ISTI-CNR, Italy
Tiziana Margaria	University of Limerick and LERO, Ireland
Radu Mateescu	Inria, France
Jaco van de Pol	Aarhus University, Denmark, and University of Twente, The Netherlands

Abstracts of Invited Talks

A Survey of Bidding Games on Graphs

Guy Avni and Thomas A. Henzinger

IST Austria

Abstract. A graph game is a two-player zero-sum game in which the players move a token throughout a graph to produce an infinite path, which determines the winner or payoff of the game. In *bidding games*, both players have budgets, and in each turn, we hold an "auction" (bidding) to determine which player moves the token. In this survey, we consider several bidding mechanisms and study their effect on the properties of the game. Specifically, bidding games, and in particular bidding games of infinite duration, have an intriguing equivalence with *random-turn* games in which in each turn, the player who moves is chosen randomly. We show how minor changes in the bidding mechanism lead to unexpected differences in the equivalence with random-turn games.

Keywords: Bidding games · Richman bidding · Poorman bidding · Mean-payoff · Parity

Extended Abstract

This is an extended abstract of [3].

Two-player zero-sum games on graphs games have deep connections to foundations of logic [14] as well as numerous practical applications, e.g., verification [9], reactive synthesis [13], and reasoning about multi-agent systems [2]. The game proceeds by placing a token on one of the vertices and allowing the players to move it throughout the graph to produce an infinite trace, which determines the winner or payoff of the game.

Several "modes of moving" the token have been studied. The most well-studied mode is *turn-based* games, in which the players alternate turns in moving the token. Other modes include *stochastic games* and *concurrent games*.

We study the *bidding* mode of moving. Abstractly speaking, both players have budgets, and in each turn, we hold an "auction" (bidding) to determine which player moves the token. In this survey, we consider several concrete bidding mechanisms and study the properties of the bidding games that they give rise to.

We emphasize that bidding is a mode of moving the token and bidding games can be studied in combination with any objective. We focus on three objectives: *reachability*, *parity*, and *mean-payoff*. We start by surveying results on reachability games that

This research was supported in part by the Austrian Science Fund (FWF) under grant Z211-N23 (Wittgenstein Award).

were obtained in two papers in the 1990s [10, 11]. We then turn to summarize more recent results on infinite-duration games. Our most interesting results are for mean-payoff games. In a nutshell, reachability bidding games with a specific bidding mechanism called *Richman bidding* were shown to be equivalent to a class of games called *random-turn games* [12]. We show a generalized equivalence between mean-payoff bidding games and random-turn games. While the equivalence for finite-duration games holds only for Richman bidding, for mean-payoff games, equivalences with random-turn games hold for a wide range of bidding mechanisms.

In all mechanisms that we consider, both players simultaneously submit "legal" bids that do not exceed their available budgets, and the higher bidder moves the token. We focus on three orthogonal distinctions between the mechanisms:

Who Pays: We consider *first-price* bidding in which only the higher bidder pays his bid, and *all-pay* bidding in which both players pay their bids.
Who is the Recipient: Two mechanisms were defined in [10]: in *Richman bidding* (named after David Richman), payments are made to the other player, and in *poorman bidding* the payments are made to the "bank" thus the money is lost. A third payment scheme called *taxman* spans the spectrum between Richman and poorman: for a fixed constant $\tau \in [0,1]$, portion τ is paid to the bank and portion $1-\tau$ is paid to the other player.
Which Bids Are Allowed: In *discrete bidding* [8], the budgets are given in coins and the minimal positive bid a player can make is one coin. Unless stated otherwise, we consider *continuous bidding* in which bids can be arbitrarily small.

To state our results we need several definitions. The central quantity in bidding games is the *ratio* between the players' budgets. For $i \in \{1,2\}$, let B_i denote Player i's budget. Player i's ratio is then $B_i/(B_1+B_2)$. The random-turn game that corresponds to a bidding game G w.r.t. $p \in [0,1]$, denoted $\mathtt{RT}(G,p)$, is similar to G only that instead of bidding, in each turn, the player who moves is chosen according to a (biased) coin toss that favors Player 1 with probability p and Player 2 with probability $1-p$. When G is a mean-payoff bidding game, then $\mathtt{RT}(G,p)$ is a mean-payoff stochastic game, and its value is the expected payoff under optimal play. We focus on strongly-connected games in which the value does not depend on the initial vertex.

In this survey we summarize the main results of the following sequence of papers:

- In [4], we show that mean-payoff first-price Richman bidding games are equivalent to un-biased random-turn games: the optimal payoff a player can guarantee in a bidding game G does not depend on the initial ratio and equals the value of $\mathtt{RT}(G,0.5)$.
- In [5], we show that contrary to Richman bidding, the initial ratio matters in mean-payoff first-price poorman bidding games: the optimal payoff a player can guarantee in a bidding game G w.r.t. an initial ratio $r \in (0,1)$ equals the value of $\mathtt{RT}(G,r)$.
- In [6], we unify the previous two results and show that the optimal payoff a player can guarantee in a mean-payoff first-price taxman bidding game G with taxman parameter $\tau \in [0,1]$, and initial ratio $r \in (0,1)$ equals the value of $\mathtt{RT}\left(G, \frac{r+\tau\cdot(1-r)}{1+\tau}\right)$.

– In [1], we study qualitative infinite-duration discrete Richman bidding games and study tie-breaking mechanisms that guarantee determinacy.
– In [7], we study reachability all-pay poorman bidding. Even though they are technically significantly more challenging than reachability first-price bidding games, we can still obtain simple yet powerful results on this model.

Acknowledgments. We would like to thank all our collaborators Milad Aghajohari, Ventsislav Chonev, Rasmus Ibsen-Jensen, Ismäel Jecker, Petr Novotný, Josef Tkadlec, and ore Žikelić; we hope the collaboration was as fun and meaningful for you as it was for us.

References

1. Aghajohari, M., Avni, G., Henzinger, T.A.: Determinacy in discrete-bidding infinite-duration games. In: Proceedings of the 30th CONCUR, LIPIcs, vol. 140, pp. 20:1–20:17 (2019)
2. Alur, R., Henzinger, T.A., Kupferman, O.: Alternating-time temporal logic. J. ACM **49**(5), 672–713 (2002)
3. Avni, G., Henzinger, T.A.: A survey of bidding games on graphs. In: Proceedings of the 31st CONCUR, LIPIcs, vol. 171 (2020)
4. Avni, G., Henzinger, T.A., Chonev, V.: Infinite-duration bidding games. J. ACM **66**(4), 31:1–31:29 (2019)
5. Avni, G., Henzinger, T.A., Ibsen-Jensen, R.: Infinite-duration poorman-bidding games. In: Christodoulou, G., Harks, T. (eds.) WINE 2018. LNCS, vol, 11316, pp 21–36. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04612-5_2
6. Avni, G., Henzinger, T.A., Žikelić.: Bidding mechanisms in graph games. In: Proceedings of the 44th MFCS, LIPIcs, vol. 138, pp. 11:1–11:13 (2019)
7. Avni, G., Ibsen-Jensen, R., Tkadlec, J.: All-pay bidding games on graphs. In: Proceedings of the 34th AAAI (2020)
8. Develin, M., Payne, S.: Discrete bidding games. Electron. J. Comb. **17**(1), R85 (2010)
9. Emerson, A.E., Jutla, C.S., Sistla, P.A.: On model-checking for fragments of μ-calculus. In: Proceedings of the 5th CAV, pp. 385–396 (1993)
10. Lazarus, A.J., Loeb, D.E., Propp, J.G., Stromquist, W.R., Ullman, D.H.: Combinatorial games under auction play. Games Econ. Behav. **27**(2), 229–264 (1999)
11. Lazarus, A.J., Loeb, D.E., Propp, J.G., Ullman, D.: Richman games. Games No Chance **29**, 439–449 (1996)
12. Peres, Y., Schramm, O., Sheffield, S., Wilson, D.B.: Tug-of-war and the infinity laplacian. J. Amer. Math. Soc. **22**, 167–210 (2009)
13. Pnueli, A., Rosner, R.: On the synthesis of a reactive module. In: Proceedings of the16th POPL, pp. 179–190 (1989)
14. Rabin, M.O.: Decidability of second order theories and automata on infinite trees. Trans. AMS, **141**, 1–35 (1969)

Applying Formal Methods in Industrial Railway Applications at Thales

Stefan Resch

Thales Austria GmbH, Vienna, Austria
stefan.resch@thalesgroup.com
www.thalesgroup.com

1 Introduction

The application of formal methods is intended to improve software quality. While common tools that perform static code analysis such as Coverity [3] are well known and applied in the industry, this talk presents three use cases at Thales that leverage formal methods and the according tools to an even larger extend.

The conditions for applying formal method tools for safety critical software in the railway domain are defined by the CENELEC EN 50128 [2] standard. This standard highly recommends the use of formal methods for safety relevant projects for the highest safety integrity levels (SIL) of SIL3 and SIL4. The CENELEC EN 50128 categorizes tools into three different types from T1 to T3 depending on whether they can introduce faults into the safety critical software. Here tools related to formal methods usually are of type T2, since they are used for verification and may fail to identify a fault, but cannot introduce them themselves. This requires that, when used in a safety relevant project, (1) the selection of the tool and its assigned category are justified, (2) potential failures are identified, as well as measures to handle such failures, (3) the tool has a specification or handbook, (4) it is ensured that only justified versions of the tools are used and (5) this justification is also performed when switching versions of the tool.

Each of the use cases presented in the following sections has a different focus, illustrating the vast potential of formal methods. They demonstrate that while formal methods may pose a significant overhead at design time they provide an overall benefit when used in the right context.

2 Use Case: ERTMS Hybrid Level 3

ERTMS Hybrid Level 3 is a concept that enables an increase of track capacity in the railway network by reusing regular signaling and interlocking interfaces to integrate into existing systems while benefiting from the continuous supervision of the trains in network via radio. [4] The specification of this concept was analyzed and validated

using a formal model in B [1] and executed at runtime in Pro-B [7]. It was subsequently successfully used in a field demonstration controlling real trains.

3 Use Case: Checking ETCS Level 1 Line Side Data

One of the challenges when deploying the new European Train Control System (ETCS) lines lies in the complexity of the configuration data. We use the tool Emerald for checking the ETCS Level 1 lineside configuration data against rules derived from a customer specific "Book of Rules". Emerald internally uses B and Pro-B and is developed and maintained by Thales, since it is a highly specific application. The advantage of Emerald's approach is that many data preparations errors can be caught early on during development, before starting the verification phase of the projects. This tool is actively being used in the current projects.

4 Use Case: TAS Control Plaform

The method of model-checking was used to model and develop fault-tolerant and safety-critical modules for TAS Control Platform, a platform for railway control applications up to safety integrity level (SIL) 4. [8] By model-checking modules in TLA+ [5] and PlusCal [6] core safety and liveness properties of a distributed fault-tolerant protocol were analyzed. A translator from PlusCal to C bridges the gap between model and code.

References

1. Abrial, J.R.: The B Book - Assigning Programs to Meanings. Cambridge University Press, August 1996
2. CENELEC: EN 50128-Railway Applications: Software for Railway Control and Protection Systems. European Committee for Electrotechnical Standardization (2011)
3. Coverity Scan. https://scan.coverity.com/
4. Hansen, D., Leuschel, M., Körner, P., Krings, S., Naulin, T., Nayeri, N., Schneider, D., Skowron, F.: Validation and real-life demonstration of ETCS hybrid level 3 principles using a formal B model. Int. J. Softw. Tools Technol. Transf. 1–18 (2020)
5. Lamport, L.: Specifying systems: the TLA+ language and tools for hardware and software engineers. Addison-Wesley Longman Publishing Co., Inc. (2002)
6. Lamport, L.: The PlusCal algorithm language. Theoretical Aspects of Computing-ICTAC 2009, pp. 36–60 (2009)
7. Leuschel, M., Butler, M.: ProB: an automated analysis toolset for the B method. Int. J. Soft. Tools Technol. Transf. **10**(2), 185–203 (2008)
8. Resch, S., Paulitsch, M.: Using TLA+ in the development of a safety-critical fault-tolerant middleware. In: 2017 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), pp. 146–152. IEEE (2017)

Safe Reinforcement Learning using Probabilistic Shields

Nils Jansen[1], Bettina Könighofer[2,3], Sebastian Junges[4], Alexandru Serban[1], and Roderick Bloem[2]

[1] Radboud University, Nijmegen, The Netherlands
[2] Graz University of Technology, Institute IAIK, Austria
[3] Silicon Austria Labs, TU-Graz SAL DES Lab, Austria
[4] RWTH Aachen University, Aachen, Germany

Abstract. We target the efficient construction of a safety shield for decision making in scenarios that incorporate uncertainties. Markov decision processes (MDPs) are prominent models to capture such planning problems. This paper concerns the efficient construction of a safety shield for reinforcement learning. We specifically target scenarios that incorporate uncertainty and use Markov decision processes (MDPs) as the underlying model to capture such problems. Reinforcement learning (RL) is a machine learning technique that can determine near-optimal policies in MDPs that may be unknown before exploring the model. However, during exploration, RL is prone to induce behavior that is undesirable or not allowed in safety- or mission-critical contexts. We introduce the concept of a probabilistic shield that enables RL decision-making to adhere to safety constraints with high probability. We employ formal verification to efficiently compute the probabilities of critical decisions within a safety-relevant fragment of the MDP. These results help to realize a shield that, when applied to an RL algorithm, restricts the agent from taking unsafe actions, while optimizing the performance objective. We discuss tradeoffs between sufficient progress in the exploration of the environment and ensuring safety. In our experiments, we demonstrate on the arcade game PAC-MAN that the learning efficiency increases as the learning needs orders of magnitude fewer episodes.

1 Introduction

A major open challenge for systems employing reinforcement learning (RL) is the *safety* of decision-making. In particular during the exploration phase – when an agent chooses random actions in order to examine its surroundings – it is important to avoid actions that may cause unsafe outcomes. The area of *safe exploration* investigates how RL agents may be forced to adhere to safety requirements during this phase. A suite of methods that deliver theoretical guarantees are so-called *safety-shields*. Shields prevent an agent from taking unsafe actions at runtime. To this end, the performance objective is extended with a constraint specifying that unsafe states should *never* be visited. This new safety objective ensures there are no violations during the exploration phase.

We propose to incorporate constraints that enforce safety violations to occur *only with small probability*. If an action increases the probability of a safety violation by more than a threshold δ with respect to the optimal safety probability, the shield blocks the action. Consequently, an agent augmented with a shield is *guided* to satisfy the safety objective during exploration (or as long as the shield is used). The shield is *adaptive* with respect to δ, as a high value for δ yields a stricter shield, a smaller value a more permissive shield. The value for δ can be changed on-the-fly, and may depend on the individual minimal safety probabilities at each state. Moreover, in case there is not suitable safe action with respect to δ, the shield can always pick the optimal action as a fallback. We base our formal notion of a probabilistic shield on MDPs, which constitute a popular modeling formalism for decision-making under uncertainty and is widely used in model-based RL. We assess safety by means of probabilistic *temporal logic constraints* that limit, for example, the probability to reach a set of critical states in the MDP. In order to assess the risk of one action, we (1) construct a behavior model for the environment using model-based RL. By plugging this model into any concrete scenario, we obtain an MDP. To construct the shield, we (2) use a model-based verification technique known as *model checking* that assesses whether a system model satisfies a specification. In particular, we obtain precise *safety probabilities of any possible decision* within the MDP. These probabilities can be looked up efficiently and compared to the threshold δ. The shield then readily (3) augments either model-free or model-based RL.

2 Problem Statement

Setting. We define a setting where one controllable agent (the *avatar*) and a number of uncontrollable agents (the *adversaries*) operate within an *arena*. The arena is a compact, high-level description of the underlying model. From this arena, the potential states and actions of all agents may be inferred. For safety considerations, the reward structure can be neglected, effectively reducing the state space for our model-based safety computations. Some (combinations of) agent positions are safety-critical, as they e.g., correspond to collisions or falling off a cliff. A safety property may describe reaching such positions, or use any other property expressible in (the safety fragment of) temporal logic. To encode a *performance criterion*, we associate edges of the arena with a *token* function, indicating the status of some edge. Tokens can be (de-) activated and have an associated *reward* earned upon taking edges with an active token.

Application. We designed the setting to be applicable to a series of scenarios. As an example, take a factory floor plan with several corridors. The nodes of the arena describe crossings, and the edges the corridors with machines. The adversaries are (possibly autonomous) transporters moving parts within the factory. The avatar models a service unit moving around and inspecting machines where an issue has been raised (as indicated by a token), while accounting for the behavior of the adversaries. Corridors might be to narrow for multiple (facing) robots, which poses a safety critical situation. Several notions of cost can be induced by the tokens, either as long as they

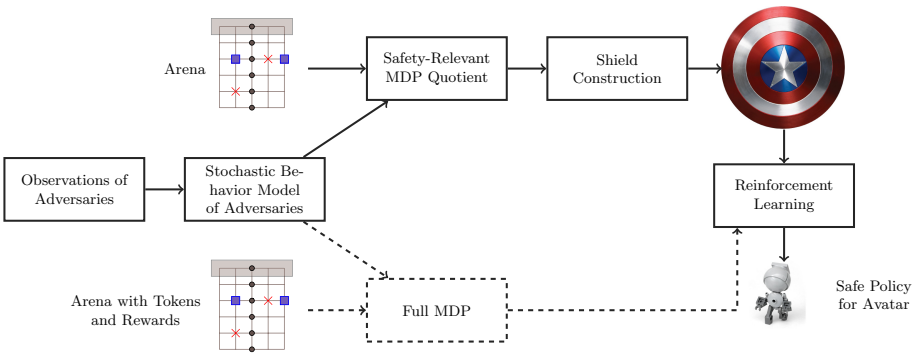


Fig. 1. Workflow of the Shield Construction

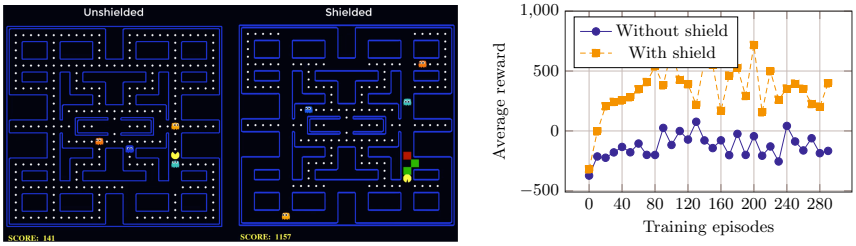
are present (costs of a broken machine) or for removing the tokens (costs for inspecting the machine).

Problem. Consider an environment described by an arena as above and a safety specification. We assume stochastic behaviors for the adversaries, e.g, obtained using RL in a training environment. The underlying model is then an MDP: the avatar executes an action, and upon this execution the next exact positions (the state of the system) are determined stochastically.

We compute a δ-shield that prevents avatar decisions that violate this specification by more than a threshold δ with respect to the optimal safety probability. We evaluate the shield using a model-based or model-free RL avatar that aims to optimize the performance. The shield therefore has to handle an intricate tradeoff between strictly focussing on (short and midterm) safety and performance.

3 Constructing Shields for MDPs

We outline the workflow of our approach in Figure 1. We employ a separation of concerns between the model-based shield construction and potentially model-free RL. First, based on observations in arbitrary arenas, we construct a general (stochastic) *behavior model* for each adversary. Combining these models with a concrete arena yields an MDP. At this point, we ignore the token function (and necessarily the unknown reward function), so the MDP may be seen as a quotient of the full MDP that models the real system within which we only assess safe behavior. We therefore call the MDP the *safety-relevant quotient*. The real scenario incorporates the *token function*. Rewards may be known or only be observed during learning. The underlying *full MDP* including tokens constitutes an *exponential blowup of the safety-relevant quotient*, rendering probabilistic model checking or planning practically infeasible. In the workflow, using the safety-relevant MDP, we construct a *shield* using probabilistic model checking. RL now aims to maximize the reward according to the original scenario, while unsafe actions are blocked by the shield.



(a) Video still for PAC-MAN (b) Scores for PAC-MAN

Fig. 2. Scenario and results for PAC-MAN

4 Implementation and Experiments

For our experiments we consider the arcade game PAC-MAN. The task is to eat *food* in a *maze* and not get eaten by *ghosts*. We model each instance of the game as an arena, where PAC-MAN is the avatar and the ghosts are adversaries. The safety specification is that the avatar *not gets eaten* with a high probability. Tokens represent the food at each position in the maze, such that food is either present or already eaten. We learn the ghost behavior from the original PAC-MAN game for each ghost. Transferring the resulting stochastic behavior to any arena (without tokens) yields the safety-relevant MDP. For that MDP, we compute a shield via the model checker `Storm` for a horizon of 10 steps. Our implementation uses an approximate Q-learning agent (using $\alpha = 0.2$, $\gamma = 0.8$ and $\varepsilon = 0.05$) with the following feature vector: (1) how far away the next food is, (2) whether a ghost collision is imminent, and (3) whether a ghost is one step away. Figure 2(a) shows a still of a series of *videos* we created[1]. Each video compares how RL performs either shielded or unshielded on a PAC-MAN instance. In the shielded version, we indicate the risk of potential decisions by the colors green (low), orange (medium), and red (high). Figures 2(b) depict the *scores* obtained during RL. We see a large difference in scores due to the fact that PAC-MAN is often saved by the shield.

5 Conclusion and Future Work

We developed the concept of shields for MDPs. Utilizing probabilistic model checking, we maintained probabilistic safety measures during reinforcement learning. We addressed inherent scalability issues and provided means to deal with typical trade-off between safety and performance. Our experiments showed that we improved the state-of-the-art in safe reinforcement learning. For future work, we will extend the applications to more arcade games and employ deep recurrent neural networks as means of decision-making. Another interesting direction is to explore (possibly model-free) learning of shields, instead of employing model-based model checking.

[1] http://shieldrl.nilsjansen.org

Contents

Temporal Logic and Model Checking