Theoretical Computer Science 467 (2013) 53–67

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Optimal time-abstract schedulers for CTMDPs and continuous-time Markov games

Markus N. Rabe [a,*], Sven Schewe [b]

[a] Universität des Saarlandes, Germany

[b] University of Liverpool, United Kingdom

ARTICLE INFO

Article history:
Received 26 January 2012
Received in revised form 25 September 2012
Accepted 2 October 2012
Communicated by D. Sannella

Keywords:
Continuous-time Markov decision processes
Continuous-time Markov games
Optimal control
Time-bounded reachability

ABSTRACT

We study time-bounded reachability in continuous-time Markov decision processes (CTMDPs) and games (CTGs) for time-abstract scheduler classes. Reachability problems play a paramount rôle in probabilistic model checking. Consequently, their analysis has been studied intensively, and approximation techniques are well understood. From a mathematical point of view, however, the question of approximation is secondary compared to the fundamental question whether or not optimal control exists.

In this article, we demonstrate the existence of optimal schedulers for the time-abstract scheduler classes for CTMDPs. For CTGs, we distinguish two cases: the simple case where both players face the same restriction to use time-abstract strategies (symmetric CTGs) and the case where one player is a completely informed adversary (asymmetric CTGs). While for the former case optimal strategies exist, we prove that for asymmetric CTGs there is not necessarily a scheduler that attains the optimum.

It turns out that for CTMDPs and symmetric CTGs optimal time-abstract schedulers have an amazingly simple structure: they converge to a memoryless scheduling policy after a finite number of steps. This allows us to compute time-abstract strategies with finite memory.

1. Introduction

Markov decision processes (MDPs) are a framework that incorporates both nondeterministic and probabilistic choices. They are used in a variety of applications such as the control of manufacturing processes or queueing systems [1–3]. We study a real time version of MDPs, continuous-time Markov decision processes (CTMDPs), which are a natural formalism for modelling in scheduling [4,2] and stochastic control theory [1]. Also, CTMDPs are the most simple representative of a number of equivalent model types (e.g. [5,6]).

The analysis of CTMDPs usually concerns the different possibilities to resolve the nondeterminism. Typical questions cover qualitative as well as quantitative properties, such as: "Can the nondeterminism be resolved by a scheduler such that a predefined property holds?" or respectively "Which scheduler optimises a given objective function?".

Nondeterminism is either always hostile or always supportive in CTMDPs. Continuous-time Markov games (CTGs) provide a generalisation of CTMDPs by partitioning the set of locations into locations where the nondeterminism is resolved angelically (supportive nondeterminism) and locations where the nondeterminism is resolved demonically (hostile nondeterminism) [7–10].

* Corresponding author. Tel.: +49 6813025637.
E-mail address: markus.norman.rabe@googlemail.com (M.N. Rabe).

0304-3975/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2012.10.001

In this article, we study the *time-bounded reachability probability* problem [11,8,12–14,2,15] in CTMDPs and CTGs. For CTMDPs, time-bounded reachability is the standard control problem to construct a scheduler that controls the Markov decision process such that the likelihood of reaching the goal region within a given time bound is maximised/minimised, and to determine the probability. For games, both the angelic and the demonic nondeterminism needs to be resolved at the same time.

The obtainable quality of the resulting scheduler naturally depends on its power to observe the run of the system and on its ability to store and process this information. The commonly considered schedulers classes and their basic connections have been discussed in the literature [13,16]. Thereof, we focus on schedulers that have no direct access to time, the time-abstract schedulers. Within time-abstract schedulers, the classes that can observe the history, its length, or nothing at all – marked H (for history-dependent), C (for counting), and P (for positional) – are distinguished.

These classes form a simple inclusion hierarchy (H $\supset$ C $\supset$ P) and in general they yield different maximum reachability probabilities. However, it is known that for *uniform* CTMDPs – CTMDPs that have a uniform transition rate λ for all their actions – the maximum time-bounded reachability probabilities of classes H and C coincide [11].

Given its importance for applications like model checking, the time-bounded reachability problem for Markov decision processes has been intensively studied. While previous research focused on *approximating* optimal scheduling policies [11,12,17,14], the existence of optimal schedulers for time-abstract scheduler classes has been demonstrated in our technical report and a subsequent publication [18,19], on which Section 3 is partly based. Brázdil et al. [8] have independently provided a similar result for continuous-time Markov games.

Contributions. In Section 3, we establish the existence of optimal counting (C) and optimal history dependent (H) schedulers in *uniform* CTMDPs and lift it to general CTMDPs. We also show that one cannot improve the quality of optimal scheduling by using randomised schedulers.

Our solution builds on the observation that, if time has almost run out, we can use a greedy strategy that optimises our chances to reach our goal in fewer steps rather than in more steps. We show that a memoryless greedy scheduler exists, and is indeed optimal after a certain step bound. The existence of an optimal scheduler is then implied by the finite number of remaining candidates—it suffices to search among those schedulers that deviate from the greedy strategy only in a finite preamble. We demonstrate these techniques on an example in Section 3.5.

In Section 4, we discuss the impact of restricting one player to time-abstract strategies while the other player has full access to time (*asymmetric* CTGs). In contrast to the single player case, the time-abstract player can benefit from randomisation and from considering the history. As a result, it is not guaranteed that there are optimal strategies for the time-abstract player in asymmetric CTGs.

Finally, we show in Section 5 that our lifting argument naturally extends to symmetric CTGs, with the notable exception of counting schedulers, which may benefit from randomisation.

2. Preliminaries

A *continuous-time Markov game* $\mathcal{M}$ is a tuple $(L, L_A, L_D, Act, \mathbf{R}, \nu)$ with a finite set of locations $L = L_A \dot\cup L_D$, a finite set of actions Act, a rate matrix $\mathbf{R} : (L \times Act \times L) \to \mathbb{Q}_{\geqslant 0}$, and an initial probability distribution ν over the locations. We define the total exit rate for a location l and an action a as $\mathbf{R}(l, a, L) = \sum_{l' \in L} \mathbf{R}(l, a, l')$. We require that, for all locations $l \in L$, there must be an action $a \in Act$ such that $\mathbf{R}(l, a, L) > 0$, and we call such actions *enabled*. We define $Act(l)$ to be the set of enabled actions in location l. If there is only one enabled action per location, a CTG $\mathcal{M}$ is a continuous-time Markov chain [20]. As usual, we assume the goal region to be absorbing, and we use $\mathbf{P}(l, a, l') = \frac{\mathbf{R}(l,a,l')}{\mathbf{R}(l,a,L)}$ to denote the time-abstract transition probability.

We are especially interested in CTMDPs, that is CTGs with only one player ($L = L_A$ or $L = L_D$), as they constitute an important class on their own. In order to keep our results accessible to a broader audience, we will first prove our results on CTMDPs in Section 3 and we will provide the generalisation to full CTGs in Section 5. For an example automaton on which we also demonstrate our method, we refer to Section 3.5.

Uniform CTGs. We call a CTG uniform with rate λ if, for every location l and actions $a \in Act(l)$, the total exit rate $\mathbf{R}(l, a, L)$ is λ. In this case the probability $p_{\lambda t}(n)$ that there are exactly n discrete events (transitions) in time t is Poisson distributed: $p_{\lambda t}(n) = e^{-\lambda t} \cdot \frac{(\lambda t)^n}{n!}$.

We define the *uniformisation* of a CTG $\mathcal{M} = (L, L_A, L_D, Act, \mathbf{R}, \nu)$ as the uniform CTG $\mathcal{U} = (L^{\mathcal{U}}, L_A^{\mathcal{U}}, L_D^{\mathcal{U}}, Act, \mathbf{R}^{\mathcal{U}}, \nu^{\mathcal{U}})$. The locations of $\mathcal{U}$ contain an additional *unobservable* copy $l_{\mathcal{U}}$ of every location l: $L^{\mathcal{U}} = \bigcup_{l \in L} \{l, l_{\mathcal{U}}\}$. The partition into the sets $L_A^{\mathcal{U}} = \bigcup_{l \in L_A} \{l, l_{\mathcal{U}}\}$ and $L_D^{\mathcal{U}} = \bigcup_{l \in L_D} \{l, l_{\mathcal{U}}\}$ carries over from $\mathcal{M}$. The new rate matrix is defined as $\mathbf{R}^{\mathcal{U}}(l, a, l') = \mathbf{R}(l, a, l')$ and $\mathbf{R}^{\mathcal{U}}(l_{\mathcal{U}}, a, l') = \mathbf{R}(l, a, l')$ for all $l, l' \in L$ and for all $a \in Act$, $\mathbf{R}^{\mathcal{U}}(l, a, l_{\mathcal{U}}) = \lambda - \mathbf{R}(l, a, L)$ and $\mathbf{R}^{\mathcal{U}}(l_{\mathcal{U}}, a, l_{\mathcal{U}}) = \lambda - \mathbf{R}(l, a, L)$ for all locations $l \in L$. Finally, the initial distribution is extended by 0-entries for the new locations: $\nu^{\mathcal{U}}(l) = \nu(l)$ if $l \in L$, and $\nu^{\mathcal{U}}(l) = 0$ otherwise.

The idea behind this uniformisation is quite simple: in contrast to the traditional uniformisation, we introduce additional copies for each location that indicate whether the last transition that was taken was only due to the uniformisation. If we enter an *observable* (i.e. not unobservable) location, we would have done this also in the non-uniformised automaton. We

refer to Section 3.5 for an example of this construction. (Note that we left out unobservable locations that are not relevant for the example.)

Paths. A *timed path* of a CTG $\mathcal{M}$ is a finite sequence in $(L \times Act \times \mathbb{R}_{\geqslant 0})^* \times L = Paths(\mathcal{M})$. We write $l_0 \xrightarrow{a_0,t_0} l_1 \xrightarrow{a_1,t_1} \cdots \xrightarrow{a_{n-1},t_{n-1}} l_n$ for a timed path π, and we require $t_{i-1} < t_i$ for all $i < n$. The t_i denote the system's time when the events happen. The corresponding *time-abstract path* is defined as $l_0 \xrightarrow{a_0} l_1 \xrightarrow{a_1} \cdots \xrightarrow{a_{n-1}} l_n$. We use $Paths_{abs}(\mathcal{M})$ to denote the set of all such projections and $|\cdot|$ to count the number of actions in a path. Concatenation of paths π, π' will be written as $\pi \circ \pi'$ if the last location of π is the first location of π'.

Schedulers. The system's behaviour is not fully determined by the CTG, we additionally need a scheduler (also called policy) that resolves the nondeterminism. We usually consider a scheduler to consist of the two players' strategies (that is, a strategy is a partial scheduler) that control the behaviour once the system is in one of their locations—L_A denotes the angelic player's locations, while L_D contains the demonic player's locations. We refer to their strategies as $S_X : Paths^X(\mathcal{M}) \to Dist(Act)$, where $X \in \{A, D\}$ and $Paths^X(\mathcal{M})$ is the set of paths that end with a location in L_X. We use the natural one-to-one mapping between a pair of strategies $\mathcal{S}_A$, $\mathcal{S}_D$ for the two players and the combined scheduler $\mathcal{S}_{A+D}$ for the CTG:

$$\mathcal{S}_{A+D}(\pi) = \begin{cases} \mathcal{S}_A(\pi) & \text{if } \pi \in Paths^A(\mathcal{M}) \\ \mathcal{S}_D(\pi) & \text{if } \pi \in Paths^D(\mathcal{M}). \end{cases}$$

When analysing properties of a CTG, such as the reachability probability, we quantify over a class of strategies. In this article, we consider the following scheduler classes, which differ in their power to observe and distinguish events:

- *Timed history-dependent* (TH) schedulers that map the system's history and the current time to decisions, $Paths(\mathcal{M}) \times \mathbb{R}_{\geqslant 0} \to \mathcal{C}$
- *Time-abstract history-dependent* (H) schedulers that map time-abstract paths to decisions, $Paths_{abs}(\mathcal{M}) \to \mathcal{C}$
- *Time-abstract hop-counting* (C) schedulers that map locations and the length of the paths to decisions, $L \times \mathbb{N} \to \mathcal{C}$
- *Positional* (P) or memoryless schedulers that map locations to decisions. $L \to \mathcal{C}$

Choices $\mathcal{C}$ are either randomised (R), in which case $\mathcal{C}$ is the set of distributions over enabled actions Act, or are restricted to deterministic (D) choices, that is $\mathcal{C} = Act$. Where it is necessary to distinguish randomised and deterministic versions we will add a postfix to the scheduler class, for example HD and HR.

For a given timed path π and finitely many intervals that form a partition $\mathcal{J}$ of time, called the *cylindrical set of paths* (or cylindrification) $[\pi]_{\mathcal{J}}$ contains all paths whose transition times are in the same equivalence classes (w.r.t. $\mathcal{J}$) as those of π. A scheduler is called *cylindrical*, if for some partition $\mathcal{J}$ of time into intervals it has constant decisions for all paths π, π' having the same cylindrification ($[\pi]_{\mathcal{J}} = [\pi']_{\mathcal{J}}$).

Probability space for Markov games. We define the probability space for a sufficiently large interval $[0, t_{\max}]$, $t_{\max} \in \mathbb{R}_{\geq 0}$, on finite paths of a CTG $\mathcal{M}$ under a measurable scheduler in 2 steps: First, we define the probability space on finite paths of $\mathcal{M}$ under a cylindrical scheduler (in the interval $[0, t_{\max}]$) as the completion of the trivial probability space on cylindrical sets of paths. Second, another completion on the class of cylindrical strategies then yields the full class of (measurable) TH strategies. For a scheduler $\mathcal{S}$, we use $Pr_{\mathcal{S}}$ to denote the corresponding probability measure on paths of $\mathcal{M}$. See [10] for details.

Note that the resulting probability space is defined on *finite* paths that have no continuation in the time interval $[0, t_{\max}]$, unlike the more common construction via the Borel σ-algebra [16]. Thus, for the definition of the reachability probability (see below), it is important to consider the probability that for a finite path (or set thereof) there is no further transition after their last transition until $t_{\max}$.

Time-bounded reachability probability. We consider the *time-bounded reachability probability* problem. That is, given a Markov game $\mathcal{M}$, a goal region $G \subseteq L$, and a time bound $T \in \mathbb{R}_{\geq 0}$, we are interested in the set of paths $reach_{\mathcal{M}}(G, T)$ that reach a location in the goal region precisely[1] at time T:

$$reach_{\mathcal{M}}(G,T) = \left\{ \sigma \in Paths(\mathcal{M}) \mid \sigma = l_0 \xrightarrow{a_0,t_0} l_1 \ldots l_n \text{ with } l_n \in G \wedge t_{n-1} \leq T \text{ or } \exists i < n.\ l_i \in G \wedge t_{i-1} \leq T \leq t_i \right\}.$$

We are particularly interested in *optimising* its probability and in finding the corresponding pair of strategies: $\sup_{\mathcal{S}_A \in \mathrm{TP}} \inf_{\mathcal{S}_D \in \mathrm{TP}} Pr_{\mathcal{S}_{A+D}}(reach_{\mathcal{M}}(G,T))$, which is commonly referred to as the *maximum* time-bounded reachability probability problem in the case of CTMDPs.

We use 'max' instead of 'sup' ('min' and 'inf', respectively) to indicate that this value is taken for some *optimal scheduler* $\mathcal{S}$ of this class.

[1] Note that we could significantly simplify this notation by using the assumption that the goal regions are absorbing. We restrain from doing so, however, in order to simplify the definitions to come.

Given a scheduler $\mathfrak{s}$, we define $Pr^G_{\mathfrak{s}}(l, t)$ to be the probability under this scheduler of being in the goal region G at time T assuming we start in location l and that $T - t$ time units have passed already (or, t time units are left). That is, $Pr^G_{\mathfrak{s}}(l, t)$ is the *conditional* probability $Pr_{\mathfrak{s}}\,(reach_{\mathcal{M}}(G, T) \mid reach_{\mathcal{M}}(\{l\}, T - t))$. Using this definition, we introduce the following notations:

- $Pr^G_{\mathfrak{s}}(t) = \sum_{l \in L} \nu(l) Pr^G_{\mathfrak{s}}(l, t)$ $(= Pr_{\mathfrak{s}}(reach_{\mathcal{M}}(G, t)))$ denotes the probability of reaching the goal region G assuming that only time t is left,
- $f : L \times [0, T] \to [0, 1]$ denotes the *optimal* probability to be in the goal region at the time bound, assuming that we start in location l and that only t time units are left: $f(l, t) = \sup_{\mathfrak{s}_A \in \mathrm{TP}} \inf_{\mathfrak{s}_D \in \mathrm{TP}} Pr^G_{\mathfrak{s}_{A+D}}(l, t)$,
- $Pr^G_{\mathfrak{s}}(t; k)$ denotes the probability of reaching the goal region G in time t *and* in at most k discrete steps, and
- $PR_{\mathfrak{s}}(\pi, t)$ is the probability to traverse the time-abstract path π within time t.

Step probability vector. Given a scheduler $\mathfrak{s}$ and a location l for a CTG $\mathcal{M}$, we define the *step probability vector* $d_{l,\mathfrak{s}}$ of infinite dimension. An entry $d_{l,\mathfrak{s}}[i]$ for $i \geq 0$ denotes the conditional probability to reach goal region G from location l with the i-th step, assuming that exactly i steps occur:

$$d_{l,\mathfrak{s}}[i] = Pr_{\mathfrak{s}}\Big(last(\sigma) \in G,\ \text{but } last(\sigma \downarrow_{i-1}) \notin G \,\Big|\, |\sigma| = i\Big).$$

3. Optimal time-abstract schedulers

In this section, we show that for CTMDPs there are *optimal* schedulers for the time-abstract scheduler classes (CD, CR, HD, and HR). Moreover, we prove that there are optimal schedulers that become positional after a small number of steps. We also show that randomisation does not yield any advantage: deterministic schedulers are as good as randomised ones. This also provides a procedure to precisely determine the time-bounded reachability probability, because we can now reduce this problem to computing the time-bounded reachability probability of a continuous-time Markov chain [21].

Our proof consists of two parts. We first consider the class of uniform CTMDPs, because we can use Poisson distributions to describe the number of steps taken within a given time bound. For uniform CTMDPs it is already known that the supremum over the bounded reachability collapses for all time-abstract scheduler classes from CD to HR [11]. It therefore suffices to show that there is a CD scheduler which takes this value.

In the non-uniform case the time-abstract path contains more information about the remaining time than its length only, and bounded reachability of history-dependent and counting schedulers usually deviate (see [11]). Thus, in a second step, we extend this result to general CTMDPs by constructing optimal CD and optimal HD schedulers that, as in the uniform case, converge against a positional scheduler after a finite number of steps.

We start this section with the introduction of *greedy schedulers*, that are HD schedulers that favour to reach G in a small number of steps over the possibility to reach G with a larger number of steps. The positional schedulers against which the optimal CD and HD schedulers converge are such greedy schedulers.

3.1. Greedy schedulers

The objective we consider in this subsection is to maximise, for a goal region G and time bound T, the time-bounded reachability probability $Pr^G_{\mathfrak{s}}(T)$ with respect to a particular scheduler class such as HD. Unfortunately, this optimisation problem is rather difficult to solve. Therefore, we start with analysing the special case of having little time left and establish the existence of simple optimal strategies for this case. If the remaining time converges to 0, we can exploit that the probability to take two or more steps declines faster than the probability to take exactly one further step. Thus, any increase of the likelihood of reaching the goal region sooner dominates the potential impact of reaching it later in this case.

Time-abstract schedulers have no direct access to the time, but they can infer a probability distribution over the remaining time from the time-abstract history (or its length). It turns out that after sufficiently many steps, the probability to be in a time point sufficiently close to T is very high, compared to the probability to have more time left. (The distribution of time passed is an Erlang distribution, and the distribution of time passed *provided* that time has not run out is therefore a truncated Erlang distribution, see Fig. 1). Thus, also for time-abstract schedulers, reaching the goal region in few steps dominates the potential impact of reaching it in many steps (after a sufficiently long preamble).

This motivates the introduction of greedy schedulers. Schedulers are called greedy, if they (greedily) look for short-term gain, and favour it over any long-term effect. A notion of greedy schedulers that optimise the reachability within the first k steps have been exploited in the efficient analysis of CTMDPs [11]. To understand the principles of optimal control, however, a simpler form of greediness proves to be more appropriate: We call an HD scheduler *greedy* if it maximises the step probability vector of every location l with respect to the lexicographic order (for example $(0, 0.2, 0.3, \dots) >_{lex} (0, 0.1, 0.4, \dots)$). To prove the existence of greedy schedulers, we draw from the fact that the supremum $d_l = \sup_{\mathfrak{s} \in HD} d_{l,\mathfrak{s}}$ obviously exists, where the supremum is to be read as a supremum with respect to the lexicographic order. An action $a \in Act(l)$ is called *greedy* for a location $l \notin G$ if it satisfies $shift(d_l) = \sum_{l' \in L} \mathbf{P}(l, a, l') d_{l'}$, where $shift(d_l)$ shifts the vector by one position (that is, $shift(d_l)[i] = d_l[i + 1]$ $\forall i \in \mathbb{N}$). For locations l in the goal region G, all enabled actions $a \in Act(l)$ are greedy.

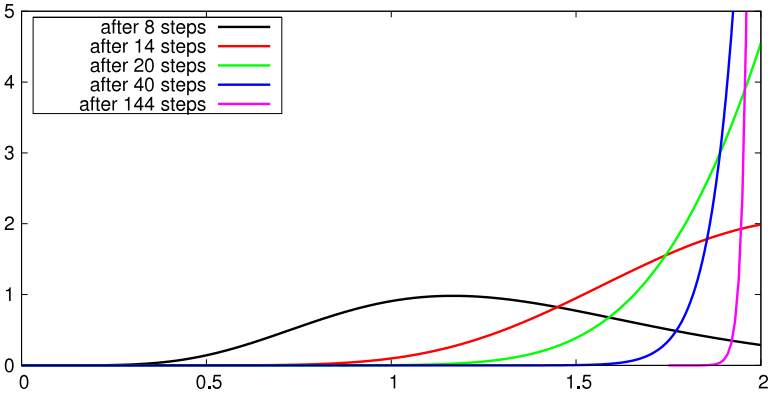

Fig. 1. The probability density functions of the Erlang distribution with rate $\lambda = 6$ and different shape parameters (i.e. number of steps), 'truncated' at time bound $T = 2$. That is, we consider the *conditional* Erlang distribution, assuming we have not exceeded the bound T (as the behaviour after T is irrelevant).

Lemma 3.1. *Greedy schedulers exist, and they can be described as the class of schedulers that choose a greedy action upon every reachable time-abstract path.*

Proof. It is plain that, for every non-goal location $l \notin G$ it holds $shift(d_l) \geq \sum_{l' \in L} \mathbf{P}(l, a, l') d_{l'}$ for every action a, and that equality must hold for some. That is, greedy actions exist. Using the definition of greedy actions, we can show by induction over all $i \in \mathbb{N}$ that for a scheduler $\mathcal{S}$ that always chooses greedy actions it holds that $d_l[i] = d_{l,\mathcal{S}}[i]$. This (constructively) proves the existence of greedy schedulers.

Considering the proof by induction above, we can immediately see that other schedulers that choose a non-greedy action on one or more reachable time-abstract paths (that do not contain goal locations) yield smaller step probability vectors. Thus, the schedulers that choose greedy actions characterise the class of greedy schedulers. □

This allows in particular to fix a positional *standard greedy scheduler* by fixing an arbitrary greedy action for every location.

To determine the set of greedy actions, let us consider a deterministic scheduler $\mathcal{S}$ that starts in a location l with a non-greedy action a. Then $shift(d_{l,\mathcal{S}}) \leq \sum_{l' \in L} \mathbf{P}(l, a, l') d_{l'}$ holds, where the sum $\sum_{l' \in L} \mathbf{P}(l, a, l') d_{l'}$ corresponds to the scheduler choosing the non-greedy action a at location l and acting greedy in all further steps. Let $d_{l,a}$ be the step probability vector of such schedulers. It naturally satisfies $shift(d_{l,a}) = \sum_{l' \in L} \mathbf{P}(l, a, l') d_{l'}$.

We know that $d_{l,\mathcal{S}} \leq d_{l,a} < d_l$. Hence, there is not only a difference between $d_{l,\mathcal{S}}$ and d_l, this difference will not occur at a higher index than the first difference between the newly defined $d_{l,a}$ and d_l. The finite number of locations and actions thus implies the existence of a bound k on the occurrence of this first difference between $d_{l,a}$ and d_l as well as $d_{l,\mathcal{S}}$ and d_l. While the existence of such a k suffices to demonstrate the existence of optimal schedulers, we show in Section 3.4 that this constant $k < |L|$ is smaller than the CTMDP itself.

Having established such a bound k, it suffices to compare schedulers up to this bound. This provides us with the greedy actions, and also with the initial sequence $d_{l,a}[0], d_{l,a}[1], \ldots, d_{l,a}[k]$ for all locations l and actions a. Consequently, we can determine a positive lower bound $\mu > 0$ for the first non-zero entry of the vectors $d_l - d_{l,\mathcal{S}}$ (considering all non-greedy schedulers $\mathcal{S}$). We call this lower bound μ the *discriminator* of the CTMDP. Intuitively, the discriminator μ represents the minimal advantage of the greedy strategy over non-greedy strategies.

3.2. Uniform CTMDPs

In this subsection, we show that every CD or HD scheduler for a uniform CTMDP can be transformed into a scheduler that converges to the standard greedy scheduler while preserving or improving the reachability probability.

In the quest for an optimal scheduler, it is useful to consider the fact that the maximal reachability probability can be computed using the step probability vector, because the likelihood that a particular number of steps happen in time t is independent of the scheduler:

$$Pr^G_{\mathcal{S}}(t) = \sum_{l \in L} \nu(l) \sum_{i=0}^{\infty} d_{l,\mathcal{S}}[i] \cdot p_{\lambda t}(i). \tag{1}$$

Moreover, the Poisson distribution $p_{\lambda t}$ has the useful property that the probability of taking k steps is falling very fast. We define the *greed bound* $n_{\mathcal{M}}$ to be a natural number, for which

$$\mu\, p_{\lambda t}(n) \geq \sum_{i=1}^{\infty} p_{\lambda t}(n+i) \quad \forall n \geq n_{\mathcal{M}} \tag{2}$$

holds. It suffices to choose $n_{\mathcal{M}} \geq \frac{2\lambda t}{\mu}$ since it implies $\mu p_{\lambda t}(n) \geq 2p_{\lambda t}(n+1)$, $\forall n > n_{\mathcal{M}}$ (which yields (2) by simple induction). Such a greed bound implies that the decrease in likelihood of reaching the goal region in few steps caused by

making a non-greedy decision after the greed bound dwarfs any potential later gain. We use this observation to improve any given CD or HD scheduler $\mathfrak{S}$ that makes a non-greedy decision after $\geq n_{\mathcal{M}}$ steps by replacing the behaviour after this history by a greedy scheduler. Finally, we use the interchangeability of greedy schedulers to introduce a scheduler $\overline{\mathfrak{S}}$ that makes the same decisions as $\mathfrak{S}$ on short histories and follows the standard greedy scheduling policy once the length of the history reaches the greed bound. For this scheduler, we show that $Pr^G_{\overline{\mathfrak{S}}}(t) \geq Pr^G_{\mathfrak{S}}(t)$ holds.

Theorem 3.2. *For uniform CTMDPs, there is an optimal scheduler for the classes CD and HD that converges to the standard greedy scheduler after $n_{\mathcal{M}}$ steps.*

Proof. Let us consider any HD scheduler $\mathfrak{S}$ that makes a non-greedy decision after a time-abstract path π of length $|\pi| \geq n_{\mathcal{M}}$ with last location l. If the path ends in, or has previously passed, the goal region, or if the probability of the history π is 0, that is, if it cannot occur with the scheduling policy of $\mathfrak{S}$, then we can change the decision of $\mathfrak{S}$ on every path starting with π arbitrarily – and in particular to the standard greedy scheduler – without altering the reachability probability.

If $PR_{\mathfrak{S}}(\pi, t) > 0$, then we change the decisions of the scheduler $\mathfrak{S}$ for paths with prefix π such that they comply with the standard greedy scheduler. We call the resulting HD scheduler $\mathfrak{S}'$ and analyse the change in reachability probability using Eq. (1):

$$Pr^G_{\mathfrak{S}'}(t) - Pr^G_{\mathfrak{S}}(t) = PR_{\mathfrak{S}}(\pi, t) \cdot \sum_{i=0}^{\infty} (d_l[i] - d_{l,\mathfrak{S}_\pi}[i]) \cdot p_{\lambda t}(|\pi| + i),$$

where $\mathfrak{S}_\pi : \pi' \mapsto \mathfrak{S}(\pi \circ \pi')$ is the HD scheduler which prefixes its input with the path π and then calls the scheduler $\mathfrak{S}$. The greedy criterion implies $d_l > d_{l,\mathfrak{S}_\pi}$ with respect to the lexicographic order, and after rewriting the upper equation (for some $j > 0$):

$$Pr^G_{\mathfrak{S}'}(t) - Pr^G_{\mathfrak{S}}(t) = PR_{\mathfrak{S}}(\pi, t) \cdot \left(\mu p_{\lambda t}(|\pi| + j) + \sum_{i>j}^{\infty} (d_l[i] - d_{l,\mathfrak{S}_\pi}[i]) \cdot p_{\lambda t}(|\pi| + i) \right)$$

we can apply Eq. (2) to deduce that the difference $Pr^G_{\mathfrak{S}'}(t) - Pr^G_{\mathfrak{S}}(t)$ is non-negative.

Likewise, we can concurrently change the scheduling policy to the standard greedy scheduler for all paths of length $\geq n_{\mathcal{M}}$ for which the scheduler $\mathfrak{S}$ makes non-greedy decisions. In this way, we obtain a scheduler $\mathfrak{S}''$ that makes non-greedy decisions only in the first $n_{\mathcal{M}}$ steps, and yields a (not necessarily strictly) better time-bounded reachability probability than $\mathfrak{S}$.

Since all greedy schedulers are interchangeable without changing the time-bounded reachability probability (and even without altering the step probability vector), we can modify $\mathfrak{S}''$ such that it follows the standard greedy scheduling policy after $\geq n_{\mathcal{M}}$ steps, resulting in a scheduler $\overline{\mathfrak{S}}$ that comes with the same time-bounded reachability probability as $\mathfrak{S}''$. Note that $\overline{\mathfrak{S}}$ is counting if $\mathfrak{S}$ is counting.

Hence, the supremum over the time-bounded reachability of all CD/HD schedulers is equivalent to the supremum over the bounded reachability of CD/HD schedulers that deviate from the standard greedy scheduler only in the first $n_{\mathcal{M}}$ steps. This class is finite, and the supremum over the time-bounded reachability probability is, therefore, the maximal time-bounded reachability probability obtained by one of its representatives. □

Hence, we have shown the existence of a – simple – optimal time-bounded CD scheduler. Using the fact that the suprema over the time-bounded reachability probability coincide for CD, CR, HD, and HR schedulers [11], we can infer that such a scheduler is optimal for all of these classes.

Corollary 3.3. $\max_{\mathfrak{S} \in CD} Pr^G_{\mathfrak{S}}(t) = \max_{\mathfrak{S} \in HR} Pr^G_{\mathfrak{S}}(t)$ *holds for all uniform CTMDPs $\mathcal{M}$.* □

3.3. Non-uniform CTMDPs

Reasoning over non-uniform CTMDPs is harder than reasoning over uniform CTMDPs, because the likelihood of seeing exactly k steps does not adhere to the simple Poisson distribution, but depends on the precise history. Even if two paths have the same length, they may imply different probability distributions over the time passed so far. Knowing the time-abstract history therefore provides a scheduler with more information about the system's state than merely its length. As a result, it is simple to construct example CTMDPs, for which history-dependent and counting schedulers can obtain different time-bounded reachability probabilities [11].

Now, we extend the results from the previous subsection to general CTMDPs. We show that simple optimal CD/HD schedulers exist, and that randomisation does not yield an advantage:

$$\max_{\mathfrak{S} \in CD} Pr^G_{\mathfrak{S}}(t) = \max_{\mathfrak{S} \in CR} Pr^G_{\mathfrak{S}}(t) \quad \text{and} \quad \max_{\mathfrak{S} \in HD} Pr^G_{\mathfrak{S}}(t) = \max_{\mathfrak{S} \in HR} Pr^G_{\mathfrak{S}}(t).$$

To obtain this result, we work on the uniformisation $\mathcal{U}$ of $\mathcal{M}$ instead of working on $\mathcal{M}$ itself. We argue that the behaviour of a general CTMDP $\mathcal{M}$ can be viewed as the observable behaviour of its uniformisation $\mathcal{U}$, using a scheduler that does not *see* the new transitions and locations. Schedulers from this class can then be replaced by (or viewed as) schedulers that do not *use* the additional information. And finally, we can approximate schedulers that do not use the additional information

by schedulers that do not use it initially, where initially means until the number of visible steps–and hence in particular the number of steps–exceeds the greed bound $n_{\mathcal{U}}$ of the uniformisation $\mathcal{U}$ of $\mathcal{M}$. Comparable to the argument from the proof of Theorem 3.2, we show that we can restrict our attention to the standard greedy scheduler after this initial phase, which leads again to a situation where considering a finite class of schedulers suffices to obtain the optimum.

Lemma 3.4. *The greedy decisions and the step probability vector coincide for the observable and unobservable copy of each location in the uniformisation $\mathcal{U}$ of any CTMDP $\mathcal{M}$.*

Proof. The observable and unobservable copy of each location reach the same successors under the same actions with the same transition rate. □

We can therefore choose a positional *standard greedy scheduler* whose decisions coincide for the observable and unobservable copy of each location. For the *uniformisation* $\mathcal{U}$ of a CTMDP $\mathcal{M}$, we define the function $vis : Paths_{abs}(\mathcal{U}) \to Paths_{abs}(\mathcal{M})$ that maps a path π of $\mathcal{U}$ to the corresponding path in $\mathcal{M}$, the *visible path*, by deleting all unobservable locations and their directly preceding transitions from π. (Note that all paths in $\mathcal{U}$ start in an observable location.) We call a scheduler *n-visible* if its decisions only depend on the visible path and coincide for the observable and unobservable copy of every location for all paths containing up to n visible steps. We call a scheduler *visible* if it is n-visible for all $n \in \mathbb{N}$.

We call an HD/HR scheduler an (n-)visible HD/HR scheduler if it is (n-)visible. An (n-)visible HD/HR scheduler will be called a visible CD/CR scheduler if its decisions depend only on the length of the visible path and the last location, and it will be called an n-visible CD/CR scheduler if its decisions depend only on the length of the visible path (and the last location) for all paths containing up to n visible steps. The respective classes are denoted with prefixes accordingly, for example, n-vCD. Note that (n-)visible counting schedulers are not necessarily counting schedulers.

It is a simple observation that we can study visible CD, CR, HD, and HR schedulers on the uniformisation $\mathcal{U}$ of a CTMDP $\mathcal{M}$ instead of studying CD, CR, HD, and HR schedulers on $\mathcal{M}$.

Lemma 3.5. *$\mathcal{S} \mapsto \mathcal{S} \circ vis$ is a bijection from CD, CR, HD, or HR schedulers of a CTMDP $\mathcal{M}$ onto* visible *CD, CR, HD, or HR, respectively, schedulers for the uniformisation $\mathcal{U}$ of $\mathcal{M}$ that preserves the time-bounded reachability probability: $Pr^G_{\mathcal{S}}(t) = Pr^G_{\mathcal{S}\circ vis}(t)$.* □

At the same time, copying the argument from the proof of Theorem 3.2, an $n_{\mathcal{U}}$-visible CD or HD scheduler $\mathcal{S}$ can be adjusted to the $n_{\mathcal{U}}$-visible CD or HD scheduler $\overline{\mathcal{S}}$ that deviates from $\mathcal{S}$ only in that it complies with the standard greedy scheduler for $\mathcal{U}$ after $n_{\mathcal{U}}$ visible steps, without decreasing the time-bounded reachability probability. These schedulers are visible schedulers from a finite sub-class, and hence some representative of this class takes the optimal value. We can, therefore, construct optimal CD and HD schedulers for every CTMDP $\mathcal{M}$.

Lemma 3.6. *The following equations hold for the uniformisation $\mathcal{U}$ of a CTMDP $\mathcal{M}$:*

$$\max_{\mathcal{S}\in n_{\mathcal{U}}-vCD} Pr^G_{\mathcal{S}}(t) = \max_{\mathcal{S}\in vCD} Pr^G_{\mathcal{S}}(t) \quad \textit{and} \quad \max_{\mathcal{S}\in n_{\mathcal{U}}-vHD} Pr^G_{\mathcal{S}}(t) = \max_{\mathcal{S}\in vHD} Pr^G_{\mathcal{S}}(t).$$

Proof. We have shown in Theorem 3.2 that turning to the standard greedy scheduling policy after $n_{\mathcal{U}}$ or more steps can only increase the time-bounded reachability probability. This implies that we can turn to the standard greedy scheduler after $n_{\mathcal{U}}$ *visible* steps.

The scheduler resulting from this adjustment does not only remain $n_{\mathcal{U}}$-visible, it becomes a visible CD and HD scheduler, respectively. Moreover, it is a scheduler from the finite subset of CD or HD schedulers, respectively, whose behaviour may only deviate from the standard scheduler within the first $n_{\mathcal{U}}$ visible steps. □

To prove that optimal CD and HD schedulers are also optimal CR and HR schedulers, respectively, we first prove the simpler lemma that this holds for k-bounded reachability.

Definition 3.7. We define *k-bounded reachability* to be the probability to reach the goal region in k or less steps. Accordingly, *k-optimal schedulers* optimise k-bounded reachability.

Lemma 3.8. *k-optimal CD or HD schedulers are also k-optimal CR or HR schedulers, respectively.*

Proof. For a CTMDP $\mathcal{M}$ we can turn an arbitrary CR or HR scheduler $\mathcal{S}$ into a CD or HD scheduler $\mathcal{S}'$ with a time and k-bounded reachability probability that is at least as good as the one of $\mathcal{S}$ by first determinising the scheduler decisions from the $(k+1)$st step onwards – this has obviously no impact on k-bounded reachability – and then determinising the remaining randomised choices.

Replacing a single randomised decision on a path π (for history-dependent schedulers) or on a set of paths Π (for counting schedulers) that end(s) in a location l is safe, because the time and k-bounded reachability probability of a scheduler is an affine combination – the affine combination defined by $\mathcal{S}(\pi)$ and $\mathcal{S}(|\pi|, l)$, respectively – of the $|Act(l)|$ schedulers resulting from determinising this single decision. Hence, we can pick one of them whose time and k-bounded reachability probability is at least as high as the one of $\mathcal{S}$.

As the number of these randomised decisions is finite ($\leq k\,|L|$ for CR, and $\leq k^{|L|}$ for HR schedulers), this results in a deterministic scheduler after a finite number of improvement steps. □

Theorem 3.9. *Optimal CD schedulers are also optimal CR schedulers.*

Proof. First, for $n \to \infty$ the probability to reach the goal region G in exactly n or more than n steps converges to 0, independent of the scheduler. Together with Lemma 3.8, this implies

$$\sup_{\mathfrak{S}\in CR} Pr^G_{\mathfrak{S}}(t) = \lim_{n\to\infty} \sup_{\mathfrak{S}\in CR} Pr^G_{\mathfrak{S}}(t;n) = \lim_{n\to\infty} \sup_{\mathfrak{S}\in CD} Pr^G_{\mathfrak{S}}(t;n) \leq \max_{\mathfrak{S}\in CD} Pr^G_{\mathfrak{S}}(t),$$

where equality is implied by $CD \subseteq CR$. □

Analogously, we can prove the similar theorem for history-dependent schedulers:

Theorem 3.10. *Optimal HD schedulers are also optimal HR schedulers.* □

3.4. Constructing optimal schedulers

The proof of the existence of an optimal scheduler is not constructive in two aspects. First, the computation of a positional greedy scheduler requires a bound for k, which indicated the maximal depth up to which we have to compare the step probability vectors before we can ascertain equality. Second, we need an exact method to compare the quality of two (arbitrary) schedulers.

A bound for k. The first property is captured in the following lemma. Without this lemma, we could only provide an algorithm that is guaranteed to converge to an optimal scheduler, but would be unable to determine whether an optimal solution has already been reached, because we would not know when to stop when comparing step probability vectors. In this lemma, we show that it suffices to check for equivalence of two step probability vectors up to position $|L| - 2$. As discussed in Section 3.1, this enables us to identify greedy actions and thus to *compute* the discriminator μ and consequently the greed bound $n_{\mathcal{M}}$.

Lemma 3.11. *For every location l of a uniform CTMDP $\mathcal{M}$, the position of the first difference between d_l and any $d_{l,a}$ is bounded by $|L| - 2$. That is, $|L| - 2$ is an upper bound for the smallest k that satisfies*

$$\forall l \in L,\ a \in Act(l) :\ d_l \neq d_{l,a} \Rightarrow \exists k' \leq k :\ d_l[k] > d_{l,a}[k].$$

Proof. The techniques we exploit in this proof draw from linear algebra, and are, while simple, a bit unusual in this context. We first turn to the simpler notion of Markov chains by resolving the nondeterminism in accordance with the positional standard greedy scheduler $\mathfrak{S}$ whose existence was shown in Section 3.1.

We first lift the step probability vector from locations to linear combinations over locations (for example distributions), where $d_\nu = \sum_{l\in L} \nu(l) d_l$ is, for a function $\nu : L \to \mathbb{R}$, the linear combination of the step probability vectors of the individual locations.

In this proof, we define two functions $\nu, \nu' : L \to \mathbb{R}$ to be equivalent if their step probability vectors $d_\nu = d_{\nu'}$ are equal. Further, we call them i-step equivalent, denoted $\nu \sim_i \nu'$, if their step probability vectors are equal up to position i (that is, $\forall 0 \leq j \leq i.\ d_\nu[j] = d_{\nu'}[j]$).

Let, for all $i \in \mathbb{N}$,

$$S_i = \{\nu - \nu' \mid \nu, \nu' \text{ are distributions with } \nu \sim_i \nu'\}$$

be the spanning set obtained from the equivalence classes of i-step equivalent distributions, and let D_i be the vector space it spans. (Addition and scalar multiplication are defined location-wise.)

As the elements of the D_i are vectors with $|L|$ components, the dimension of each D_i is at most $|L|$. Naturally, $D_i \supseteq D_{i+1}$ always holds, as $i + 1$ step equivalence implies i-step equivalence, which in turn implies $S_i \supseteq S_{i+1}$.

Next, we show that D_0 has $|L| - 2$ dimensions, and that $D_i = D_{i+i}$ implies that a fixed point is reached, which together implies that $D_{|L|-2} = D_j$ holds for all $j \geq |L| - 2$.

- D_0 has $|L| - 2$ dimensions: The elements δ of D_0 can also be considered as the multitudes of differences $\delta = \lambda(\nu - \nu')$ of *distributions* $\nu, \nu' : L \to [0, 1]$ that are equally likely in the goal region (due to 0-step equivalence; $d_\nu[0] = d_{\nu'}[0]$).

  The fact that ν and ν' are distributions implies $\sum_{l\in L} \nu(l) = 1$ and $\sum_{l\in L} \nu'(l) = 1$, and hence $\sum_{l\in L} \delta(l) = 0$. Further, the fact that ν and ν' are equally likely in the goal region implies $\sum_{l\in G} \nu(l) = \sum_{l\in G} \nu'(l)$, and hence $\sum_{l\in G} \delta(l) = 0$. Thus, D_0 has $|L| - 2$ dimensions. (Assuming $G \neq L$ and $G \neq \emptyset$, but otherwise every scheduler has equal quality.)
- Once we have constructed D_i, we can construct the vector space O_i that contains a vector δ if it is a multitude $\delta = \lambda(\nu - \nu')$ of differences $\nu - \nu'$ of distributions, such that $shift(d_\nu)$ and $shift(d_{\nu'})$ are i-step equivalent, that is, $shift(d_\nu) - shift(d_{\nu'}) \in D_i$.

  The transition from step probability vectors to the *shift* of them is a simple linear operation, which transforms the distributions according to the transition matrix of the embedded DTMC. Hence, we can obtain O_i from D_i by a linear transformation of the vector space.
- Two-step probability vectors are $i + 1$-step equivalent if (1) they are i-step equivalent, and (2) their shifts are i-step equivalent. Therefore $D_{i+1} = D_i \cap O_i$ can be obtained by an intersection of the two vector spaces D_i and O_i.

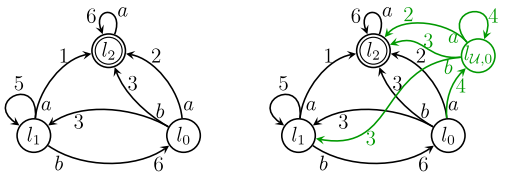

Fig. 2. The example CTMDP $\mathcal{M}$ (left) and the reachable part of its uniformisation $\mathcal{U}$ (right).

This reaffirms that the vector spaces are shrinking, that is, $D_0 \supseteq D_1 \supseteq \cdots \supseteq D_{|L|-2} \supseteq \cdots$. Moreover, $D_i = D_{i+1}$ implies $D_i = D_i \cap O_i$ and hence $D_i \subseteq O_i$, as well as $O_{i+1} = O_i$, and consequently $D_{i+2} = D_i \cap O_i = D_i$.

As D_0 is an $|L| - 2$ dimensional vector space, and inequality ($D_i \neq D_{i+1}$) implies the loss of at least one dimension, a fixed point is reached after at most $|L| - 2$ steps. Two distributions are therefore i-step equivalent for all $i \in \mathbb{N}$ if, and only if, they are $(|L| - 2)$-step equivalent. As two distributions are equivalent if, and only if, they are i-step equivalent for all $i \in \mathbb{N}$, they are equivalent if, and only if, they are $(|L| - 2)$-step equivalent.

Having established this, we apply it on the distribution $\nu_{l,a}$ obtained in one step from a position $l \notin G$ when choosing the action a, as compared to the distribution ν_l obtained when choosing the action according to the positional greedy scheduler.

Now, $d_l > d_{l,a}$ holds if, and only if, $\mathit{shift}(d_l) = d_{\nu_l} > d_{\nu_{l,a}} = \mathit{shift}(d_{l,a})$, which implies $d_{\nu_l}[k'] > d_{\nu_{l,a}}[k']$ for some $k' \leq |L| - 2$, and hence $d_l[k] > d_{l,a}[k]$ for some $k < |L|$. □

Comparing schedulers. So far, we have narrowed down the set of candidates for the optimal scheduler to a finite number of schedulers. To determine the optimal scheduler, it now suffices to have a comparison method for their reachability probabilities.

The combination of each of these schedulers with the respective CTMDP can be viewed as a *finite* continuous-time Markov *chain* (CTMC) since they behave like a positional scheduler after $n_{\mathcal{M}}$ steps. Aziz et al. [21] have shown that the time-bounded reachability probability of CTMCs are computable (and comparable) finite sums $\sum_{i \in I} \eta_i e^{\delta_i}$, where the individual η_i and δ_i are algebraic numbers.

We conclude with a constructive extension of our results:

Corollary 3.12. *We can effectively construct optimal CD, CR, HD, and HR schedulers.* □

Corollary 3.13. *We can compute the time-bounded reachability probability of optimal schedulers as finite sums $\sum_{i \in I} \eta_i e^{\delta_i}$, where the η_i and δ_i are algebraic numbers.* □

Complexity. These corollaries lean on the precise CTMC model checking approach of Aziz et al. [21], which only demonstrates the effective decidability of this problem. We deem it unlikely that a complexity for finding optimal strategies can be provided prior to determining the respective CTMC model checking complexity.

3.5. Example

To exemplify our proposed construction, let us consider the example CTMDP $\mathcal{M}$ depicted in Fig. 2. As $\mathcal{M}$ is not uniform, we start with constructing the uniformisation $\mathcal{U}$ of $\mathcal{M}$ (cf. Fig. 2).

$\mathcal{U}$ has the uniform transition rate $\lambda = 6$. Independent of the initial distribution of $\mathcal{M}$, the unobservable copies of l_1 and l_2 are not reachable in $\mathcal{U}$, because the initial distribution of a uniformisation assigns all probability weight to observable locations, and the transition rate of all enabled actions in l_1 and l_2 in $\mathcal{M}$ is already λ. (Unobservable copies of a location l are only reachable from the observable and unobservable copy of l upon enabled actions a with non-maximal exit rate $\mathbf{R}(l, a, L) \neq \lambda$.)

Disregarding the unreachable part of $\mathcal{U}$, there are only 8 positional schedulers for $\mathcal{U}$, and only 4 of them are visible (that is, coincide on l_0 and $l_{\mathcal{U},0}$). They can be characterised by $\mathcal{S}_1 = \{l_0 \mapsto a,\ l_1 \mapsto a\}$, $\mathcal{S}_2 = \{l_0 \mapsto a,\ l_1 \mapsto b\}$, $\mathcal{S}_3 = \{l_0 \mapsto b,\ l_1 \mapsto a\}$, and $\mathcal{S}_4 = \{l_0 \mapsto b,\ l_1 \mapsto b\}$. In order to determine a greedy scheduler, we first determine step probability vectors:

For l_0: $d_{l_0,\mathcal{S}_1} = d_{l_0,\mathcal{S}_2} = (\frac{1}{3}, \frac{5}{9}, \frac{19}{27}, \ldots)$, $d_{l_0,\mathcal{S}_3} = (\frac{1}{2}, \frac{7}{12}, \frac{43}{72}, \ldots)$, $d_{l_0,\mathcal{S}_4} = (\frac{1}{2}, \frac{1}{2}, \frac{3}{4}, \ldots)$.

For l_1: $d_{l_1,\mathcal{S}_1} = d_{l_1,\mathcal{S}_3} = (\frac{1}{6}, \frac{7}{36}, \frac{71}{216}, \ldots)$, $d_{l_1,\mathcal{S}_2} = (0, \frac{1}{3}, \frac{5}{9}, \ldots)$, $d_{l_1,\mathcal{S}_4} = (0, \frac{1}{2}, \frac{1}{2}, \ldots)$.

Note that, in the given example, it suffices to compute the step probability vector for a single step to determine that $\mathcal{S}_3$ is optimal (w.r.t. the greedy optimality criterion); in general, it suffices to consider as many steps as the CTMDP has locations. Since deviating from S_3 decreases the chance to reach the goal location l_2 in a single step by $\frac{1}{6}$ both from l_0 and l_1, the discriminator $\mu = \frac{1}{6}$ is easy to compute.

Our coarse estimation provides a greed bound of $n_{\mathcal{U}} = \lceil 72 \cdot T \rceil$, where T is the time bound, but $n_{\mathcal{U}} = \lceil 42 \cdot T \rceil$ suffices to satisfy Eq. (2). Fig. 1 depicts the probability distribution over time the scheduler may assume for the case that $T = 2$ and different step counts; 144 steps correspond to $n_{\mathcal{U}}$.

When seeking optimal schedulers from any of the discussed classes, we can focus on the finite set of those schedulers that comply with $\mathcal{S}_3$ after $n_{\mathcal{U}}$ (visible) steps. In the previous subsection, we described how the precise model checking technique of Aziz et al. [21] can be exploited to turn the existence proof into an effective technique for the construction of optimal schedulers.

4. Asymmetric games

In this section, we discuss CTGs where one player is restricted to using a time-abstract scheduling policy, while his opponent can use time-dependent scheduling policies. This is a natural assumption, as one of the two players often refers to our means to control the behaviour of these systems, which may suffer from restrictions like not being able to observe time, while the other player refers to an antagonist or an abstraction. The latter is therefore usually assumed to be unrestricted or very powerful. In *asymmetric CTGs*, we thus assume that this player may draw from the full power of time-dependent schedulers and may choose its strategy depending on the choice of the time-dependent player. The problems discussed in this section, therefore, target the following objectives

$$\sup_{\mathcal{S}_A \in X} \inf_{\mathcal{S}_D \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{G}}(G, T)), \tag{3}$$

and

$$\inf_{\mathcal{S}_D \in X} \sup_{\mathcal{S}_A \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{G}}(G, T)), \tag{4}$$

where $X \in \{CD, CR, HR, HD\}$ is a time-abstract scheduler class.

The first question that arises is whether or not optimal schedulers for the time-abstract player do exist. We give a positive answer to this question for all time-abstract scheduler classes in Section 4.1. The structure of these schedulers, however, is not as simple as for the CTMDPs discussed in the previous section (or for the symmetric case, cf. Section 5): schedulers that are positional in the limit are generally insufficient for all of the time-abstract scheduler classes.

The second question that arises is whether or not counting strategies provide, at least for uniform CTGs, results equivalent to strategies that may use history, and whether or not deterministic strategies are as good as randomised ones. Different to the results for CTMDPs (or for the symmetric case discussed in Section 5), the answer to both questions is negative: in Section 4.2, we provide examples, where even positional randomised strategies improve over history dependent deterministic ones, as well as cases, where history dependent deterministic strategies improve over counting randomised ones.

4.1. Existence and structure of optimal strategies

We start this section with an existence proof for optimal time-abstract schedulers: for all time-abstract scheduler classes, there is a scheduler, for which the value of Eq. (3) is taken. In Theorem 4.1 we provide a non-constructive existence proof for a scheduler $\mathcal{S}_A$ that satisfies $\inf_{\mathcal{S}_D \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{G}}(G, T)) = \sup_{\mathcal{S}_A \in X} \inf_{\mathcal{S}_D \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{G}}(G, T))$. Likewise, we demonstrate the existence of a scheduler, for which the value of Eq. (4) is taken.

Theorem 4.1. *Given a Markov game, a goal region, and a time bound, there is a scheduler $\mathcal{S}$ in every class $X \in \{CD, HD, CR, HR\}$, such that the supremum / infimum from Eqs. (3) and (4) is taken.*

Proof. We can inductively construct an optimal strategy for the time-abstract player by successively fixing the decisions the player makes on all paths of length i, starting with $i = 0$. For the classes CD and HD of deterministic counting and history-dependent schedulers and an initial strategy that fixes all decisions up to some step $i - 1$ (where -1 is the base case where nothing is fixed), there are only finitely many continuations for step i. We partition the set $\mathcal{S}_{i-1}$ of strategies that comply with the first $i - 1$ fixed decisions in a finite number of sets $\mathcal{S}^j_{i-1}$ of strategies such that each of them complies with one of the continuations in step i. The supremum (resp. infimum) over the finite union of sets is the maximum (resp. minimum) over the suprema (resp. infima) of the sets:

$$\sup_{\mathcal{S}_A \in \bigcup_j \mathcal{S}^j_{i-1}} f(\mathcal{S}_A) = \max_j \sup_{\mathcal{S}_A \in \mathcal{S}^j_{i-1}} f(\mathcal{S}_A)$$

holds when j ranges over a finite domain like the decisions in step i, where $f(\mathcal{S}) = \inf_{\mathcal{S}_D \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}\big(reach_{\mathcal{M}}(G, T)\big)$. Thus, we can select a maximising parameter j and fix it as an optimal decision for step i.

For the randomised scheduler classes CR and CD and an initial strategy that fixes all decisions up to some step $i - 1$, where -1 serves again as the base case where nothing is fixed, the set of continuations for step i is compact, and the value of the sup inf (resp. inf sup) term is continuous in the choice made at step i. Hence the optimum is taken on this compact set.

The observation that the likelihood of seeing more than i discrete events converges to 0 when i grows towards infinity closes the argument that a strategy constructed this way will indeed take the value of the sup inf term of Eq. (3) or the value of the inf sup term of Eq. (4), respectively. □

The proof of Theorem 4.1 is remarkably simple. Indeed, it is far simpler than the proof of the existence of optimal strategies for CTMDPs provided in the previous section. The reason for this is that the claim is much weaker: The proof is neither constructive, nor does it provide an argument for a potential conversion against a limit stable strategy.

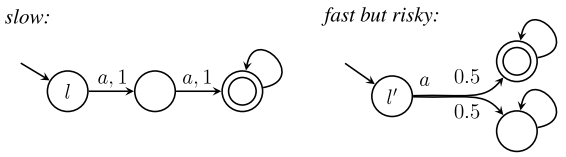

Fig. 3. Two simple uniform CTGs (CTMCs, in fact) with transition rate $\lambda = 1$. On the left, the time-bounded reachability probability for l is equal to the Erlang(2, 1)-distribution (*cumulative* density function). On the right, the time-bounded reachability probability for l' is one half of the exponential distribution. Both functions intersect when $t \approx 1.26$ time units remain.

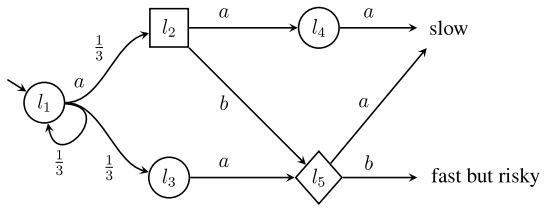

Fig. 4. A Markov Game $\mathcal{G}$ that serves as an example that there is no limit-optimal HD scheduler for asymmetric games. The box-shaped location is controlled by the time-dependent player, while the diamond-shaped location is controlled by the time-abstract player. The labels *slow* and *fast but risky* refer to the CTMCs in Fig. 3.

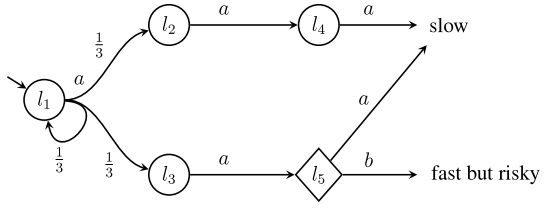

Fig. 5. The CTMDP $\mathcal{M}$ that results when fixing the action a in location l_2.

We now show that such an extension cannot exist: there are Markov games, for which no strategy that converges to a positional strategy after finitely many steps is optimal. This does not imply the impossibility of optimal strategies, but it shows that the proof technique we used for the single player case cannot succeed and that the structure of optimal strategies may be more complex in asymmetric games.

Theorem 4.2. *There is a uniform Markov game, a goal region and a time bound such that there is no optimal limit-positional deterministic or randomised strategy for the time-abstract player that is optimal for the HD or HR strategies.*

Proof. We consider the case the time-abstract player aims to maximise the time-bounded reachability (Eq. (3)). We fix a sufficiently large time bound T (i.e. $T = 1000$) and a uniform Markov game $\mathcal{G}$, see Fig. 4, for which we demonstrate the impossibility of an optimal limit-positional strategy for the time-abstract player. The proof builds on the fact that in the given CTG, the time-abstract player must distinguish whether its location was entered via l_2 or via l_3 in order to attain the optimal reachability probability.

It is easy to see, that the time-abstract player can use his knowledge about the history to effectively take away the option of playing b from the choices of the time-dependent player: by choosing a whenever l_5 is entered from l_2, the game in Fig. 4 reduces to the CTMDP $\mathcal{M}$ in Fig. 5.

For the restricted class of time-abstract schedulers for the CTG $\mathcal{G}$ that shows this behaviour, the supremum of their time-bounded reachability probabilities is not worse than the reachability of the full class of time-abstract schedulers, since we could consider this operation simply as a restriction of the time-*dependent* player's available strategies.

As we eliminated all choices of the time-dependent player the game reduced to a CTMDP with only a time-abstract player. Using the results from the previous section, we can determine a time-abstract counting deterministic (CD) strategy for CTMDP $\mathcal{M}$ that converges to a positional strategy after finitely many steps (i.e. is limit-positional). It can be translated back to an HD strategy for the CTG $\mathcal{G}$ that is optimal also for the CTG, but is not limit-positional any more: it chooses a whenever l_5 was reached via l_2, and after a finite number of steps it chooses action b for all other paths. Let $\mathcal{S}^*$ denote this strategy for the CTG in Fig. 4.

It remains to show that any limit-positional strategy $\mathcal{S}_{lp}$ for the CTG in Fig. 4 would be inferior, but we can prove this easily: Let k be the step number after which $\mathcal{S}_{lp}$ becomes positional. Assigning a positive probability to action a in the positional part (after step k) would obviously decrease the reachability probability on any path that traverses l_3 compared to strategy $\mathcal{S}^*$. For paths that lead through location l_2 the time-dependent player always has the choice to ignore his option for action b, which would not affect the probability compared to strategy $\mathcal{S}^*$. Also for the remaining paths that visit l_3 and reach l_5 before the step bound k, the $\mathcal{S}_{lp}$ cannot achieve a higher probability than $\mathcal{S}^*$. Thus, a limit-positional strategy cannot assign action a a positive probability value.

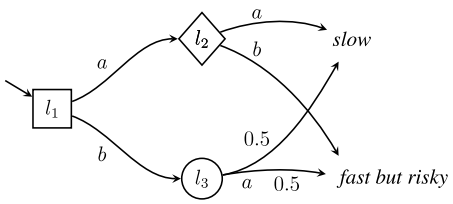

Fig. 6. The CTG of Theorem 4.4. To ease understanding, we split the game into three parts. Whenever we reach the labels *slow* or *fast but risky* we enter the initial state of the respective automata of Fig. 3. The box shaped location is controlled by the time-dependent player while the diamond shaped location is controlled by the time-abstract player. At round locations no decision can be made.

Choosing b positionally after step bound k, also leads to a decreased reachability probability. For any k there is a positive (possibly tiny) probability that location l_2 is reached *after* k steps but with $T' \geq 100$ time units left. For these cases it is beneficial for the time-dependent player to choose action b, as it leads with probability ≈ 0.5 to the goal region, while action a shows a reachability of almost 1. For all other paths that visit location l_5, the reachability probability cannot improve over $\mathcal{S}^*$. Thus, there is no limit-positional strategy that is optimal for HD or HR.

For the case that the time-abstract player has the objective to minimise the time-bounded reachability, we simply swap the 'slow' and 'fast but risky' parts in Figs. 4 and 5 and argue accordingly. □

4.2. Randomisation and history can help

In the previous subsection, we have shown that, different to the case of CTMDPs discussed in Section 3, schedulers that are positional in the limit are insufficient for maximising (or minimising) the time-bounded reachability probability. In this subsection, we show that both, randomisation and history, are already needed in simple games for the construction of optimal strategies.

Revisiting the proof of the previous Theorem, we have already provided an example of a game where HD schedulers are superior to CR schedulers.

Theorem 4.3. *History-dependent schedulers (HD) may have an advantage over counting schedulers (CR), even for uniform CTGs. That is, there is a uniform Markov game $\mathcal{G}$, a goal region G, and a time bound T such that the following inequalities hold:*

$$\sup_{\mathcal{S}_A \in HD} \inf_{\mathcal{S}_D \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{M}}(G,T)) > \sup_{\mathcal{S}_A \in CR} \inf_{\mathcal{S}_D \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{M}}(G,T)),$$

$$\inf_{\mathcal{S}_D \in HD} \sup_{\mathcal{S}_A \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{M}}(G,T)) < \inf_{\mathcal{S}_D \in CR} \sup_{\mathcal{S}_A \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{M}}(G,T)).$$

We now show that, vice versa, CR schedulers may be superior to HD schedulers. What is more, we show that even *positional* randomised (PR) schedulers may be superior to HD schedulers, even for uniform Markov games.

Theorem 4.4. *Randomised schedulers may have an advantage over non-randomised schedulers in uniform Markov games: there is a uniform Markov game $\mathcal{G}$, a goal region G, and a time bound T such that*

$$\sup_{\mathcal{S}_A \in PR} \inf_{\mathcal{S}_D \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{G}}(G,T)) > \sup_{\mathcal{S}_A \in HD} \inf_{\mathcal{S}_D \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{G}}(G,T)),$$

and

$$\inf_{\mathcal{S}_D \in PR} \sup_{\mathcal{S}_A \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{G}}(G,T)) < \inf_{\mathcal{S}_D \in HD} \sup_{\mathcal{S}_A \in THR} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(reach_{\mathcal{G}}(G,T)).$$

Proof. We begin with constructing a Markov game to prove the first inequality. Consider the Markov game $\mathcal{G} = (L, L_A, L_D, Act, \mathbf{R}, \nu)$ and goal region of Fig. 6, where $l_1 \in L_D$, $l_2 \in L_A$, and $\nu(l_1) = 1$. Let $\mathcal{G}$ have a uniform rate $\lambda = 1$ and let the time bound be fixed as $T = 1000$.

There is only one possible time-abstract path to from the initial location l_1 to l_2, such that there are effectively only two history-dependent *deterministic* time-abstract schedulers: the positional schedulers that choose on this history a and b, respectively. We call them $\mathcal{S}_a$ and $\mathcal{S}_b$, respectively.

Similar to the situation in the proof of Theorem 4.2, the optimal positional randomised scheduler is the scheduler that selects a and b with likelihood 0.5: it effectively makes l_2 and l_3 equivalent; we call this scheduler $\mathcal{S}_{ab}$. However, we do not have to prove the optimality of this scheduler, it suffices to show that $\mathcal{S}_{ab}$ is superior to $\mathcal{S}_a$ and to $\mathcal{S}_b$.

If the remaining time is large, say ≥ 100, the time aspect of reaching the goal region becomes irrelevant compared to the fact which of the two sub-automata, *slow* or *fast but risky*, we reach. Assuming that the time-dependent player chooses action a for all times, strategy $\mathcal{S}_b$ leads to a reachability probability of ≈ 0.5, whereas $\mathcal{S}_{ab}$ yields a probability of ≈ 0.75 and we obviously chose the best possible timed counter strategy for $\mathcal{S}_{ab}$. Thus, we showed that $\mathcal{S}_{ab}$ is superior to $\mathcal{S}_b$.

For the comparison of $\mathcal{S}_a$ and $\mathcal{S}_{ab}$, we have to consider a more involved counter-strategy $\mathcal{S}_t$ of the time-dependent player. Let $\mathcal{S}_t$ choose action a if the remaining time is less than 0.001, and action b if more time remains.

For the fixed counter-strategy $\mathcal{S}_t$, the evolution of the system for the time interval [0, 999.999] is independent of the strategy of the time-abstract player. Using independence of probabilistic events in our system we can consider the case that no jump happened in that interval in separation:

Location l_1 yields a higher probability for strategy $\mathcal{S}_{ab}$ than for strategy $\mathcal{S}_a$ for remaining time 0.001, since the probability that sufficiently many steps occur dominates the constant factor of 0.75 we get from the probabilistic choices of successor locations. In our case, the probability that exactly 3 steps occur in 0.001 time units is $\approx 1.67 \cdot 10^{-10}$, whereas the probability that 4 or more steps happen is $\approx 4.17 \cdot 10^{-14}$. Thus, the total difference in probability between $\mathcal{S}_a$ and $\mathcal{S}_{ab}$ is small, but positive.

This demonstrates that for $\mathcal{S}_a$ and $\mathcal{S}_b$ the infimum over all time-dependent strategies is lower than for $\mathcal{S}_{ab}$, which proves our claim for the first inequality. The second inequality can be proven with the exact same example. □

5. Symmetric games

In this section, we extend our results to continuous-time Markov games in which the two players have antagonistic objectives – the angelic player tries to maximise the time-bounded reachability probability while the demonic player tries to minimise it – but, different to the previous section, both players have no direct access to time. For a given CTG $\mathcal{M}$, a goal region G, and a time bound T, we establish the existence and computability of optimal strategies for both players in the time-abstract scheduler classes for the time-bounded reachability probability problem. That is, there is a pair of strategies for the following term that constitutes a Nash equilibrium.

$$\sup_{\mathcal{S}_A} \inf_{\mathcal{S}_D} Pr^G_{\mathcal{S}_{A+D}}(t), \quad \text{or} \quad \inf_{\mathcal{S}_D} \sup_{\mathcal{S}_A} Pr^G_{\mathcal{S}_{A+D}}(t), \tag{5}$$

where equality of both versions is guaranteed by [8, Theorem 3].

For *uniform* CTGs, this claim has recently been shown by Brázdil, Forejt, Krčál, Křetínský, and Kučera:

Theorem 5.1. *[8] For a given uniform CTGs $\mathcal{M}$, a goal region G and a time bound T, we can compute a bound $n_{\mathcal{M}}$ (comparable to our greed bound) and a memoryless deterministic greedy strategy $\mathcal{S} : L \to Act$, such that following $\mathcal{S}$ is optimal for both players with respect to CD after $n_{\mathcal{M}}$ steps.*

That is, optimal (counting) strategies for uniform Markov games have a similarly simple structure as those for CTMDPs. Now, we extend these results to history-dependent (HD and HR) schedulers:

Theorem 5.2. *[8] The optimal CD strategies from Theorem 5.1 (that is, for uniform CTGs) are also optimal for HR.*

Although this theorem has been proven already in [8], we give a similar, but less technical, proof in this work.

Proof. Let us assume the minimiser plays in accordance with her optimal CD strategy. Let us further assume that the maximiser has an HR strategy that yields a better result than his CD strategy. Then it must improve over his optimal CD strategy by a margin of some ε.

Let us define $p(k, l)$ as the maximum of the probabilities to still reach the goal region in the future that the maximiser can reach under the paths of length k which end in location l with the *better* history-dependent strategy. Further, let $h_l(k)$ be a path where this optimal value is taken. (Note that our goal region is absorbing.) The decision this HR scheduler takes is an affine combination of deterministic decisions, and the quality (the probability of reaching the goal region in the future) is the respective affine combination of the outcome of these pure decisions. Hence, there is at least one pure decision that (not necessarily strictly) improves over the randomised decision.

As our CTG is uniform, we can improve this history-dependent scheduler by changing all decisions it makes on a path $\pi = \pi'_l \circ \pi'$ that start with a path π'_l of length 2 ending in a location l, to the decisions it made upon the path $h_l(2) \circ \pi'$. (The improvement is not necessarily strict.) We then improve it further (again not necessarily strictly) by turning to the improved pure decision. The resulting strategy is initially counting – it depends only on the length of the history and the current location – and deterministic for paths up to length 2.

Having constructed a history-dependent scheduler that is initially counting and deterministic for paths up to length k, we repeat this step for paths $\pi = \pi'_l \circ \pi'$ that start with a history π'_l of length $k+1$, where we replace the decision made by our initially k counting and deterministic scheduler by the decision made on $h_l(k+1) \circ \pi'$, and then further to its deterministic improvement. This again leads to a – not necessarily strict – improvement.

Once the probability of making at least k steps falls below ε, any deterministic counting scheduler that agrees on the first k steps with a history-dependent scheduler from this sequence (which is initially counting and deterministic for at least k steps) improves over the counting scheduler we started with for the maximiser, which contradicts its optimality.

A similar argument can be made for the minimiser. □

Our argument that infers the existence of optimal strategies for general CTMDPs from the existence of optimal strategies for uniform CTMDPs does not depend on the fact that we have only one player with a particular objective. In fact, it can be lifted easily to Markov games.

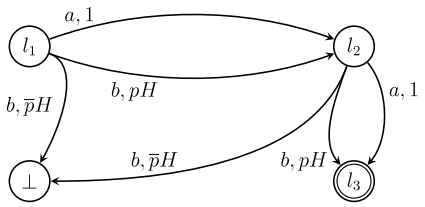

Fig. 7. Different to uniform CTGs, optimal decisions for the time-abstract player may need to be randomised in general CTGs. In the depicted game, a counting reachability player owns the location l_2 while his adversary (who can be time-dependent or time-abstract) owns the location l_1. We consider a time-bound $T = 1$ and a goal location l_3. The factor H is a huge factor, say 10^{10}, such that selecting an action b essentially results in an immediate transition. If a is always chosen, at least one discrete transition occurs with likelihood $p_1 = \frac{e-1}{e}$ and the second a transition occurs with likelihood $p_2 = \frac{e-2}{e-1}$ under the assumption that the first discrete transition is taken. Choosing $p = \sqrt{p_1 p_2} \approx 0.514$ then has the following effect: If the time-abstract reachability player chooses a (he always acts after one step), then the optimal strategy of the safety player is to choose a, resulting in an overall reachability probability of approximately $p_1 p_2$. If the time-abstract reachability player chooses b, then the optimal strategy of the safety player is to choose b (initially and long enough that we can approximate it by assuming it to be an immediate transition), resulting again in an overall reachability probability of approximately $p_1 p_2$. If the counting reachability player uses a mixed strategy, choosing each action with a 50% probability, then both pure strategies of the safety player (playing always a and playing always b) result in the same probability of $\frac{p_1}{2}(p + p_2) > p_1 p_2$. This probability is also the optimal time-bounded reachability.

Theorem 5.3. *For a Markov game $\mathcal{M}$, a goal region G, and a time bound T, optimal CD, CR, and HD schedulers exist, and an optimal HD scheduler is also optimal in the class of HR schedulers.*

Proof sketch. We start again with the uniformisation $\mathcal{U}$ of the Markov game $\mathcal{M}$. By Theorem 5.1, there is a deterministic memoryless greedy strategy for both players in $\mathcal{U}$ that is optimal after $n_{\mathcal{U}}$ steps. We argue along the same lines as for CTMDPs:

- We study the *visible* strategies on the uniformisation $\mathcal{U}$ of $\mathcal{M}$. Like in the constructions from Section 3.3, we use a bijection *vis* from the *visible strategies* on $\mathcal{U}$ onto the strategies of $\mathcal{M}$, which preserves the time-bounded reachability.
- We define $n_{\mathcal{U}}$-visible strategies analogously to the $n_{\mathcal{U}}$-visible schedulers to be those strategies, which can use the additional information provided by $\mathcal{U}$ after $n_{\mathcal{U}}$ visible steps have passed.

  After $n_{\mathcal{U}}$ visible steps, the class of $n_{\mathcal{U}}$-visible strategies clearly contains the deterministic greedy strategies described in the previous theorems of this section, as they can use all information after step $n_{\mathcal{U}}$. Using Theorem 5.1 we can deduce that, for both players, it suffices to seek an optimal $n_{\mathcal{U}}$-visible strategy in the subset of those strategies that turn to the *standard greedy strategy* after $n_{\mathcal{U}}$ *visible* steps.
- Locations l and their counterparts $l_{\mathcal{U}}$ have exactly the same exit rates for all actions, and therefore a greedy-optimal memoryless strategy will pick the same action for both locations (up to equal quality of actions). This directly implies that the standard greedy scheduler is a visible strategy, and with it all $n_{\mathcal{U}}$-visible strategies that turn to the standard greedy strategy after $n_{\mathcal{U}}$ visible steps are visible strategies. Hence, an optimal strategy for the class of $n_{\mathcal{U}}$-visible strategies that turns to the standard greedy strategy after $n_{\mathcal{U}}$ visible steps is also optimal for the class of visible strategies (time-abstract strategies in $\mathcal{M}$, respectively).
- For deterministic strategies, this class is finite, which immediately implies the existence of an optimum in this class (using Eq. (5)).

For history-dependent schedulers, randomised strategies again cannot provide an advantage over deterministic ones, because their outcome is just an affine combination of the outcome of the respective pure strategies, and the extreme points are taken at the fringe. (Technically, we can start with any randomised strategy and replace one randomised decision after another by a pure counterpart, improving the quality of the outcome – not necessarily strictly – for the respective player.)

For counting schedulers, randomisation can lead to an advantage. The intuitive reason for this is that a history-dependent strategy can use the history to infer a probability distribution over the time that still remains, whereas a counting strategy can only infer a probability distribution for a given counter-strategy. Fig. 7 provides an example and a technical explanation.

In case of CR schedulers, we can show the existence of optimal counting strategies by an argument similar to the one from Theorem 4.1. □

Computing optimal strategies. For the classes CD, HD, and HR this leaves us with a finite number of candidates. The optimal strategy can therefore – at least in principle – be found by applying a brute force approach: For all these deterministic strategies, we can compute and compare the reachability probabilities using the algorithm of Aziz et al. [21], which allows for identifying the deterministic strategies that mark an optimal Nash equilibrium.

6. Conclusions

We have demonstrated the existence of optimal control for time-bounded safety and reachability objective for time-abstract schedulers. For continuous-time Markov decision processes, we showed in Section 3 that finite optimal control exists, and that allowing for randomisation does not improve the overall result.

There are two natural extensions to games: From a purely theoretical point of view, it is interesting to consider games where both players face the same restriction. For this case we have shown in Section 5 that all results from CTMDPs extend to these games, using essentially the same techniques.

From a practical point of view, however, it is more natural to consider asymmetric games where only one player has restricted access to time: In such games, one of the players usually represents a controller or our control over the behaviour of a system, while his opponent represents an adversary used to abstract from the behaviour of an environment. To be conservative, such an antagonist should be unrestricted, and should therefore be able to make her decisions based on all information, including the information unavailable to the control. Such asymmetric continuous-time Markov games are studied in Section 4.

Surprisingly, optimal strategies asymmetric games do not necessarily show the simple structure we exploit in the symmetric case. We showed that there are instances of asymmetric CTGs for which optimal strategies cannot be limit-positional. Additionally, we showed that further results do not carry over from the symmetric case: Randomisation may increase the reachability probability, and even for uniform CTGs, history dependent strategies may improve over counting strategies.

In both symmetric and asymmetric games, CD schedulers can be strictly weaker than CR scheduler, as we showed in the simple example from Fig. 7. The intuitive difference to CTMDPs is that for them, we can (for a fixed strategy) infer the distribution of the time passed upon a particular history, while we cannot make a similar reasoning in games unless we know the strategy of our opponent. In CTMDPs, our 'opponent' is deterministic, and we can simply choose an optimal response. In games, a deterministic decision can be used by our opponent to our disadvantage, as seen in the example from Fig. 7.

Interestingly, the precise distribution upon a history can again be inferred for history dependent scheduler in symmetric games, which leads to HD and HR scheduler being equally powerful.

The need for randomisation sets time-abstract scheduling for games apart from time-dependent setting [10] and CTMDPs discussed in Section 3 alike.

Acknowledgments

We would like to thank Vojtech Forejt for pointing out an error in a previous version and the anonymous reviewers for their constructive critique.

This work was partly funded by the Engineering and Physical Science Research Council (EPSRC) through the grant EP/H046623/1 'Synthesis and Verification in Markov Game Structures', and by the German Research Foundation (DFG) under the project SpAGAT (grant no. FI 936/2-1) in the priority program 'Reliably Secure Software Systems – RS3'.

References

[1] E.A. Feinberg, Continuous time discounted jump Markov decision processes: a discrete-event approach, Mathematics of Operations Research 29 (2004) 492–524.
[2] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley-Interscience, 1994.
[3] L.I. Sennott, Stochastic Dynamic Programming and the Control of Queueing Systems, Wiley-Interscience, 1999.
[4] J. Bruno, P. Downey, G.N. Frederickson, Sequencing tasks with exponential service times to minimize the expected flow time or makespan, Journal of the ACM 28 (1981) 100–113.
[5] H. Hermanns, Interactive Markov Chains and the Quest for Quantified Quality, in: LNCS, vol. 2428, Springer-Verlag, 2002.
[6] M.A. Marsan, G. Balbo, G. Conte, S. Donatelli, G. Franceschinis, Modelling with Generalized Stochastic Petri Nets, John Wiley & Sons, 1995.
[7] R. Bellman, Dynamic Programming, Princeton University Press, 1957.
[8] T. Brázdil, V. Forejt, J. Krcál, J. Kretínský, A. Kucera, Continuous-time stochastic games with time-bounded reachability, in: Proceedings of FSTTCS'09, pp. 61–72.
[9] J. Filar, K. Vrieze, Competitive Markov decision processes, Springer-Verlag New York, Inc., New York, NY, USA, 1996.
[10] M. Rabe, S. Schewe, Finite optimal control for time-bounded reachability in CTMDPs and continuous-time Markov games, Acta Informatica 48 (2011) 291–315.
[11] C. Baier, H. Hermanns, J.-P. Katoen, B.R. Haverkort, Efficient computation of time-bounded reachability probabilities in uniform continuous-time Markov decision processes, Theoretical Computer Science 345 (2005) 2–26.
[12] P. Buchholz, E.M. Hahn, H. Hermanns, L. Zhang, Model checking algorithms for CTMDPs, in: Proceedings of CAV'11, pp. 225–242.
[13] M.R. Neuhäußer, M. Stoelinga, J.-P. Katoen, Delayed nondeterminism in continuous-time Markov decision processes, in: Proceedings of FOSSACS'09, pp. 364–379.
[14] M.R. Neuhäußer, L. Zhang, Time-bounded reachability probabilities in continuous-time Markov decision processes, in: Proceedings of QEST'10, pp. 209–218.
[15] L. Zhang, H. Hermanns, E.M. Hahn, B. Wachter, Time-bounded model checking of infinite-state continuous-time Markov chains, in: Proceedings of ACSD'08, pp. 98–107.
[16] N. Wolovick, S. Johr, A characterization of meaningful schedulers for continuous-time Markov decision processes, in: Proceedings of FORMATS'06, pp. 352–367.
[17] J. Fearnley, M. Rabe, S. Schewe, L. Zhang, Efficient approximation of optimal control for Markov games, in: Proceedings of FSTTCS'11, Leibniz International Proceedings in Informatics.
[18] M. Rabe, S. Schewe, Optimal Schedulers for Time-Bounded Reachability in CTMDPs, Reports of SFB/TR 14 AVACS, Nr. 55, 2009. http://www.avacs.org.
[19] M. Rabe, S. Schewe, Optimal time-abstract schedulers for CTMDPs and Markov games, in: Proceedings of QAPL'10, pp. 144–158.
[20] V.G. Kulkarni, Modeling and Analysis of Stochastic Systems, Chapman & Hall, Ltd., London, UK, 1995.
[21] A. Aziz, K. Sanwal, V. Singhal, R. Brayton, Model-checking continuous-time Markov chains, Transactions on Computational Logic 1 (2000) 162–170.