Optimal Time-Abstract Schedulers for CTMDPs and Markov Games*

Markus Rabe
Universität des Saarlandes
rabe@cs.uni-saarland.de

Sven Schewe
University of Liverpool
sven.schewe@liverpool.ac.uk

We study time-bounded reachability in continuous-time Markov decision processes for time-abstract scheduler classes. Such reachability problems play a paramount rôle in dependability analysis and the modelling of manufacturing and queueing systems. Consequently, their analysis has been studied intensively, and techniques for the approximation of optimal control 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.

We demonstrate the existence of optimal schedulers for the time-abstract scheduler classes for all CTMDPs. Our proof is constructive: We show how to compute optimal time-abstract strategies with finite memory. It turns out that these optimal schedulers have an amazingly simple structure—they converge to an easy-to-compute memoryless scheduling policy after a finite number of steps.

Finally, we show that our argument can easily be lifted to Markov games: We show that both players have a likewise simple optimal strategy in these more general structures.

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 [12, 5] or queueing systems [16]. We study a real time version of MDPs, continuous-time Markov decision processes (CTMDPs), which are a natural formalism for modelling in scheduling [4, 12] and stochastic control theory [5]. CTMDPs can also be seen as a unified framework for different stochastic model types used in dependability analysis [15, 12, 9, 7, 10].

The analysis of CTMDPs usually concerns the different possibilities to resolve the nondeterminism by means of a scheduler (also called strategy). 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?".

As a slight restriction, nondeterminism is either always hostile or always supportive in CTMDPs. Markov games [6] provide a generalisation of CTMDPs by disintegrating the control locations into locations where the nondeterminism is resolved angelically (supportive nondeterminism) and control locations where the nondeterminism is resolved demonically (hostile nondeterminism).

In this paper, we study the *maximal time-bounded reachability problem* [12, 2, 18, 10, 11, 3] in CTMDPs and Markov games. Time-bounded reachability is the standard control problem to construct a scheduler that controls the Markov decision process such that the likelihood of reaching a goal region

*This work was partly supported by the German Research Foundation (DFG) as part of the Transregional Collaborative Research Center "Automatic Verification and Analysis of Complex Systems" (SFB/TR 14 AVACS) and by the Engineering and Physical Science Research Council (EPSRC) through grant EP/H046623/1 "Synthesis and Verification in Markov Game Structures".

A. Di Pierro & G. Norman (Eds): 8th Workshop on Quantitative Aspects of Programming Languages (QAPL 2010)
EPTCS 28, 2010, pp. 144–158, doi:10.4204/EPTCS.28.10

within a given time bound is maximised, 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 scheduling policy naturally depends on the power a scheduler has 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 [10, 17]. Thereof, we consider those schedulers that have no direct access to time, the time-abstract schedulers. The time-abstract scheduler classes that can observe the history, its length, or nothing at all, are marked H (for history-dependent), C (for hop-counting), and P (for positional), respectively.

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 the maximum reachability probabilities of classes H and C coincide [2]. Uniform CTMDPs have a uniform transition rate λ for all their actions.

Optimal schedulers. Given its practical importance, the bounded reachability problem for Markov decision processes (and their deterministic counterpart the *Markov chains*) has been intensively studied [1, 2, 10, 3].

While previous research focused on *approximating* optimal scheduling policies [2], the existence of optimal schedulers for all scheduler classes has been demonstrated in Rabe's master thesis [13, 14], on which this paper is partly based. Meanwhile, Brazdil et al. [3] have independently provided a similar result for *uniform* Markov games, that is, for games that use the same transition rate for all actions.

Contribution. We start with a report on our work on counting (C) and history dependent (H) schedulers in *uniform* CTMDPs. Although the case of the counting schedulers could by now be inferred as a corollary from the existence of optimal counting strategies in Markov games [3], we decided to present it for 2.5 reasons: Firstly, it requires only marginal extra effort. Secondly, CTMDPs have been an important object of study for decades whereas Markov games are comparably new, and we think that our proof can provide insights in particular to readers that are not familiar with games. Finally, it was developed independently and at the same time.

We then show how our result on uniform CTMDPs can be lifted to general CTMDPs, and that randomisation cannot improve the quality of optimal scheduling. In Section 4, we show that our lifting argument naturally extends to Markov games: We show that there are optimal time-abstract counting and history dependent schedulers with finite memory for general Markov games and that—as for CTMDPs—randomisation cannot improve optimal scheduling for either player.

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.

The extension to non-uniform CTMDPs (and Markov games) builds upon a simple uniformisation technique and draws from a class of schedulers that are (partially) blind to the additional information introduced by the uniformisation. With the help of this scheduler class, we successively demonstrate that it is optimal (in the game case for both players) to turn to a fixed memoryless greedy strategy after a finite number of steps that is easy to compute. Hence, we can focus on scheduling policies that deviate from this scheduling policy only on a finite preamble. It then suffices to exclude that randomisation can improve the result (for either player) to reduce the candidate strategies to a finite set, and hence to infer the existence of simple optimal strategies for the non-uniform case as well.

2 Continuous-Time Markov Decision Processes

A *continuous-time Markov decision process* $\mathcal{M}$ is a tuple $(L, Act, \mathbf{R}, \nu, B)$ with a finite set of locations L, a finite set of actions Act, a rate matrix $\mathbf{R} : (L \times Act \times L) \to \mathbb{Q}_{\geqslant 0}$, an initial distribution $\nu \in Dist(L)$, and a goal region $B \subseteq L$. We define the total exit rate for a location l and an action a as $\mathbf{R}(l,a,L) = \sum_{l' \in L} \mathbf{R}(l,a,l')$. For a CTMDP 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 CTMDP $\mathcal{M}$ is a continuous-time Markov chain [8]. If multiple actions are available, we need to resolve the nondeterminism by means of a scheduler (also called strategy or scheduling policy). 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.

Uniform CTMDPs. We call a CTMDP uniform with rate λ if, for every location l and action $a \in Act(l)$, 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* $\mathcal{U}$ of a CTMDP $\mathcal{M}$ as the uniform CTMDP obtained by the following transformation steps. We create a copy $l_{\mathcal{U}}$ for every $l \in L$ and obtain $L_{\mathcal{U}} = \bigcup_{l \in L} \{l, l_{\mathcal{U}}\}$. We call the new copies unobservable, and all locations $l \in L$ observable. Let λ be the maximal total exit rate in $\mathcal{M}$. The new rate matrix $\mathbf{R}_{\mathcal{U}}$ extends $\mathbf{R}$ by first adding the rate $\mathbf{R}_{\mathcal{U}}(l,a,l_{\mathcal{U}}) = \lambda - \mathbf{R}(l,a,L)$ for every location $l \in L$ and action $a \in Act$ of $\mathcal{M}$, and by then copying the outgoing transitions from every observable location l to its unobservable counterpart $l_{\mathcal{U}}$, while the other components remain untouched. The intuition behind this uniformisation technique is that it enables us to distinguish whether a step would have occurred in the original automaton or not.

Paths. A *timed path* in CTMDP $\mathcal{M}$ 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 sequence 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 CTMDP, we additionally need a scheduler that resolves the nondeterminism that occurs in locations where multiple actions are enabled. When analysing properties of a CTMDP, such as the reachability probability, we usually quantify over a class of schedulers. In this paper, we consider the following common scheduler classes, which differ in their power to observe and distinguish events:

- *Time-abstract history-dependent* (H) schedulers that map time-abstract paths to decisions. $Paths_{abs}(\mathcal{M}) \to D$
- *Time-abstract hop-counting* (C) schedulers that map locations and the length of the path to decisions. $L \times \mathbb{N} \to D$
- *Positional* (P) or memoryless schedulers that map locations to decisions. $L \to D$

Decisions D are either randomised (R), in which case $D = Dist(Act)$ is the set of distributions over enabled actions, or are restricted to deterministic (D) choices, that is $D = Act$. Where it is necessary to distinguish randomised and deterministic versions we will add a postfix to the scheduler class, for example HD and HR. We restrict all scheduler classes to those schedulers creating a measurable probability space (cf. [17]).

Induced Probability Space. We build our probability space in the natural way: we first define the probability measure for cylindric sets of paths that start with

$$l_0 \xrightarrow{a_0,t_0} l_1 \xrightarrow{a_1,t_1} \cdots \xrightarrow{a_{n-1},t_{n-1}} l_n,$$

with $t_j \in I_j$ for all $j < n$, and for non-overlapping open intervals $I_0, I_1, \ldots, I_{n-1}$, to be the usual probability that a path starts with these actions for a given randomised scheduler $\mathcal{S}$, and such that $\mathcal{S}(l_0 \xrightarrow{a_0,t_0} \ldots \xrightarrow{a_{i-1},t_{i-1}} l_i)$ is equivalent for all $(t_0, \ldots, t_{i-1}) \in I_0 \times \ldots \times I_{i-1}$:

$$\int_{t_0 \in I_0, t_1 \in I_1, \ldots, t_{n-1} \in I_{n-1}} \prod_{i=0}^{n-1} \mathcal{S}(l_0 \xrightarrow{a_0,t_0} \ldots \xrightarrow{a_{i-1},t_{i-1}} l_i)(a_i) \cdot \mathbf{R}(l_i, a_i, l_{i+1}) \cdot e^{-\mathbf{R}(l_i,a_i,L)(t_i - t_{i-1})},$$

assuming $t_{-1} = 0$.

From this basic building block, we build our probability measure for measurable sets of paths and measurable schedulers in the usual way (cf. [17]).

Time-Bounded Reachability Probability. For a given CTMDP $\mathcal{M} = (L, Act, \mathbf{R}, \nu, B)$ and a given measurable scheduler $\mathcal{S}$ that resolves the nondeterminism, we use the following notations for the probabilities:

- $Pr^{\mathcal{M}}_{\mathcal{S}}(l,t)$ is the probability of reaching the goal region B in time t when starting in location l,
- $Pr^{\mathcal{M}}_{\mathcal{S}}(t) = \sum_{l \in L} \nu(l) Pr^{\mathcal{M}}_{\mathcal{S}}(l,t)$ denotes the probability of reaching the goal region B in time t,
- $Pr^{\mathcal{M}}_{\mathcal{S}}(t;k)$ denotes the probability of reaching the goal region B in time t *and* in at most k discrete steps, and
- $PR^{\mathcal{M}}_{\mathcal{S}}(\pi,t)$ is the probability to traverse the time-abstract path π within time t.

As usual, the supremum of the time-bounded reachability probability over a particular scheduler class is called the time-bounded reachability of $\mathcal{M}$ for this scheduler class, and we use 'max' instead of 'sup' to indicate that this value is taken for some *optimal scheduler* $\mathcal{S}$ of this class.

Step Probability Vector. Given a scheduler $\mathcal{S}$ and a location l for a CTMDP $\mathcal{M}$, we define the *step probability vector* $d_{l,\mathcal{S}}$ of infinite dimension. An entry $d_{l,\mathcal{S}}[i]$ for $i \geq 0$ denotes the probability to reach goal region B in up to i steps from location l (not considering any time constraints).

3 Optimal Time-Abstract Schedulers

In this section, we show that *optimal* schedulers exist for all natural time-abstract classes, that is, for CD, CR, HD, and HR. Moreover, we show that there are optimal schedulers that become positional after a small number of steps, which we compute with a simple algorithm. We also show that randomisation

does not yield any advantage: deterministic schedulers are as good as randomised ones. Our proofs are constructive, and thus allow for the construction of optimal schedulers. This also provides the first procedure to precisely determine the time-bounded reachability probability, because we can now reduce this problem to solving the time-bounded reachability problem of continuous-time Markov chains [1].

Our proof consists of two parts. We first consider the class of uniform CTMDPs, which are much simpler to treat in the time-abstract case, 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 [2]. It therefore suffices to show that there is a CD scheduler which takes this value.

We then show that a similar claim holds for CD and HD scheduler in the general class of not necessarily uniform CTMDPs. In this case, it also holds that there are simple optimal schedulers that converge against a positional scheduler after a finite number of steps, and that randomisation does not improve the time-bounded reachability probability. However, 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 [2] for a simple example).

We start this section with the introduction of *greedy schedulers*, HD schedulers that favour reachability in a small number of steps over reachability with a larger number of steps; the positional schedulers against which the CD and HD schedulers converge are such greedy schedulers.

3.1 Greedy Schedulers

The objective we consider is to maximise time-bounded reachability $Pr_{\mathcal{S}}^{\mathcal{M}}(l,t)$ for every location l 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 (that is, the remaining time t is close to 0).

Time-abstract schedulers have no direct access to the time, but they can infer the distribution over the remaining time from the time-abstract history (or its length). When examining the resulting Poisson distribution one can easily see that for large step numbers the probability to take more than one further step 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 in further steps (after sufficiently many steps).

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. Greedy schedulers that optimise the reachability within the first k steps have been exploited in the efficient analysis of CTMDPs [2]. 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 $d_l = \sup_{\mathcal{S}\in HD} d_{l,\mathcal{S}}$ 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 $l \notin B$ 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 B, all enabled actions $a \in Act(l)$ are greedy.

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 $l \notin B$, $shift(d_l) \geq \sum_{l'\in L} \mathbf{P}(l,a,l')d_{l'}$ holds for every action a, and that equality must hold for some.

For a scheduler $\mathcal{S}$ that always chooses greedy actions, a simple inductive argument shows that $d_l[i] = d_{l,\mathcal{S}}[i]$ holds for all $i \in \mathbb{N}$, while it is easy to show that $d_l > d_{l,\mathcal{S}}$ holds if $\mathcal{S}$ deviates from greedy decisions upon a path that is possible under its own scheduling policy and does not contain a goal location. □

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 $\mathit{shift}(d_{l,\mathcal{S}}) \leq \sum_{l' \in L} \mathbf{P}(l,a,l') d_{l'}$ holds true, 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 $d_{l,a} = \sum_{l' \in L} \mathbf{P}(l,a,l') d_{l'}$ denote the step probability vector of such schedulers.

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 Subsection 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 this standard greedy scheduler.

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:

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

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

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

holds true. It suffices to choose $n_{\mathcal{M}} \geq \frac{2\lambda t}{\mu}$ since it implies $\mu p_{\lambda t}(n) \geq 2 p_{\lambda t}(n+1)$, $\forall n > n_{\mathcal{M}}$ (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 $\mathcal{S}$ 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 $\overline{\mathcal{S}}$ that makes the same decisions as $\mathcal{S}$ 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 $Pr^{\mathcal{M}}_{\overline{\mathcal{S}}}(t) \geq Pr^{\mathcal{M}}_{\mathcal{S}}(t)$ holds true.

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 $\mathcal{S}$ 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 $\mathcal{S}$, then we can change the decision of $\mathcal{S}$ on every path starting with π arbitrarily—and in particular to the standard greedy scheduler—without altering the reachability probability.

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

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

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

$$Pr^{\mathcal{M}}_{\mathcal{S}'}(t) - Pr^{\mathcal{M}}_{\mathcal{S}}(t) = PR^{\mathcal{M}}_{\mathcal{S}}(\pi,t) \cdot \left(\mu p_{\lambda t}(|\pi|+j) + \sum_{i>j}^{\infty} (d_l[i] - d_{l,\mathcal{S}_\pi}[i]) \cdot p_{\lambda t}(|\pi|+i) \right) \quad \text{(for some } j > 0\text{)}$$

we can apply Equation 2 to deduce that the difference $Pr^{\mathcal{M}}_{\mathcal{S}'}(t) - Pr^{\mathcal{M}}_{\mathcal{S}}(t)$ 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 $\mathcal{S}$ makes non-greedy decisions. In this way, we obtain a scheduler $\mathcal{S}''$ 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 $\mathcal{S}$.

Since all greedy schedulers are interchangeable without changing the time-bounded reachability probability (and even without altering the step probability vector), we can modify $\mathcal{S}''$ such that it follows the standard greedy scheduling policy after $\geq n_{\mathcal{M}}$ steps, resulting in a scheduler $\overline{\mathcal{S}}$ that comes with the same time-bounded reachability probability as $\mathcal{S}''$. Note that $\overline{\mathcal{S}}$ is counting if $\mathcal{S}$ 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 bounded reachability is therefore the maximal bounded reachability 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 [2], we can infer that such a scheduler is optimal for all of these classes.

Corollary 3.3 $\max_{\mathcal{S}\in CD} Pr^{\mathcal{M}}_{\mathcal{S}}(t) = \max_{\mathcal{S}\in HR} Pr^{\mathcal{M}}_{\mathcal{S}}(t)$ *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 [2].

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

$$\max_{\mathcal{S}\in CD} Pr^{\mathcal{M}}_{\mathcal{S}}(t) = \max_{\mathcal{S}\in CR} Pr^{\mathcal{M}}_{\mathcal{S}}(t) \qquad \text{and} \qquad \max_{\mathcal{S}\in HD} Pr^{\mathcal{M}}_{\mathcal{S}}(t) = \max_{\mathcal{S}\in HR} Pr^{\mathcal{M}}_{\mathcal{S}}(t).$$

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

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

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 $\mathcal{M}$:*

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

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.

Lemma 3.7 *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 3.8 *Optimal CD schedulers are also optimal CR schedulers.*

Proof First, for $n \to \infty$ the probability to reach the goal region B in exactly n or more than n steps converges to 0, independent of the scheduler. Together with Lemma 3.7, this implies

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

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

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

Theorem 3.9 *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 until 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, as we never know when to stop when comparing step probability vectors. In this lemma, however, we show that it suffices to check for equivalence of two step probability vectors only up to position $|L|-2$. As discussed in Subsection 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 3.10 *Given a uniform CTMDP $\mathcal{M}$, the smallest k that satisfies $\forall l \in L,\ a \in Act(l).\ d_l \neq d_{l,a} \Rightarrow \exists k' \leq k.\ d_l[k'] > d_{l,a}[k']$ is bounded by $|L|-2$.*

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 $\mathcal{S}$ whose existence was shown in Subsection 3.1.

We first lift the step probability vector from locations to distributions, where $d_\nu = \sum_{l \in L} \nu(l) d_l$ is, for a distribution $\nu : L \to [0,1]$, the affine combination of the step probability vectors of the individual locations. In this proof, we define two distributions $\nu, \nu' : L \to [0,1]$ to be equivalent, if their step probability vectors $d_\nu = d_{\nu'}$ are equal. Further, we call them i-step equivalent if they are equal up to position i ($\forall j \leq i.\ d_\nu[j] = d_{\nu'}[j]$).

In order to argue with vector spaces, we extend these definitions to arbitrary vectors $\nu : L \to \mathbb{R}$ (instead of $\nu : L \to [0,1]$).

Let D_i be the vector space spanned by i-step equivalent distributions ν, ν' over L. Naturally, $D_i \supseteq D_{i+1}$ always holds, as $i+1$ step equivalence implies i-step equivalence. In addition 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 $D_{|L|-2} = D_j$ for all $j \geq |L|-2$.

- D_0 has $|L|-2$ dimensions: D_0 is the vector space that contains 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 equivalency; $d_\nu[0] = d_{\nu'}[0]$).

  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 $\sum_{l \in B} \nu(l) = \sum_{l \in B} \nu'(l)$, and hence $\sum_{l \in B} \delta(l) = 0$. Thus, D_0 has $|L|-2$ dimensions. (Assuming $B \neq L, B \neq \emptyset$, 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 $\mathit{shift}(d_\nu)$ and $\mathit{shift}(d_{\nu'})$ are i-step equivalent, that is, $\mathit{shift}(d_\nu) - \mathit{shift}(d_{\nu'}) \in D_i$.

  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 simple 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 shift 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.

Naturally, this implies that the vector spaces are shrinking, that is, $D_0 \supseteq D_1 \supseteq \ldots \supseteq D_{|L|-2} \supseteq \ldots$, and that $D_i = D_{i+1}$ implies that a fixed point is reached. (It implies $O_i = O_{i+1}$ and hence $D_i = D_j$ $(\forall j \geq i)$ by a simple inductive argument.)

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. That is, two distributions 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 $l \notin B$ 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. [1] 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 3.11 *We can effectively construct optimal CD, CR, HD, and HR schedulers.* □

Corollary 3.12 *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 rely on the precise CTMC model checking approach of Aziz et al. [1], 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 Figure 1. As $\mathcal{M}$ is not uniform, we start with constructing the uniformisation $\mathcal{U}$ of $\mathcal{M}$ (cf. Figure 1).

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

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

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

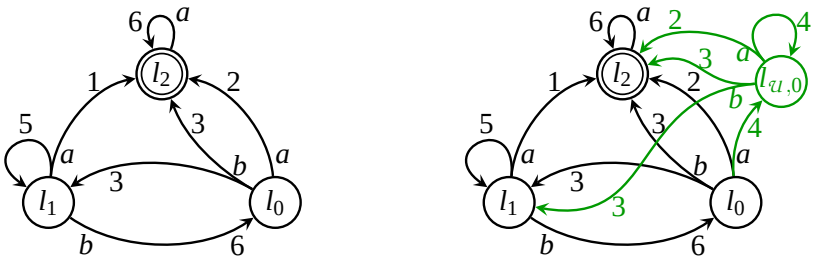

Figure 1: The example CTMDP $\mathcal{M}$ (left) and the reachable part of its uniformisation $\mathcal{U}$ (right).

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 Equation (2).

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 describe how the precise model checking technique of Aziz et al. [1] can be exploited to turn the existence proof into an effective technique for the construction of optimal schedulers.

4 Extension to Continuous-Time Markov Games

Markov decision processes can easily be extended to continuous-time Markov games (CTGs) $\mathcal{G} = (L_A, L_D, Act, \mathbf{R}, \nu, B)$ by disintegrating the set of locations into game positions of a maximiser (L_A, angelic game positions) and a minimiser (L_D, demonic game positions). These two players have antagonistic objectives to maximise and minimise the time-bounded reachability probability. These games are closely related to the CTMDP framework, and we define, for a given Markov game $\mathcal{G}$, the *underlying* CTMDP $\mathcal{M} = (L_A \dot{\cup} L_D, Act, \mathbf{R}, \nu, B)$. CTGs are called *uniform* if their underlying CTMDP is uniform.

The players can choose an action upon the entrance to one of their locations, and, as with schedulers for CTMDPs, they may have limited access to the timed history of the system. We only consider time-abstract strategies $S_X : Paths^X_{abs}(\mathcal{G}) \to Dist(Act)$ for both players, where paths are defined over the underlying CTMDP, and $Paths^X_{abs}(\mathcal{G})$ (for $X \in \{A, D\}$) is the set of paths that end with a location in L_X.

Obviously, there is a one-to-one mapping between *combined strategies*

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

of a CTG and schedulers of the underlying CTMDP.

For a given CTG and a pair of strategies $\mathcal{S}_A$, $\mathcal{S}_D$ we define the according probability space equivalent to the probability space of the underlying CTMDP with the combined strategy S_{A+D}. Then, the time-bounded reachability probability can be formulated for CTGs as follows:

$$\sup_{\mathcal{S}_A} \inf_{\mathcal{S}_D} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(t) = \inf_{\mathcal{S}_D} \sup_{\mathcal{S}_A} Pr^{\mathcal{G}}_{\mathcal{S}_{A+D}}(t) \tag{3}$$

where equality is guaranteed by [3, Theorem 3.1].

For uniform CTGs, a theorem similar to Theorem 3.2 has recently been shown:

Theorem 4.1 *[3] For uniform CTGs $\mathcal{G}$ with counting strategies, we can compute a bound $n_{\mathcal{G}}$ (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 after $n_{\mathcal{G}}$ 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 4.2 *The optimal CD strategies from Theorem 4.1 (that is, for uniform CTGs) are also optimal HR strategies.*

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.

Theorem 4.3 *For a Markov game $\mathcal{G}$, we can effectively construct optimal CD and HD schedulers, which are also optimal CR and HR schedulers, respectively, and 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.*

Proof sketch We start again with the uniformisation $\mathcal{U}$ of the Markov game $\mathcal{G}$. By Theorem 4.1, there is a deterministic memoryless greedy strategy for both players in $\mathcal{U}$ that is optimal after $n_{\mathcal{U}}$ steps. Hence, we can argue along the same lines as for CTMDPs:

- We study the *visible* strategies on the uniformisation $\mathcal{U}$ of $\mathcal{G}$. Like in the constructions from Section 3.3, we use a bijection *vis* from the visible strategies on $\mathcal{U}$ onto the strategies of $\mathcal{G}$, 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 4.1 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 turn to the standard greedy strategy after $n_{\mathcal{U}}$ visible steps is also optimal for the class of visible strategies (time-abstract strategies in $\mathcal{G}$, respectively).
- For deterministic strategies, this class is finite, which immediately implies the existence of an optimum in this class (using Equation 3).

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.)

Consequently, we are left with a finite set of history dependent or counting candidate strategies, respectively, and the result can—at least in principle—be found by applying a brute force approach: For each of these deterministic strategies, we can compute the reachability probability using the algorithm of Aziz et al. [1], which allows for identifying the deterministic strategies that mark an optimal Nash equilibrium. □

References

[1] Adnan Aziz, Kumud Sanwal, Vigyan Singhal, and Robert Brayton. Model-checking continuous-time Markov chains. *Transactions on Computational Logic*, 1(1):162–170, 2000.

[2] Christel Baier, Holger Hermanns, Joost-Pieter Katoen, and Boudewijn R. Haverkort. Efficient computation of time-bounded reachability probabilities in uniform continuous-time Markov decision processes. *Theoretical Computer Science*, 345(1):2–26, 2005.

[3] Tomas Brazdil, Vojtech Forejt, Jan Krcal, Jan Kretinsky, and Antonin Kucera. Continuous-time stochastic games with time-bounded reachability. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2009)*, pages 61–72, 2009.

[4] J. Bruno, P. Downey, and G. N. Frederickson. Sequencing Tasks with Exponential Service Times to Minimize the Expected Flow Time or Makespan. *Journal of the ACM*, 28(1):100–113, 1981.

[5] Eugene A. Feinberg. Continuous Time Discounted Jump Markov Decision Processes: A Discrete-Event Approach. *Mathematics of Operations Research*, 29(3):492–524, 2004.

[6] Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer-Verlag New York, Inc., New York, NY, USA, 1996.

[7] H. Hermanns. *Interactive Markov Chains and the Quest for Quantified Quality. LNCS 2428*. Springer-Verlag, 2002.

[8] Vidyadhar G. Kulkarni. *Modeling and Analysis of Stochastic Systems*. Chapman & Hall, Ltd., London, UK, 1995.

[9] M. A. Marsan, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis. Modelling with Generalized Stochastic Petri Nets. *SIGMETRICS Performance Evaluation Review*, 26(2):2, 1998.

[10] Martin R. Neuhäußer, Mariëlle Stoelinga, and Joost-Pieter Katoen. Delayed Nondeterminism in Continuous-Time Markov Decision Processes. In *Proceedings of FOSSACS '09*, pages 364–379, 2009.

[11] Martin R. Neuhäußer and Lijun Zhang. Time-Bounded Reachability in Continuous-Time Markov Decision Processes. Technical report, 2009.

[12] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, April 1994.

[13] Markus Rabe. Optimal Schedulers for Time-Bounded Reachability in CTMDPs. Master's thesis, Saarbrücken Graduate School of Computer Science, Saarbrücken, Germany, September 2009.

[14] Markus Rabe and Sven Schewe. Optimal Schedulers for Time-Bounded Reachability in CTMDPs. Reports of SFB/TR 14 AVACS 55, SFB/TR 14 AVACS, October 2009. http://www.avacs.org.

[15] William H. Sanders and John F. Meyer. Reduced Base Model Construction Methods for Stochastic Activity Networks. In *Proceedings of PNPM'89*, pages 74–84, 1989.

[16] Linn I. Sennott. *Stochastic Dynamic Programming and the Control of Queueing Systems*. Wiley-Interscience, 1999.

[17] Nicolás Wolovick and Sven Johr. A Characterization of Meaningful Schedulers for Continuous-Time Markov Decision Processes. In *Proceedings of FORMATS'06*, pages 352–367, 2006.

[18] L. Zhang, H. Hermanns, E. M. Hahn, and B. Wachter. Time-bounded model checking of infinite-state continuous-time Markov chains. In *Proceedings of ACSD'08*, pages 98–107, 2008.