Finite Optimal Control for Time-Bounded Reachability in CTMDPs and Continuous-Time Markov Games

Markus N. Rabe[1] and Sven Schewe[2]

[1] Universität des Saarlandes
[2] University of Liverpool

Abstract. We establish the existence of optimal scheduling strategies for time-bounded reachability in continuous-time Markov decision processes, and of co-optimal strategies for continuous-time Markov games. Furthermore, we show that optimal control does not only exist, but has a surprisingly simple structure: the optimal schedulers from our proofs are deterministic and timed positional, and the bounded time can be divided into a finite number of intervals, in which the optimal strategies are positional. That is, we demonstrate the existence of *finite* optimal control. Finally, we show that these pleasant properties of Markov decision processes extend to the more general class of continuous-time Markov games, and that both early and late schedulers show this behaviour.

1 Introduction

Continuous-time Markov decision processes (CTMDPs) are a widely used framework for dependability analysis and for modelling the control of manufacturing processes [8, 17, 24], because they combine real-time aspects with probabilistic behaviour and non-deterministic choices. CTMDPs can also be viewed as a framework that unifies different stochastic model types [5, 12, 13, 17, 22].

While CTMDPs allow for analysing worst-case and best-case scenarios, they fall short of the demands that arise in many real control problems, as they disregard the different nature that non-determinism can have depending on its source: some sources of non-determinism are supportive, while others are hostile, and in a realistic control scenario, we face both types of non-determinism at the same time: supportive non-determinism can be used to model the influence of a controller on the evolution of a system, while hostile non-determinism can capture abstraction or unknown environments. We therefore consider a natural

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), the project SpAGAT in the DFG priority programme RS^3, and by the Engineering and Physical Science Research Council (EPSRC) through grant EP/H046623/1 "Synthesis and Verification in Markov Game Structures".

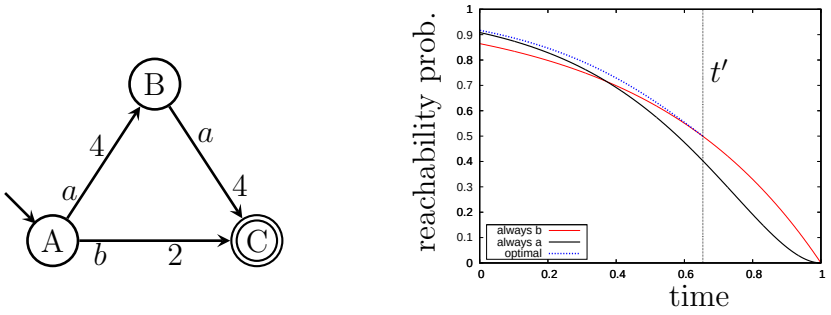

Fig. 1. A CTMDP and the reachability probabilities for all positional schedulers with time bound $t_{\max} = 1$. Time $t' = t_{\max} - \frac{1}{2}\log(2)$ is the optimal time to switch to action b.

extension of CTMDPs: continuous-time Markov games (CTMGs) that have two players with opposing objectives [5, 6, 9–11, 26].

The analysis of CTMDPs and CTMGs requires us to resolve the non-deterministic choices by means of a scheduler (which consists of a pair of strategies in the case of CTMGs), and typically tries to optimise a given objective function.

Contributions. In this article, we study the *time-bounded reachability problem*, which recently enjoyed much attention [3, 6, 15, 16, 27, 28] due to its relevance for quantitative model checking. Time-bounded reachability in CTMDPs is the standard control problem to construct a scheduler that controls the Markov decision process such that the probability of reaching a goal region within a given time bound is maximised (or minimised), and to determine the value. For CTMGs, time-bounded reachability reduces to finding a Nash equilibrium, that is, a pair of strategies for the players, such that each strategy is optimal for the chosen strategy of her opponent.

This article has three main contributions:

- First, we extend the common model of CTMDPs by adding discrete locations, which are passed in 0 time. This generalisation of the model is mainly motivated by avoiding the discussion about the appropriate scheduler class.
- The second contribution of this article is the answer to an intriguing research question: we show that optimal control of CTMDPs exists for time-bounded reachability and safety objectives. Moreover, we show that optimal control can always be *finite*.
- Our third contribution is to lift these results to continuous-time Markov games.

Related Work. Optimal control in CTMDPs clearly depends on the observational power we allow our schedulers to have when observing a run. In the literature, various classes of schedulers with different restrictions on what they can observe [6, 15, 20, 25] are considered. We focus on the most general class of schedulers, schedulers that can fully observe the system state and may change their decisions at any point in time (*late schedulers*, cf. [5, 14, 26]). To be able to transfer our results to the class of schedulers that fix their decisions when

entering a location (early schedulers [25]), we introduce discrete locations that allow for a translation from early to late schedulers (see Section 5.1).

Due to their practical importance, time-bounded reachability for continuous-time Markov models has been studied intensively (for example [2–6, 15, 16, 27, 28]). However, most previous research focussed on *approximating* optimal control. The existence of optimal control has only been known for the artificial class of *time-abstract* schedulers [6, 20], which assume that the scheduler has no access whatsoever to a clock. In a work independent of the reports [18, 19] this article is based on, Neuhäußer and Zhang showed the existence of optimal timed positional schedulers for the restricted class of locally uniform CTMDPs [16].

While efficient approximation [21, 7, 16] is of interest for a practitioner, being unable to determine whether or not optimal control *exists*—or if the optimal quality which was described more than half a century ago by Bellman [5] may be a limit that cannot be reached—is very dissatisfying from a theoretical point of view.

Technique in a Nutshell. Pursuing a different research question, we exploit proof techniques that differ from those frequently used in the analysis of CTMDPs. Our proofs build mainly on topological arguments: the proof that demonstrates the existence of measurable optimal schedulers, for example, shows that we can fix the decisions of an optimal scheduler successively on closures of open sets (yielding only measurable sets), and the lift to finiteness uses local optimality of positional schedulers in open left and right environments of arbitrary points of times and the compactness of the considered time interval.

For our proofs it turned out to be much more convenient to use a Lebesgue measure rather than the more widespread Borel measure. While this is but a minor technical decision from a practical point of view, it also shows that the particular choice of measure should be driven by convenience only. Indeed, our proof of the existence of finite optimal control shows that there are optimal solutions in the weakest class of cylindrical schedulers, which form but the starting point for the definition of a measure, be it a Borel or a Lebesgue measure.

Structure of the Article. We follow a slightly unorthodox order of proofs for a mathematical article: we start with a special case in Section 3 and generalise the results later. Besides keeping the proofs simple, this approach is chosen because the simplest case, CTMDPs, is the classical case, and we assume that a wider audience is interested in results for these structures. In Section 4, we strengthen this result by demonstrating that optimal control does not only exist, but can be found among schedulers with finitely many switching points and positional strategies between them. In Section 5, we lift this result to single player games (CTMDPs with discrete locations, that is). In the final section, we generalise the existence theorem for finite optimal control to finite co-optimal strategies for continuous-time Markov games.

2 Preliminaries

A continuous-time Markov game is a tuple $(L, L_d, L_c, L_r, L_s, Act, \mathbf{R}, \mathbf{P}, \nu)$, consisting of

- a finite set L of locations, which is partitioned
  - into a set L_d of *discrete* locations and a set L_c of *continuous* locations, and
  - into sets L_r and L_s of locations owned by a *reachability* and a *safety* player, respectively,
- a finite set Act of actions,
- a rate matrix $\mathbf{R} : (L_c \times Act \times L) \to \mathbb{Q}_{\geqslant 0}$,
- a discrete transition matrix $\mathbf{P} : (L \times Act \times L) \to \mathbb{Q}_{\geqslant 0} \cap [0, 1]$, and
- an initial distribution $\nu \in Dist(L)$,

that satisfies the following side-conditions: for all continuous locations $l \in L_c$, there must be an action $a \in Act$ such that $\mathbf{R}(l, a, L) := \sum_{l' \in L} \mathbf{R}(l, a, l') > 0$; we call such actions *enabled*. For actions enabled in continuous locations, we require $\mathbf{P}(l, a, l') = \frac{\mathbf{R}(l,a,l')}{\mathbf{R}(l,a,L)}$, and we require $\mathbf{P}(l, a, l') = 0$ for the remaining actions. For discrete locations, we require that either $\mathbf{P}(l, a, l') = 0$ holds for all $l' \in L$, or that $\sum_{l' \in L} \mathbf{P}(l, a, l') = 1$ holds. Like in the case of continuous locations, we call the latter actions *enabled* and require the existence of at least one enabled action for each discrete location $l \in L_d$.

The idea behind discrete-time locations is that they execute immediately. We therefore do not permit cycles of only discrete-time locations (counting every positive rate of any action as a transition). This restriction is stronger than it needs to be, but it simplifies our proofs, and the simpler model is sufficient for our means.

Intuitively, it is the objective of the reachability player to maximise the probability to reach a goal region in a predefined time t_0, while it is the objective of the safety player to minimise this probability. (Hence, it is a zero-sum game.)

We are particularly interested in (traditional) CTMDPs. They are single player CTMGs, where either all positions belong to the reachability player ($L = L_r$), or to the safety player ($L = L_s$), without discrete locations ($L_d = \emptyset$ and $L_c = L$).

2.1 Paths

A *timed path* π in a CTMG $\mathcal{M}$ is a finite sequence in $L \times (Act \times \mathbb{R}_{\geqslant 0} \times L)^* = Paths$. 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 $0 \leq t_{i-1} \leq t_i \leq T$ for all $i < n$, where n is the length of the path and T is an arbitrary time bound for the system. (For technical reasons, we define the system's behaviour only in an interval $[0, T]$). The t_i denote the system's time when the action a_i is selected and a discrete transition from l_i to l_{i+1} takes place. Concatenation of paths π, π' will be written

as $\pi \circ \pi'$ if the last location of π is the first location of π' and the transition times are ordered correctly. We call a timed path a *complete* timed path when we want to stress that this path describes a complete system run, not to be extended by further transitions.

Further, we use $Paths_n$ to denote the set of all timed paths of length n.

2.2 Schedulers and Strategies

The nondeterminism in the system needs to be resolved by a scheduler, which maps paths to decisions. The power of schedulers is determined by their ability to observe and distinguish paths, and thus by their domain. In this article, we consider the following common scheduler classes:

- *Timed history-dependent* (TH) schedulers, $Paths(\mathcal{M}) \times \mathbb{R}_{\geqslant 0} \to Dist(Act)$, that map timed paths and the remaining time to distributions over actions.
- *Timed positional* (TP) schedulers, $L \times \mathbb{R}_{\geqslant 0} \to Dist(Act)$, that map locations and the remaining time to distributions over actions.
- *Positional* (P) or memoryless schedulers, $L \to Dist(Act)$, that map locations to distributions over actions.

We call a scheduler deterministic, if it selects one action with probability 1 and all other actions with probability 0. For convenience, we will sometimes consider the set of actions (instead of the set of distributions over actions) as the codomain of deterministic schedulers (e.g. $L \to Act$ for positional deterministic schedulers). Where it is necessary to distinguish between randomised and deterministic classes we indicate this via the postfixes R and D, respectively; for example THR and THD.

Strategies. In case of CTMGs, a scheduler consists of the two participating players' strategies, which can be seen as functions $Paths_p(\mathcal{M}) \times \mathbb{R}_{\geqslant 0} \to Dist(Act)$, where $Paths_p(\mathcal{M})$ denotes, for $p \in \{r, s\}$, the paths ending on the position of the reachability or safety player, respectively. As for general schedulers, we can introduce restrictions on what players are able to observe.

Late and early scheduling. The main motivation to introduce discrete locations was to avoid the discussion whether a scheduler has to fix its decision as to which action it chooses, upon entering a location (early), or whether such a decision can be revoked while staying in the location (late). For example, the general measurable schedulers discussed in [25] have only indirect access to the remaining time (through the timed path), and therefore have to decide upon entrance of a location which action they want to perform. Our definition builds on fully-timed schedulers (cf. [5]) that were recently rediscovered and formalised by Neuhäußer et al. [15], which may revoke their decision after they enter a location. (As a side result, we lift Neuhäußer's restriction of local uniformity.) The discrete locations now allow us to encode schedulers that make their decision upon entering a continuous location l by mapping the decision to a discrete location that is 'guarding the entry' to a family of continuous locations, one for each action enabled in l. (See Section 5.1 for details.)

2.3 A Primitive Probability Measure

While it is common to refer to TH schedulers as a class, the truth is that there is no straightforward way to define a time-bounded reachability probability for the complete class (cf. [25]). We therefore start with a natural subset, the cylindrical schedulers, which will be used in the following subsections to build a powerful yet measurable sub-class of TH schedulers.

Let $\mathcal{J}$ be a finite partition of the interval $[0, T]$ into intervals $I_0 = [0, t_0]$ and $I_i = (t_{i-1}, t_i]$ for $i = 1, \ldots, n$ with $t_0 \geq 0$ and $t_i > t_{i-1}$ for $i = 1, \ldots, n$. Then we denote with $[t]_{\mathcal{J}}$ the interval $I_i \in \mathcal{J}$ that contains t, called the *$\mathcal{J}$-cylindrification of t*, and we denote with $[\pi]_{\mathcal{J}} = l_0 \xrightarrow{a_0,[t'_0]_{\mathcal{J}}} l_1 \xrightarrow{a_1,[t'_1]_{\mathcal{J}}} \cdots \xrightarrow{a_{n-1},[t'_{n-1}]_{\mathcal{J}}} l_n$ the *$\mathcal{J}$-cylindrification* of the timed path $\pi = l_0 \xrightarrow{a_0,t'_0} l_1 \xrightarrow{a_1,t'_1} \cdots \xrightarrow{a_{n-1},t'_{n-1}} l_n$.

For a given finite partition $\mathcal{J}$ of the interval $[0, T]$, we call a set X of paths *$\mathcal{J}$-cylindrical* if it can be defined by $X = \{\pi' \in \mathit{Paths} \mid [\pi]_{\mathcal{J}} = [\pi']_{\mathcal{J}}\}$ for some concrete path π (that is if it is the set of timed paths with the same $\mathcal{J}$-cylindrification as π), and we call a finite partition $\mathcal{J}'$ of $[0, T]$ a *refinement* of $\mathcal{J}$ if every interval in $\mathcal{J}$ is the union of intervals in $\mathcal{J}'$.

We call a TH scheduler *$\mathcal{J}$-cylindrical* if its decisions depend only on the *cylindrification* $[\pi]_{\mathcal{J}}$ and $[t]_{\mathcal{J}}$ of π and t, respectively, and *cylindrical* if it is $\mathcal{J}$-cylindrical for some finite partition $\mathcal{J}$ of the interval $[0, T]$.

Cylindrical Sets and Primitive Probability Space. For an $\mathcal{J}'$-cylindrical scheduler $\mathcal{S}$, an $\mathcal{J}''$-cylindrical set of finite timed paths $[\pi]_{\mathcal{J}''}$, with π being just a representative $\pi = l_0 \xrightarrow{a_0,t_0} \ldots \xrightarrow{a_{n-1},t_{n-1}} l_n$, and a finite partition $\mathcal{J}$ that is a refinement of $\mathcal{J}'$ and $\mathcal{J}''$, the likelihood that a complete path is from this cylindrical set is easy to define: within each interval of $\mathcal{J}$, the likelihood that a CTMG $\mathcal{M}$ with scheduler $\mathcal{S}$ behaves in accordance with the $\mathcal{J}$-cylindrical set can—assuming compliance in all previous intervals—be checked using the same techniques as the ones used for finite Markov chains.

The likelihood that a system run (a complete timed path) is in the $\mathcal{J}$-cylindrical set of timed paths for $\mathcal{S}$ is the product $\nu(l_0) \prod_{I \in \mathcal{J}} p_I$, where P_I is the the probability to comply with the individual segments of $\mathcal{J}$. The probability p_{I_i} to comply with the i-th segment I_i of the partition $\mathcal{J}$ is the product of three multiplicands ($p_{I_i} = p_1^{I_i} \cdot p_2^{I_i} \cdot p_3^{I_i}$):

1. the probability $p_1^{I_i}$ that the actions are *chosen* in accordance with the $\mathcal{J}$-cylindrical scheduler (which is either 0 or 1 for deterministic schedulers),
2. the probability $p_2^{I_i}$ that the transitions are *taken* in accordance with the $\mathcal{J}$-cylindrical set of timed paths, provided the respective actions are chosen, which is simply the product over the individual probabilities $\mathbf{P}(l_j, a_j, l_{j+1})$ of the transitions that happen in I_i, and
3. the probability $p_3^{I_i}$ that the *right number of steps* is made in this subsequence of the $\mathcal{J}$-cylindrical set of timed paths.
   That is, $p_3^{I_i}$ is 0 if the last location is a discrete location, as the system would leave this location at the same point in time in which it was entered.

Otherwise, for a continuous location it is the difference $p_3^{I_i} = q_{\geq n} - q_{\geq n+1}$ between the likelihood that at least $n \geq 0$ transitions are made ($q_{\geq n}$), and the likelihood that at least $n+1$ transitions are made ($q_{\geq n+1}$) in the relevant subsequence of the timed path (we count only those transitions starting in continuous locations, and n is the correct number of transitions for I_i).
The likelihood to get a path of length $\geq n$ is

$$q_{\geq n} = \int_{(\tau_0,\ldots,\tau_{n-1})\in\Phi_{n,i}} \prod_{k=0}^{n-1} \lambda_k e^{-\lambda_k \tau_k} d\tau_k$$

for $\Phi_{n,i} = \{(\tau_0,\ldots,\tau_{n-1}) \in [0,T]^n \mid \sum_{j=0}^{n-1} \tau_j \leq t_i - t_{i-1}\}$ for $n > 0$, and 1 for $n = 0$. Here, t_i and t_{i-1} are the upper and lower endpoints of the interval I_i and the rates λ_k are defined as follows.
Let $\overline{l}_0, \overline{l}_1, \ldots, \overline{l}_n$ be the n continuous locations in the required order of appearance (note that n might be 0, and that the same location can occur multiple times), then for deterministic schedulers the transition rates are $\lambda_i = \mathbf{R}(\overline{l}_i, \mathcal{S}(\overline{l}_i, I_i), L)$. For a randomised scheduler $\mathcal{S}$, the transition rates are the expected transition rate $\lambda_i = \sum_{a\in Act(\overline{l}_i)} \mathcal{S}(\overline{l}_i, I_i)(a) \cdot \mathbf{R}(\overline{l}_i, a, L)$. The definition for paths of length $n+1$ runs accordingly.

Based on this, we define a straight forward *primitive probability measure* that covers the complement and finite unions of cylindrical sets of paths. Note that this does not define a *probability space* as countable unions are not covered yet.

2.4 Probability Space for Cylindrical Schedulers

Probability spaces of CTMDPs are sometimes defined in the tradition of the work of Wolovick and Johr [25], where the probability space is constructed in the following two steps: first, one defines a simple Borel measure on paths (comparable to a Borel-extension on our primitive measure from above), and second, one defines the probability space for an arbitrary (but fixed) Borel-measurable scheduler by defining the probability of a step and then using the Ionescu-Tulcea Theorem [1, Theorem 2.7.2] to obtain the unique extension to a probability space on infinite paths.

In this article, we take a step back and do the leg-work of defining a probability space directly through the simpler and more primitive completion underlying this closure: we complete the given primitive probability space using the standard completion through Cauchy sequences as known from the completion of Riemann integrable to Lebesgue integrable functions.

We do this in the following way: building on the primitive probability space defined in Section 2.3, we complete the space of cylindrical set of finite timed paths in this subsection for a *given* cylindrical scheduler, extending the primitive probability measure over complements and finite unions of cylindrical sets of paths to a probability space (covering also countable unions). Based on this definition, we define the time-bounded reachability probability for this simple class of schedulers in Section 2.5. Finally, we apply a second layer of completion

on the class of schedulers, for which we need to define a difference measure (or metric) *on* schedulers in Section 2.6.

Probability space for a given cylindrical scheduler. To obtain a suitable probability space for a given cylindrical scheduler, we complete the space of finite unions of cylindrical sets of timed paths to Cauchy sequences of finite unions of cylindrical sets of timed paths. In order to use Cauchy sequences, we need the notion of a difference measure (or metric). We define this difference measure between two sets of timed paths (each a finite disjoint union of cylindrical sets) as the probability of the symmetrical difference of the two sets. The symmetrical difference can obviously be represented as a finite disjoint union of cylindrical sets of timed paths, and thus we can use the primitive probability space we defined in Section 2.3.

Based on this definition, we define the usual equivalence class on the resulting Cauchy sequences: two Cauchy sequences $s = P_1, P_2, P_3, \ldots$ and $s' = P'_1, P'_2, P'_3, \ldots$ are equivalent if, and only if, $\lim_{n\to\infty} |P_n - P'_n| = 0$. This defines an equivalence class on Cauchy sequences, which we use to complete the space of cylindrical sets of timed paths to a space of equivalence classes of Cauchy sequences in the usual way. Thus, the we can use the quotient class of s instead of s. In particular, it trivially holds that the limit $\lim_{n\to\infty} |P_n|$ is independent from the chosen representative.

In the following we will use $Pr_S(\cdot)$ to denote the probability measure of the resulting probability space on measurable sets of timed paths. Where necessary to distinguish, we will refer to the corresponding CTMG $\mathcal{M}$ by using an additional parameter, $Pr_S^{\mathcal{M}}(\cdot)$.

2.5 Time-Bounded Reachability Probability

In this subsection, we consider the *time-bounded reachability probability* problem first for the primitive cylindrical schedulers: given a CTMG $\mathcal{M}$, a goal region $G \subseteq L$, and a time bound $0 \leq t_{\max} \leq T$,[4] we are interested in the set of paths $reach_{\mathcal{M}}(G, t_{\max})$ that reach a location in the goal region within time $t_{\max}$:

$$reach_{\mathcal{M}}(G, t_{\max}) = \Big\{ \pi \in Paths \mid \pi = l_0 \xrightarrow{a_0, t_0} l_1 \ldots l_n,\ \exists i < n.\, l_i \in G \wedge t_{i-1} \leq t_{\max} \Big\}.$$

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

Given a scheduler $\mathcal{S}$, we define $Pr_{\mathcal{S}}^{G}(\pi, t)$ to be the probability under this scheduler of visiting the goal region G within time $t_{\max}$, assuming we start with path π and that $t_{\max} - t$ time units have passed already (or, likewise,

[4] The upper time bound of the probability space T and of the reachability property $t_{\max}$ could be different. Nevertheless this is not a restriction, since we can choose T freely.

that t time units are left). That is, $Pr^G_{\mathcal{S}}(\pi, t)$ is the *conditional* probability $Pr_{\mathcal{S}}\left(reach_{\mathcal{M}}(G, t_{\max}) \mid \pi \text{ is prefix}\right)$. Similarly, for a location $l \in L$, we define

$$Pr^G_{\mathcal{S}}(l, t) = Pr_{\mathcal{S}}\left(reach_{\mathcal{M}}(G, t_{\max}) \mid \{\pi \in Paths \mid \ \text{last}(\pi) = l,\ t_{\text{len}(\pi)-1} < t\}\right).$$

As usual, the supremum of the time-bounded reachability probability over a particular scheduler class is also 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.

2.6 Completing the space of schedulers

While we have established a probability space for a given cylindrical scheduler, the class of cylindrical schedulers is not particularly strong, and, like in the recent literature on CTMDPs, we would like to prove our results for a wider class of schedulers. In order to exploit the technique of completion (as used above), we have to create a suitable metric space on cylindrical schedulers. For convenience we base the required difference measure on the time-bounded reachability probability.

A metric space *on* cylindrical schedulers. For *deterministic* schedulers D and E for a CTMG $\mathcal{M}$, we define a difference scheduler $\delta_{\{D,E\}}$ that uses the actions of D and E on every history, on which they coincide, and a fresh action a^* if D chose some action a_D and E chose an action $a_E \neq a_D$ upon this history. The new action a^* leads to a fresh continuous location g with rate $\overline{\lambda}$, where $\overline{\lambda} = \max_{l \in L, a \in Act(l)} \mathbf{R}(l, a, L)$ denotes the maximal transition rate of $\mathcal{M}$. We use $\mathcal{M}'$ to denote the adjusted CTMG used for the difference measure.

For all continuous locations l, we fix

- $\mathbf{R}(l, a^*, g) = \overline{\lambda}$ and
- $\mathbf{R}(l, a^*, l') = 0$ for all locations $l' \neq g$

for the new action, and maintain the entries to the rate matrix for the old actions. Location g has only one enabled action, say a^*, with $\mathbf{R}(g, a^*, g) = \overline{\lambda}$ and 0 for all other locations. For discrete locations, we would fix $\mathbf{P}(l, a^*, g) = 1$ (and $\mathbf{P}(l, a^*, l') = 0$ for $l' \neq g$) for the new action, and maintain the entries for the old actions in $\mathbf{P}$.

The distance d between two schedulers is then defined as the probability of $reach_{\mathcal{M}'}(\{g\}, T)$, where T is the same as for $\mathcal{M}$, using the *cylindrical* scheduler $\delta_{\{D,E\}}$.

This distance function defines a pseudometric on the cylindrical schedulers; symmetry and non-negativity obviously hold. For the triangle inequation $d(D, F) \leq d(D, E) + d(E, F)$ consider the scheduler $\delta_{\{D,E,F\}}$ that, similarly to the construction above, picks the action a^* unless all three schedulers agree. Clearly, it holds $d(D, F) \leq Pr_{\delta_{\{D,E,F\}}}(reach_{\mathcal{M}'}(\{g\}, T))$ and we can see that $Pr_{\delta_{\{D,E,F\}}}(reach_{\mathcal{M}'}(\{g\}, T)) \leq d(D, E) + d(E, F)$ as, whenever D deviates from F, scheduler E has to deviate from at least one of them.

Again, we have to resort to a carrier set of quotient classes of schedulers with distance 0 in order to satisfy $d(x,y) = 0 \Leftrightarrow x = y$ and thus to obtain a metric. (One can think of different actions on unreachable paths.)

Randomised schedulers. If the schedulers D and E from above are randomised, then we would define the distribution chosen by $\delta_{\{D,E\}}$ in line with the definition from above.

For every path ending in some location l where D chooses an action a with probability p_D^a and E chooses a with probability p_E^a then $\delta_{\{D,E\}}$ chooses a with probability $\min\{p_D^a, p_E^a\}$. ($\delta_{\{D,E,F\}}$ would choose a with probability $\min\{p_D^a, p_E^a, p_F^a\}$.) Similarly to the deterministic case, we assign the remaining probability mass to the fresh action a^* leading to the newly created location g. For continuous locations l, we set the rate to be $\mathbf{R}(l, a^*, g) = \overline{\lambda}$ (and to 0 otherwise). The new difference between two randomised schedulers is the probability to reach g within the time interval $[0, T]$ when using $\delta_{\{D,E\}}$.

Having constructed a difference measure (pseudometric), all arguments from above extend to the case of randomised schedulers.

Completion. We have defined a probability space for every cylindrical scheduler and a metric on the class of cylindrical schedulers. This allows us to complete the class of schedulers by defining a probability space for every scheduler that is a limit point of a Cauchy sequence of cylindrical schedulers.

It is easy to see that, for a CTMG $\mathcal{M}$ and schedulers D, E for $\mathcal{M}$, a set of paths $\Pi \subseteq \mathit{Paths}$ that is measurable for these schedulers must satisfy

$$\left| Pr_D(\Pi) - Pr_E(\Pi) \right| \;\leq\; d(D,E) \;.$$

Consequently, the limit probability of Π for a Cauchy sequence of cylindrical schedulers is well defined and independent of the representative.

We refer to the schedulers in this space as *measurable schedulers*. They contain the equally named class defined in [25, 15], but apply also for non-uniform CTMGs. We will use $Pr_{\mathcal{S}}(\cdot)$ to denote the measure of the probability space associated with a scheduler $\mathcal{S}$ from this much larger class.

Time-bounded reachability probabilities. Note that the definition of the metric implies that we can use the limit of the time-bounded reachability probabilities of a representative (that is, a Cauchy sequence of cylindrical schedulers) to define the time-bounded reachability probability of its quotient.

2.7 Why not via Borel σ-algebras?

As it is more common in the community to construct the probability space via Borel σ-algebras on paths, we want to give our motivation to switch to the definition introduced above.

In our opinion, the probability space should be the servant of the application, and not the other way round. (Although this may well be different in a

purely mathematical paper.) Thus, our starting point are the 'implementable' schedulers. But what does make a scheduler implementable?

We believe that cylindrical schedulers are a safe upper bound for what can be implemented. It is likely that we will face discrete-time systems in practice—as time cannot be measured with infinite precision—for which the continuous-time systems are abstractions, and an implementable scheduler has to take this into account[5].

If we consider a reasonable metric space of schedulers, such schedulers should be dense in it for the simple reason that every measure that would suggest the existence of a better scheduler would use a scheduler that cannot even be approximated by schedulers of this simple class, and hence cannot be approximated by implementable schedulers. In our view, such a measure should be considered with suspicion, because it has lost its relation to the real world and would therefore no longer be a suitable model. From this point of view, it is the most natural choice to work with a completion of the room of cylindrical schedulers.

The question of whether differential equations and derivations are justified for the resulting schedulers is, of course, of paramount importance. Note, however, that we only use these concepts for timed positional schedulers, which are fully described by the functions that map locations and points of time to actions or distributions over actions. But if these functions are Lebesgue measurable, then we can use the normal Lebesgue integral and use the standard theory of integration and derivation.

Finally, the main result of this article is that the class of cylindrical schedulers—the building blocks of our probability spaces—is sufficient for optimising time-bounded reachability probabilities, and thus the precise way how to complete the class of schedulers is not of importance.

3 Optimal Scheduling in CTMDPs

In this section we demonstrate the existence of optimal schedulers in CTMDPs. For this, we first develop upper and lower bounds ($f_{\max}, f_{\min} \in L \times \mathbb{R}_{\geq 0} \to [0,1]$) for the time-bounded reachability of any THR scheduler (Lemma 1) and then show that these bounds are actually taken by some TPD schedulers (Theorems 1 & 2). First, however, let us develop an intuition how the reachability probabilities in a CTMG evolve over time.

In the case of a fixed deterministic positional scheduler $\mathcal{S}$ for a given CTMG, we know that the reachability probabilities $Pr^G_{\mathcal{S}}(\cdot,\cdot)$ comply with the Kolmogorov backward equation for CTMCs [23]:

$$\dot{Pr}^G_{\mathcal{S}}(\cdot,t) = Q \cdot Pr^G_{\mathcal{S}}(\cdot,t) \ ,$$

[5] This view is in the tradition of the elegant and intuitive approach used by Bellman [5], who views continuous-time Markov games with bounded safety or reachability objectives as limits of sequences of discrete-time Markov games.

where $Pr^G_{\mathcal{S}}(\cdot, t)$ is the vector of probabilities $Pr^G_{\mathcal{S}}(l, t)$ for all locations $l \in L$, and matrix Q is defined by $Q[l, l'] = \mathbf{R}(l, \mathcal{S}(l), l')$ if $l \neq l'$ and $Q[l, l] = -\mathbf{R}(l, \mathcal{S}(l), L) + \mathbf{R}(l, \mathcal{S}(l), l)$.

We can interpret this equation location-wise. Thus, for deterministic positional schedulers $\mathcal{S}$, this immediately gives us the following system of equations for every location $l \in L$:

$$-\dot{Pr}^G_{\mathcal{S}}(l, t) = \sum_{l' \in L} \mathbf{R}\big(l, \mathcal{S}(l), l'\big) \cdot \big(Pr^G_{\mathcal{S}}(l', t) - Pr^G_{\mathcal{S}}(l, t)\big) \ .$$

This naturally extends to cylindrical TPD schedulers, as we can define the functions for the intervals recursively by interpreting each interval as an independent CTMC. Thus, the equation extends to TPD schedulers:

$$-\dot{Pr}^G_{\mathcal{S}}(l, t) = \sum_{l' \in L} \mathbf{R}\big(l, \mathcal{S}(l, t), l'\big) \cdot \big(Pr^G_{\mathcal{S}}(l', t) - Pr^G_{\mathcal{S}}(l, t)\big) \ ,$$

and even to TPR schedulers with small modifications:

$$-\dot{Pr}^G_{\mathcal{S}}(l, t) = \sum_{a \in Act(l)} \mathcal{S}(l, t)(a) \sum_{l' \in L} \mathbf{R}\big(l, a, l'\big) \cdot \big(Pr^G_{\mathcal{S}}(l', t) - Pr^G_{\mathcal{S}}(l, t)\big) \ .$$

Additionally, we know that, if time runs out, at time $t_{\max}$, or when we reach the goal region it obviously holds:

- $Pr^G_{\mathcal{S}}(l, t) = 1$ for all goal locations $l \in G$ and all $t \leq t_{\max}$,
- $Pr^G_{\mathcal{S}}(l, t) = 0$ for all other locations $l \notin G$ at time $t = t_{\max}$, and
- $Pr^G_{\mathcal{S}}(l, t) = 0$ for all locations $l \in L$ and for all $t > t_{\max}$.

As discussed in the literature, this enables us already to effectively approximate the time-bounded reachability for CMTCs (and also for CTMGs under (late) TPD schedulers). As we are interested in the optimal time-bounded reachability, we have to maximise/minimise over the available choices. It is not surprising—especially considering the way we define our schedulers—that pointwise optimisation leads to global optimality:

$$-\dot{f}_{\mathsf{opt}}(l, t) = \underset{a \in Act(l)}{\mathsf{opt}} \sum_{l' \in L} \mathbf{R}(l, a, l') \cdot (f_{\mathsf{opt}}(l', t) - f_{\mathsf{opt}}(l, t)) \ \text{ for } t \in [0, t_{\max}] \ ,$$

where $\mathsf{opt} \in \{\min, \max\}$. This intuitive result is not hard to prove[6]. In fact, f_{opt} dominates (is dominated by, respectively) not only TPD schedulers, but the full class of THR schedulers. The full proof is moved to Appendix A.2.

[6] The systems of non-linear ordinary differential equations used in this article are all quite obvious, and the challenge is to prove that they can be *taken* and not merely approximated. An approximative argument for these ODE's goes back to Bellman [5], but he uses a less powerful set of schedulers, and only proves that $f_{\max}$ and $f_{\min}$ can be approximated from below and above, respectively, claiming that the other direction is obvious.

Lemma 1. *For a CTMG $\mathcal{M}$ with only continuous locations the time-bounded reachability probability of any measurable THR scheduler is dominated by the function $f_{\max}$ and dominates the function $f_{\min}$.*

Proof Idea: To prove this claim for $f_{\max}$, assume that there is a scheduler that provides a strictly better time-bounded reachability probability $Pr^{\mathcal{M}}_{\mathcal{S}}(l,t) > f_{\max}(l,t)$ for some location $l \in L$ and time $t \in [0, t_{\max}]$ (in particular for $t = 0$), and hence improves over $f_{\max}(l,t)$ at this position by at least 3ε for some $\varepsilon > 0$.

$\mathcal{S}$ is a Cauchy sequence of cylindrical schedulers. Therefore we can sacrifice one ε and get an ε-close cylindrical scheduler from this sequence, which is still at least 2ε better than $f_{\max}$ at position (l,t).

As the measure for this cylindrical scheduler is a Cauchy sequence of measures for sequences with a bounded number of discrete transitions, we can sacrifice another ε to sharpen the requirement for the scheduler to reach the goal region in time *and* with at most n_ε steps for an appropriate bound $n_\varepsilon \in \mathbb{N}$, still maintaining an ε advantage over $f_{\max}$. Hence, we can compare with a finite structure, and use an inductive argument to show for paths π of shrinking length that end in any location $l' \in L$ that $f_{\max}(l',t) \leq Pr^{\mathcal{M}}_{\mathcal{S}}(\pi,t)$ holds. □

It is not very surprising that the optimal strategy is to always/point wise choose the optimising actions. The challenge is to prove that a *measurable* point wise optimal scheduler exists.

Theorem 1. *For every CTMDP there is a measurable TPD scheduler $\mathcal{S}$ that maximises the time-bounded reachability probability in the class of measurable THR schedulers.*

Proof. We construct a *measurable* scheduler S that chooses, for all $t \in [0, t_{\max}]$, an action a that maximises $\sum_{l' \in L} \mathbf{R}\big(l, \mathcal{S}(l,t), l'\big) \cdot \big(Pr^{\mathcal{M}}_{\mathcal{S}}(l',t) - Pr^{\mathcal{M}}_{\mathcal{S}}(l,t)\big)$. By Lemma 1, this guarantees that $\sum_{l \in L} \nu(l) Pr^{\mathcal{M}}_{\mathcal{S}}(l,0) = \sum_{l \in L} \nu(l) f_{\max}(l,0) = \sup_{\mathcal{S} \in THR} Pr^{\mathcal{M}}_{\mathcal{S}}(t_{\max})$ holds.

For positions outside of $[0, t_{\max}]$, the behaviour of the scheduler does not matter. $\mathcal{S}(l,t)$ can therefore be fixed to any constant decision for all $t \notin [0, t_{\max}]$.

In order to fix optimal decisions for the interval $[0, t_{\max}]$ in a location l, which we fix for the rest of the proof, we start with fixing an arbitrary order $\succ$ on the actions in $Act(l)$ and introduce, for each point in time $t \in [0, t_{\max}]$, an additional order $\succcurlyeq_t$ on the actions determined by the value of $\sum_{l' \in L} \mathbf{R}(l,a,l') \cdot \big(f_{\max}(l',t) - f_{\max}(l,t)\big)$, using $\succ$ as a tie-breaker.

We now define the following sets for every action a:

- $M_a = \{t \in [0, t_{\max}] \mid a \text{ is maximal w.r.t. } \succcurlyeq_t\}$ is the set of points t in the time interval $[0, t_{\max}]$, for which a is maximal with respect to the order $\succcurlyeq_t$.
- $C_a = \{t \in [0, t_{\max}] \mid \forall \delta > 0 \exists t' \in M_a.\ |t - t'| < \delta\}$ is the closure of M_a, and
- $D_a = C_a \setminus \bigcup_{b \succ a} C_b$ as the set of points in the time interval $[0, t_{\max}]$, action a is $\succcurlyeq_t$-better than all other actions and there is no $\succ$-better action with equal quality.

To complete the proof, we have to show that the scheduler $\mathcal{S}$, which chooses a for all $t \in D_a$, makes only decisions that maximise the gain, that is $\sum_{l' \in L} \mathbf{R}(l,a,l') \cdot \big(f_{\max}(l',t) - f_{\max}(l,t)\big)$ (and hence that $f_{\max} = Pr^{\mathcal{M}}_{\mathcal{S}}$ holds), and we have to show that the resulting scheduler is measurable. We start by showing that a is an optimal decision at all points in time contained in D_a.

First, the decision a is optimal, that is, it maximises $\sum_{l' \in L} \mathbf{R}(l,a,l') \cdot \big(f_{\max}(l',t) - f_{\max}(l,t)\big)$, in M_a by definition.

The fact that a is also optimal in the larger set C_a is then a trivial consequence of the continuity of $f_{\max}$. $D_a \subseteq C_a$ then implies that a is an optimal decision in all points in time contained in D_a.

The next relevant (and trivial) fact is that the M_a's partition $[0, t_{\max}]$ by their definition. Consequently, the C_a's cover $[0, t_{\max}]$ ($C_a \supseteq D_a$), and the D_a's again partition $[0, t_{\max}]$.

Finally, we show that the D_a's are (Lebesgue) measurable: they inherit this property from the C_a's, which are measurable because they are closed subsets of $[0, t_{\max}]$. As each D_a is constructed by a finite number of negations and intersections from these C_a's, the D_a's are measurable as well.

Our construction therefore provides us with a *measurable* scheduler, which is optimal, deterministic, and timed positional. □

By simply replacing maximisation by minimisation, sup by inf, and max by min, we can rewrite the proof to yield a similar theorem for the minimisation of time-bounded reachability, or, likewise, for the maximisation of time-bounded safety.

Theorem 2. *For a CTMDP, there is a measurable TPD scheduler $\mathcal{S}$ optimal for minimum time-bounded reachability in the class of measurable THR scheduler.*

4 Finite Optimal Control

In this section we show that, once the existence of an optimal scheduler is established, we can refine this result to the existence of a *cylindrical* optimal TPD scheduler, that is, a scheduler that switches *finitely* many times between different positional strategies. This is as close as we can hope to get to implementability as optimal points for policy switching are—like in the example from Figure 1—almost inevitably irrational. From now on we call these points *switching points*.

Our proof of Theorem 1 makes a purely topological existence claim, and therefore does not imply that a finite number of switching points suffices. In principle, this could mean that the required switching points have one or more limit points, and an unbounded number of switches is required to optimise time-bounded reachability. Figure 2 shows the derivations of the reachability probabilities for a positional scheduler (black) and a potential other scheduler. The intersections of the two curves have a limit point.

To exclude such limit points, and hence to prove the existence of an optimal scheduler with a finite number of switching points, we re-visit the differential equations that define the reachability probability, but this time to answer a

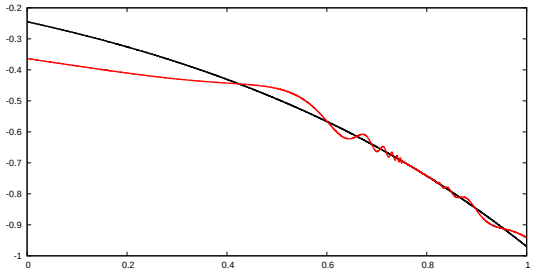

Fig. 2. Theorem 1 would not exclude that the intersection points of derivatives (the 'loss') of two positional schedulers have a limit point.

different question: can we, for every point in time t, find positional schedulers that are *locally optimal* on an ε-environment of t? If yes, then we could exploit the compactness of $[0, t_{\max}]$ to get a cover of finitely many of these open intervals and we could potentially construct a (globally) optimal scheduler using the locally optimal scheduler associated to these intervals. However, while this is possible for most points, this is *not* necessarily the case at our switching points.

In the remainder of this section, we therefore show a slightly weaker property: for every point $t \in [0, t_{\max}]$ in time, there is a positional scheduler that is locally optimal in a left ε-environment of t (that is, in a set $(t-\varepsilon, t] \cap [0, t_{\max}]$), and one that is locally optimal in a right ε-environment of t. Hence, we get an open cover of $[0, t_{\max}]$ by intervals with locally optimal schedulers that have at most one switching point. Thus, we obtain a globally optimal scheduler with a finite number of switching points.

Lemma 2. *For every CTMDP and every point in time t there is a constant scheduler $\mathcal{S}_l \in L \to Act$ ($\mathcal{S}_r$, respectively) such that it adheres to the same system of differential equations as $f_{\max}$ ($f_{\min}$, respectively) on a left (right, respectively) ε-environment of t, when using $f_{\max}(t)$ ($f_{\min}(t)$, respectively) as support point.*

Proof. We have seen that the true optimal reachability probability ($f_{\max}$) is defined by a system of differential equations. In this proof we consider the effect of starting with the 'optimal' values for a time $t \in [0, t_{\max}]$ and fix a *positional* scheduler. We then prove that there is a left-optimal scheduler among them that satisfies the same system of differential equations in a left ε-environment and a right-optimal scheduler that satisfies the same system of differential equations in a right ε-environment. Thus, there is a scheduler with at most one switching point that is locally optimal with respect to the full class of schedulers.

Given a CTMDP $\mathcal{M}$, we consider the differential equations that describe the development near the support point $f_{\max}(l, t)$ for each location l under a positional strategy D:

$$-\dot{Pr}_l^D(\tau) = \sum_{l' \in L} \mathbf{R}(l, a_l, l') \cdot \left(Pr_{l'}^D(\tau) - Pr_l^D(\tau) \right) ,$$

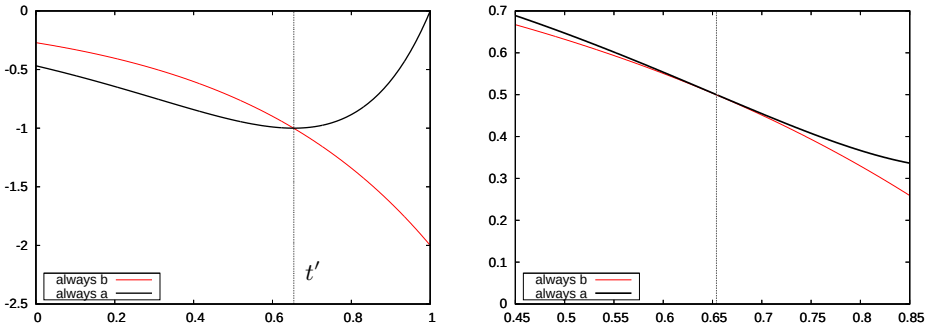

Fig. 3. The gain functions of the competing stationary strategies of Figure 1. To the left: developed *gains* $\dot{Pr}_A^{\mathcal{S}_a/\mathcal{S}_b}(t)$ from $t = t_{\max} = 1$ in order to find the only switching point $t' = t_{\max} - \frac{1}{2}\log(2)$. To the right: Developed values $Pr_A^{\mathcal{S}_a/\mathcal{S}_b}(t)$ (not gains) from t' in both directions, to show that action a is better for all $t < t'$ whereas b is better for all $t > t'$. This construction is also used to determine the existence of ϵ-environments with a stable strategy around every point t.

where a_l is the action chosen at l by D (see Figure 3 for an example).

Different to the development of the true probability, the development of these linear differential equations provides us with smooth functions. This provides us with more powerful techniques when comparing two locally positional strategies: each deterministic scheduler defines a system $\dot{y} = Ay$ of ordinary homogeneous linear differential equations with constant coefficients.

As a result, the solutions $Pr_l^D(\tau)$ of these differential equations—and hence their differences $Pr_l^{D'}(\tau) - Pr_l^D(\tau)$—can be written as finite sums $\sum_{i=1}^{n} P_i(\tau)e^{\lambda_i \tau}$, where P_i is a polynomial and the λ_i may be complex. Consequently, these functions are holomorphic.

Using the identity theorem for holomorphic functions, t can only be a limit point of the set of 0 points of $Pr_l^{D'}(\tau) - Pr_l^D(\tau)$ if $Pr_l^{D'}(\tau)$ and $Pr_l^D(\tau)$ are identical on an ε-environment of t. The same applies to their derivations: $\dot{Pr}_l^{D'}(\tau) - \dot{Pr}_l^D(\tau)$ either has no limit point in t, or $\dot{Pr}_l^{D'}(\tau)$ and $\dot{Pr}_l^D(\tau)$ are identical on an ε-environment of t.

For the remainder of the proof, we fix, for a given time t, a sufficiently small $\varepsilon > 0$ such that, for each pair of positional schedulers D and D' and every location $l \in L$, $\dot{Pr}_l^{D'}(\tau) - \dot{Pr}_l^D(\tau)$ is either < 0, $= 0$, or > 0 on the complete interval $L_\varepsilon^t = (t - \varepsilon, t) \cap [0, t_{\max}] \ni \tau$, and, possibly with different sign, for the complete interval $R_\varepsilon^t = (t, t + \varepsilon) \cap [0, t_{\max}] \ni \tau$.

We argue the case for the left ε-environment L_ε^t. In the '$>$' case for a location l, we say that D is *l-better* than D'. We call D *preferable* over D' if D' is not l-better than D for any location l, and *better* than D' if D is preferable over D' and l-better for some $l \in L$.

If D' is l-better than D in exactly a non-empty set $L_b \subset L$ of locations, then we can obviously use D' to construct a strategy D'' that is better than D by switching to the strategies of D' in exactly the locations L_b.

Since we choose our strategies from a finite domain—the deterministic positional schedulers—this can happen only finitely many times. Hence we can stepwise *strictly* improve a strategy, until we have constructed a strategy $D_{\max}$ preferable over all others.

By the definition of being preferable over all other strategies, $D_{\max}$ satisfies

$$-\dot{Pr}_l^{D_{\max}}(\tau) = \max_{a \in Act(l)} \sum_{l' \in L} \mathbf{R}(l, a, l') \cdot \left(Pr_{l'}^{D_{\max}}(\tau) - Pr_l^{D_{\max}}(\tau)\right)$$

for all $\tau \in L_\varepsilon^t$ and all $l \in L$ —it fulfils the same system of equations as $f_{\max}$.

We can use the same method for the right ε-environment R_ε^t and for the minimising strategies. □

Theorem 3. *For every CTMDP there is a cylindrical TPD scheduler $\mathcal{S}$ optimal for maximum time-bounded reachability in the class of measurable THR schedulers.*

Proof. We can use Lemma 2 to determine a scheduler *with at most one switching point* that fulfils the optimal differential equations in an ε-environment $(t-\varepsilon, t+\varepsilon) \cap [0, t_{\max}]$ of t, using $f_{\max}(t)$ as support point.

As this is possible for all points in $[0, t_{\max}]$, the sets $I_\varepsilon^t = L_\varepsilon^t \cup R_\varepsilon^t$ define a cover of open sets of the interval $[0, t_{\max}]$. Using compactness of $[0, t_{\max}]$, we infer a *finite* sub-cover of intervals with associated schedulers that have at most one switching point each and that each locally fulfil the differential equations of $f_{\max}$ on $[0, t_{\max}]$. As these open intervals have to overlap to form a cover, this establishes the existence of an optimal strategy with a finite number of switching points. □

The proof for the minimisation of the time-bounded reachability probability runs accordingly.

Theorem 4. *For every CTMDP there is a cylindrical TPD scheduler $\mathcal{S}$ optimal for minimal time-bounded reachability in the class of measurable THR scheduler.*

5 Discrete Locations

In this section, we treat the mildly more general case of single player CTMGs, which are traditional CTMDPs plus discrete locations. We reduce the problem of finding optimal measurable schedulers for CTMGs first to *simple CTMGs*, CTMGs whose discrete locations have no incoming transitions from continuous locations. (They hence can only occur initially at time 0.) The extension from CTMDPs to simple CTMGs is trivial.

Lemma 3. *For a simple single player CTMG with only a reachability (or only a safety) player, there is an optimal deterministic scheduler with finitely many switching points.*

Proof. By the definition of simple single player games, the likelihood of reaching the goal location from any continuous location and any point in time is independent of the discrete locations and their transitions. For continuous locations, we can therefore simply reuse the results from the Theorems 3 and 4.

We can only be in discrete locations at time 0, and for every continuous location l there is a fixed time-bounded reachability probability described by $f_{\mathsf{opt}}(l, 0)$. We can show that there is a timed positional (even a positional) deterministic optimal choice for the discrete locations at time $t = 0$ by induction over the maximal distance to continuous locations: if all successors have been evaluated, we can fix an optimal timed positional choice. We can therefore use discrete positions with maximal distance 1 as induction basis, and then apply an induction step from positions with distance $\leq n$ to positions with distance $n + 1$. □

Rebuilding a single player CTMG $\mathcal{M}$ to a *simple* single player CTMG $\mathcal{M}_s$ can be done in a straight forward manner; it suffices to pool all transitions taken between two continuous locations. (Here we use that there cannot be loops in discrete states.) To construct the resulting simple CTMG $\mathcal{M}_s$, we add new continuous locations for each possible time abstract path from continuous locations of the CTMG $\mathcal{M}$, and we add the respective actions: for continuous locations $l_c, l'_c \in L_c$ and discrete locations $l^d_1, \ldots, l^d_n \in L_d$ a timed path $l_c \xrightarrow{a_0,t} l^d_1 \xrightarrow{a_1,t} l^d_2 \cdots\ l^d_n \xrightarrow{a_n,t} l'_c$ translates to $l_c \xrightarrow{\mathbf{a},t} \underline{\xrightarrow{a_0} l^d_1 \xrightarrow{a_1} l^d_2 \cdots\ l^d_n \xrightarrow{a_n} l'_c}$, where the underlined part is a new continuous location. (For simplicity, we also translate a timed path $l_c \xrightarrow{a,t} l'_c$ to $l_c \xrightarrow{\mathbf{a},t} \underline{\xrightarrow{a} l'_c}$.)

The new *actions* of the resulting simple single player CTMG encode the sequences of actions of $\mathcal{M}$ that a scheduler could make in the current location plus in all possible sequences of discrete locations, until the next continuous location is reached. (Note that this set is finite, and that the scheduler makes all of these transitions at the same point of time.) If $\mathbf{a}$ encodes choices that depend only on the position (but not on this local history), $\mathbf{a}$ is called positional. For continuous locations, all old actions are deleted, and all new continuous locations that end in a location $l_c \in L_c$ get the same outgoing transitions as l_c. The rate matrix is chosen accordingly.

Adding the information about the path to locations allows us to reconstruct the timed history in the single player CTMG from a history in the constructed simple CTMG.

Theorem 5. *For a single player CTMG $\mathcal{M}$ with only a reachability (or only a safety) player, there is an optimal deterministic scheduler with finitely many switching points.*

Proof. First, every scheduler $\mathcal{S}$ for $\mathcal{M}$ can be naturally translated into a scheduler of $\mathcal{S}_s$ of $\mathcal{M}_s$, because every timed-path in $\mathcal{M}_s$ defines a timed-path in $\mathcal{M}$; the resulting time-bounded reachability probability coincides.

Let us consider a cylindrical optimal deterministic scheduler $\mathcal{S}_{\mathsf{opt}}$ for the simple Markov game, and the function f_{opt} defined by it. For the actions $\mathbf{a}$ $\mathcal{S}_{\mathsf{opt}}$ chooses, we can, for each interval in which $\mathcal{S}_{\mathsf{opt}}$, is positional, use an inductive argument similar to the one from the proof of Lemma 3 to show that we can choose a *positional* action $\mathbf{a}'$ instead. The resulting cylindrical deterministic scheduler $\mathcal{S}'_{\mathsf{opt}}$ defines the same f_{opt} (same differential equations).

Clearly, $f_{\mathsf{opt}}(l_c, t) = f_{\mathsf{opt}}(\underline{\ldots \xrightarrow{a} l_c}, t)$ holds. We use this observation to change $\mathcal{S}'_{\mathsf{opt}}$ to $\mathcal{S}''_{\mathsf{opt}}$ by choosing the action that $\mathcal{S}'_{\mathsf{opt}}$ chooses for l_c for all locations $\underline{\ldots \xrightarrow{a} l_c}$ and at each point of time. The resulting scheduler $\mathcal{S}''_{\mathsf{opt}}$ is still cylindrical and deterministic, and defines the same f_{opt} (same differential equations).

$\mathcal{S}''_{\mathsf{opt}}$ is also the mapping of a cylindrical optimal deterministic scheduler for $\mathcal{M}$. □

5.1 From Late to Early Scheduling

Our main motivation for introducing discrete transitions is that it provides framework that covers *both*, early schedulers (which have to fix an action when *entering* a location) and the late schedulers used in this article.

Late schedulers are naturally subsumed in our model, as the schedulers we assume are the more powerful late schedulers. To embed early schedulers as well, it suffices to use a simple translation: we 'split' every continuous location l_c into a fresh discrete location l_c^d, and one fresh continuous location l_c^a for each action $a \in Act(l_c)$ enabled in l_c.

Every incoming transition to l_c is re-routed to l_c^d, l_c^d has an outgoing transition a that surely leads to l_c^a ($\mathbf{P}(l_c^d, a, l_c^a) = 1$) for each action $a \in Act(l_c)$ enabled in l_c, and no other outgoing transition. In l_c^a, we have $Act(l_c^a) = \{a\}$, and the entries in $\mathbf{R}(l_c^a, a, l)$ are the entries taken from $\mathbf{R}(l_c, a, l)$ for discrete locations l, and re-routed to the respective l^d for continuous locations. Probability mass assigned to l_c is moved to l_c^d by the translation, and if l_c is a goal state, so are l_c^d and the l_c^a's.

In the initial distribution we replace every continuous location l_c by the newly created discrete distribution l_c^d.

Intuitively, every occurrence of $\xrightarrow{*,t} l_c \xrightarrow{a,t'}$ is replaced by $\xrightarrow{*,t} l_c^d \xrightarrow{a,t} l_c^a \xrightarrow{a,t'}$; l_c is the beginning of the path, $l_c \xrightarrow{a,t'}$ is replaced by $l_c^d \xrightarrow{a,t} l_c^a \xrightarrow{a,t'}$.

Obviously, there is a simple relation between optimal early schedulers of a CTMDP (or, indeed, CTMG), and optimal late schedulers in the translated CTMDP: recall that history-dependent scheduler classes do not improve reachability over timed positional schedulers. As a consequence, the existence of finite deterministic optimal control extends to early scheduling.

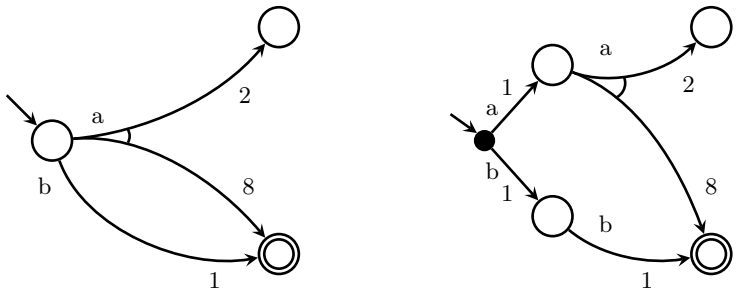

Fig. 4. An informal example, depicting the idea of the encoding of an early scheduling CTMG (left) in a late scheduling CTMG (right).

6 Continuous-Time Markov Games

In this section, we lift our results from single player to general continuous-time Markov games. In general continuous-time Markov games, we are faced with two players who have opposing objectives: a reachability player trying to maximise the time-bounded reachability probability, and a safety player trying to minimise it—we consider a 0-sum game.

Thus, all we need to do for lifting our results to games is to show that the quest for optimal strategies for single player games discussed in the previous section can be generalised to a quest for co-optimal strategies—that is, for Nash equilibria—in general games. To demonstrate this, it essentially suffices to show that it is not important whether we first fix the strategy for the reachability player and then the one for the safety player in a strategy refinement loop, or vice versa.

Let us first assume CTMGs without discrete locations.

Lemma 4. *Using the ε-environments I^t_ε from the proof of Theorem 3, we can construct a Nash equilibrium that provides co-optimal deterministic strategies for both players, such that the co-optimal strategies contain at most one strategy switch on I^t_ε.*

Proof. We describe the technique to find a constant co-optimal strategy on the right ε-environment $R^t_\varepsilon = (t, t+\varepsilon) \cap [0, t_{\max}]$ of t.

We write a constant strategy as $D = S + R$ that is composed of the actions chosen by the safety player on L_s, and the actions chosen by the reachability player on L_r. For this simple structure, we introduce a strategy improvement technique on the finite domain of deterministic choices for the respective player.

For a fixed strategy S of the safety player, we can define an optimal counter strategy $R(S)$ of the reachability player by applying the technique described in Theorem 3. (For equivalent strategies, we make an arbitrary but fixed choice.)

We call the resulting vector $(-\dot{Pr}^{\mathcal{M}}_{S+R(S)}(l, t+\frac{1}{2}\varepsilon) \mid l \in L)$ the *quality vector of S*. Now, we choose an arbitrary $\overline{S}$ for which this vector is minimal. (Note that there could, potentially, be multiple incomparable minimal elements.)

We now show that the following holds for $\overline{S}$ and all $\tau \in R^t_\varepsilon$:

$$-\dot{Pr}^{\mathcal{M}}_{\overline{S}+R(\overline{S})}(l,\tau) = \max_{a\in Act(l)} \sum_{l'\in L} \mathbf{R}(l,a,l') \cdot \big(Pr^{\mathcal{M}}_{\overline{S}+R(\overline{S})}(l',\tau) - Pr^{\mathcal{M}}_{\overline{S}+R(\overline{S})}(l,\tau)\big)$$

for all $l \in L_r$, and

$$-\dot{Pr}^{\mathcal{M}}_{\overline{S}+R(\overline{S})}(l,\tau) = \min_{a\in Act(l)} \sum_{l'\in L} \mathbf{R}(l,a,l') \cdot \big(Pr^{\mathcal{M}}_{\overline{S}+R(\overline{S})}(l',\tau) - Pr^{\mathcal{M}}_{\overline{S}+R(\overline{S})}(l,\tau)\big)$$

for all $l \in L_s$. (Note that the order between the derivation is maintained on the complete right ε-environment R^t_ε.)

The first of these claims is a trivial consequence from the proof of Theorem 3. (The result is, for example, the same if we had a single player CTMDP that, in the locations L_s of the safety player, has only one possible action: the one chosen by $\overline{S}$.)

Let us assume that the second claim does not hold. Then we choose a particular $l \in L_s$ where it is violated. Let us consider a slightly changed setting, in which the choices in l are restricted to two actions, the action a_1 chosen by $\overline{S}$, and the minimising action a_2. Among these two, one maximises, and one minimises

$$-\dot{Pr}^{\mathcal{M}}_{\overline{S}+R(\overline{S})}(l,\tau) = \min_{a\in\{a_1,a_2\}} \sum_{l'\in L} \mathbf{R}(l,a,l') \cdot \big(Pr^{\mathcal{M}}_{\overline{S}+R(\overline{S})}(l',\tau) - Pr^{\overline{S}+R(\overline{S})}(l,\tau)\big)\ .$$

Let us fix all other choices of S, and allow the reachability player to choose among a_1 and a_2 (we 'pass control' to the other player). As shown in Theorem 3, she will select an action that produces the well defined set of max equations for the resulting single player game. Hence, choosing a_1 and keeping all other choices from $R(\overline{S})$ is the optimal choice for the reachability player in this setting (as the max equations are satisfied, while they are dissatisfied for a_2).

Consequently, the quality vector for $\overline{S}$ is strictly greater than the one for the adjusted strategy. Assuming that choosing an arbitrary maximal element does not lead to a satisfaction of the min and max equations thus leads to a contradiction.

We can argue symmetrically for the left ε-environment. Note that the satisfaction of the min and max equations implies that it does not matter if we change the role of the safety and reachability player in our argumentation. □

This lemma can easily be extended to construct simple co-optimal strategies:

Theorem 6. *For CTMGs without discrete locations, there are cylindrical deterministic timed positional co-optimal strategies for the reachability and the safety player.*

Proof. First, Lemma 4 provides us with an open coverage of co-optimal strategies that switch at most once, and we can build a strategy that switches at most finitely many times from a finite sub-cover of the open space $[0, t_{\max}]$. This strategy is everywhere locally co-optimal, and forms a Nash equilibrium:

It is straight forward to cut the interval $[0, t_{\max}]$ into a finite set of subintervals $[0, t_0]$, $(t_0, t_1]$, $\ldots$, $(t_{n-1}, t_n]$ with $t_n = t_{\max}$, such that the strategy for the safety player is constant in all of these intervals. We can use the construction from Theorem 3 (note that the proof of Theorem 3 does not use that the differential equations are initialised to 0 or 1 at $t_{\max}$) to construct an optimal strategy for the reachability player: we can first solve the problem for the interval $[t_{n-1}, t_n]$, then for the interval $[t_{n-2}, t_{n-1}]$ using $f_{\mathsf{opt}}(l, t_{n-1})$ as initialisation, and so forth. A similar argument can be made for the other player.

This provides us with the same differential equations, namely:

$$-\dot{f}_{\mathsf{opt}}(l,t) = \max_{a \in Act(l)} \sum_{l' \in L} \mathbf{R}(l,a,l') \cdot (f_{\mathsf{opt}}(l',t) - f_{\mathsf{opt}}(l,t))$$

for $t \in [0, t_{\max}]$ and $l \in L_r$, and

$$-\dot{f}_{\mathsf{opt}}(l,t) = \min_{a \in Act(l)} \sum_{l' \in L} \mathbf{R}(l,a,l') \cdot (f_{\mathsf{opt}}(l',t) - f_{\mathsf{opt}}(l,t))$$

for $t \in [0, t_{\max}]$ and $l \in L_s$.

Note that all Nash equilibria need to satisfy these equations (with the exception of 0 sets, of course), because otherwise one of the players could improve her strategy. □

The extension of these results to the full class of CTMGs is straight forward: we would first reprove Theorem 5 in the style of the proof of Theorem 3 (which requires to establish the Theorem in the first place). The only extension is that we additionally get an equation $Pr^D_l(\tau) = \sum_{l' \in L} \mathbf{P}(l, a_l, l') \cdot Pr^D_{l'}(\tau)$ for every discrete location l. The details are moved to Appendix B.

Theorem 7. *For continuous-time Markov Games, there are cylindrical deterministic timed positional co-optimal strategies for the reachability and the safety player.*

As a small side result, these differential equations show us that we can, for each continuous location $l_c \in L_c$ and every action $a \in Act(l_c)$, add arbitrary values to $\mathbf{R}(l_c, a, l_c)$ without changing the bounded reachability probability for every pair of schedulers. (Only if we change $\mathbf{R}(l_c, a, l_c)$ to 0 will we have to make sure that a is not removed from $Act(l_c)$.) In particular, this implies that we can locally and globally uniformise a continuous-time Markov game if this eases its computational analysis. (Cf. [15] for the simpler case of CTMDPs.)

7 Variations

In this section, we discuss the impact of small changes in the setting, namely the impact of infinitely many states or actions, and the impact of introducing a non-absorbing goal region.

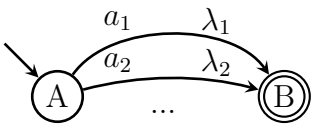

Fig. 5. An example CTMDP with infinitely many actions.

Infinitely Many States. If we allow for infinitely many states, optimal solutions may require infinitely many switching points. To see this, it suffices to use one copy of the CTMDP from Figure 1, but with rates i and $2i$ for the i-th copy, and assign an initial probability distribution that assigns a weight of 2^{-i} to the initial state A_i of the i-th copy. (If one prefers to consider only systems with bounded rates, one can choose rates $1+\frac{1}{i}$ and $2+\frac{2}{i}$.) The switching points are then different for every copy, and an optimal strategy has to select the correct switching point for every copy.

Infinitely Many Actions. If we allow for infinitely many actions, there is not even an optimal strategy if we restrict our focus to CTMDPs with two locations, an initial location and an absorbing goal location. For the CTMDP of Figure 5 with the natural numbers $\mathbb{N}$ as actions and rate $\lambda_i = 2 - \frac{1}{i}$ for the action $i \in \mathbb{N}$ if we have a reachability player and $\lambda_i = \frac{1}{i}$ if we have a safety player, every strategy $\mathcal{S}$ can be improved over by a strategy $\mathcal{S}'$ that always chooses the successor $i+1$ when of the action i chosen by $\mathcal{S}$.

Reachability at $t_{\max}$. If we drop the assumption that the goal region is absorbing, one might be interested in the marginally more general problem to be in the goal region at time t_{max} for the reachability player (safety player, respectively). For this generalisation, no substantial changes need to be made: it suffices to replace

$$f_{\mathsf{opt}}(l,t) = Pr^{\mathcal{M}}_{\mathcal{S}}(l,t) = 1 \qquad \text{for all goal locations } l \in G \text{ and all } t \leq t_{\max}$$

by

$$f_{\mathsf{opt}}(l,t_{\max}) = Pr^{\mathcal{M}}_{\mathcal{S}}(l,t_{\max}) = 1 \qquad \text{for all goal locations } l \in G.$$

(In order to be flexible with respect to this condition, the $-\dot{f}_{\mathsf{opt}}(l,t)$ are defined for goal locations as well. Note that, when all goal locations are absorbing, the value of $-\dot{f}_{\mathsf{opt}}(l,t)$ is 0 and $f_{\mathsf{opt}}(l,t)$ is 1 for all goal locations $l \in G$ and all $t \in [0, t_{\max}]$.)

References

1. Robert B. Ash and Catherine A. Doléans-Dade. *Probability & Measure Theory*. Elsevier Science, 2000.
2. Adnan Aziz, Kumud Sanwal, Vigyan Singhal, and Robert Brayton. Model-checking Continuous-Time Markov Chains. *Transactions on Computational Logic*, 1(1):162–170, 2000.
3. 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.
4. Christel Baier, Jost-Pieter Katoen, and Holger Hermanns. Approximate Symbolic Model Checking of Continuous-Time Markov Chains. In *Proceedings of CONCUR'99*, volume 1664 of *Lecture Notes in Computer Science*, pages 146–161, 1999.
5. Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.
6. Tomas Brazdil, Vojtech Forejt, Jan Krcal, Jan Kretinsky, and Antonin Kucera. Continuous-Time Stochastic Games with Time-Bounded Reachability. In *Proceedings of FSTTCS'09*, Leibniz International Proceedings in Informatics (LIPIcs), pages 61–72, 2009.
7. Peter Buchholz and Ingo Schulz. Numerical Analysis of Continuous Time Markov Decision Processes Over Finite Horizons. *Computers & Operations Research*, 38(3):651 – 659, 2011.
8. Eugene A. Feinberg. Continuous Time Discounted Jump Markov Decision Processes: A Discrete-Event Approach. *Mathematics of Operations Research*, 29(3):492–524, 2004.
9. Xianping Guo and Onésimo Hernández-Lerma. Zero-sum Games for Continuous-Time Markov Chains with Unbounded Transition and Average Payoff Rates. *Journal of Applied Probability*, 40(2):327–345, 2003.
10. Xianping Guo and Onésimo Hernández-Lerma. Zero-sum Continuous-Time Markov Games with Unbounded Transition and Discounted Payoff Rates. *Bernoulli*, 11(6):1009–1029, 2005.
11. Xianping Guo and Onésimo Hernández-Lerma. Zero-sum Games for Continuous-Time Jump Markov Processes in Polish Spaces: Discounted Payoffs. *Advances in Applied Probability*, 39(3):645–668, 2007.
12. Holger Hermanns. *Interactive Markov Chains and the Quest for Quantified Quality. LNCS 2428*. Springer-Verlag, 2002.
13. 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.
14. Bruce L. Miller. Finite State Continuous Time Markov Decision Processes with a Finite Planning Horizon. *SIAM Journal on Control*, 6(2):266–280, 1968.
15. 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.
16. Martin R. Neuhäußer and Lijun Zhang. Time-Bounded Reachability Probabilities in Continuous-Time Markov Decision Processes. In *QEST*, pages 209–218, 2010.
17. Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, April 1994.
18. Markus Rabe and Sven Schewe. Optimal Schedulers for Time-Bounded Reachability in CTMDPs. Reports of SFB/TR 14 AVACS 55, October 2009.

19. Markus Rabe and Sven Schewe. Finite Optimal Control for Time-Bounded Reachability in CTMDPs and Continuous-Time Markov Games. *CoRR*, abs/1004.4005, 2010.
20. Markus Rabe and Sven Schewe. Optimal Time-Abstract Schedulers for CTMDPs and Markov Games. In *Proceedings of QAPL*, pages 144–158, 2010.
21. Markus Rabe, Sven Schewe, and Lijun Zhang. Efficient Approximation of Optimal Control for Markov Games. *CoRR*, abs/1011.0397, 2010.
22. 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.
23. William J. Stewart. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1994.
24. Wayne Winston. Optimality of the Shortest Line Discipline. *Journal of Applied Probability*, 14(1):181–189, 1977.
25. 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.
26. Xianping Guo and Onésimo Hernández-Lerma. *Continuous-Time Markov Decision Processes*, volume 62 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, 2009.
27. Lijun Zhang, Holger Hermanns, Ernst M. Hahn, and Björn. Wachter. Time-bounded Model Checking of Infinite-state Continuous-Time Markov Chains. In *Proceedings of ACSD'08*, pages 98–107, 2008.
28. Lijun Zhang and Martin R. Neuhäußer. Model Checking Interactive Markov Chains. In *Proceedings of TACAS*, pages 53–68, 2010.

Appendix

A Optimal reachability probability

A.1 Differential Equations

The differential equations defining f_{opt} are simply the differential equations in place when a strategy is locally constant. This holds almost everywhere (everywhere but in a 0-set of positions) in case of the *cylindrical* schedulers that are the basic building blocks in the incomplete space that we have completed by considering Cauchy sequences of cylindrical schedulers.

Hence, for every cylindrical scheduler we can partition the interval $[0, t_{\max}]$ into a finite set of intervals I_0, I_1, I_2, ..., I_n as described in the preliminaries.

Within such an interval,

$$-\dot{Pr}_{\mathcal{S}}(\pi, t) = \sum_{l' \in L} \mathbf{R}(l, a, l') \cdot \Big(Pr_{\mathcal{S}}(\pi \xrightarrow{a,t} l', t) - Pr_{\mathcal{S}}(\pi, t) \Big) \qquad \text{for } t \in I_i$$

holds for deterministic schedulers, where π is a timed path that ends in l, a is the deterministic choice the scheduler makes in I_i on this history, and $\pi \xrightarrow{a,t} l'$ is its extension. For randomised schedulers it holds for $t \in I_i$,

$$-\dot{Pr}_{\mathcal{S}}(\pi, t) = \sum_{a \in Act} \mathcal{S}(\pi, t)(a) \sum_{l' \in L} \mathbf{R}(l, a, l') \cdot \Big(Pr_{\mathcal{S}}(\pi \xrightarrow{a,t} l', t) - Pr_{\mathcal{S}}(\pi, t) \Big) \ ,$$

where π is a timed path that ends in l, $\mathcal{S}(\pi, t)(a)$ is the likelihood that the cylindrical scheduler makes the decision a in I_i on this history, and $\pi \xrightarrow{a,t} l'$ is its extension.

To initialise the potentially infinite set of differential equations, we have the following initialisations:

- $Pr_{\mathcal{S}}(\pi, t) = 1$ holds for all timed histories π that contain (and hence end up in) locations $l \in G$ in the goal region and all $t \leq t_{\max}$,
- $Pr_{\mathcal{S}}(\pi, t_{\max}) = 0$ holds for all timed histories π that contain only non-goal locations $l \notin G$, and
- $Pr_{\mathcal{S}}(\pi, t) = 0$ holds for all locations $l \in L$ and all $t > t_{\max}$.

Additionally, we have to consider what happens at the intersection t_i of the fringes of I_i and I_{i+1} for $0 \leq i < n$. But obviously, we can simply first solve the differential equations for I_n, then use the values of $f(\pi, t_{n-1})$ as initialisations for the interval I_{n-1}, and so forth.

Remark: For timed positional deterministic schedulers we get

$$-\dot{Pr}_{\mathcal{S}}(l, t) = \sum_{l' \in L} \mathbf{R}(l, a, l') \cdot (Pr_{\mathcal{S}}(l', t) - Pr_{\mathcal{S}}(l, t)) \quad \text{for } t \in I_i, \text{ and}$$

$$-\dot{Pr}_{\mathcal{S}}(l, t) = \sum_{a \in Act} \mathcal{S}(l, t)(a) \sum_{l' \in L} \mathbf{R}(l, a, l') \cdot (Pr_{\mathcal{S}}(l', t) - f(l, t)) \quad \text{for } t \in I_i$$

for timed positional randomised schedulers. In both cases, the initialisation reads

- $Pr_{\mathcal{S}}(l,t) = 1$ holds for all goal locations $l \in G$ and all $t \leq t_{\max}$,
- $Pr_{\mathcal{S}}(l,t_{\max}) = 0$ holds for all non-goal locations $l \notin G$, and
- $Pr_{\mathcal{S}}(l,t) = 0$ holds for all locations $l \in L$ and all $t > t_{\max}$.

Obviously, these differential equations can also be used in the limit.

A.2 Timed positional schedulers suffice for optimal time-bounded reachability

In this section we give the full proof for Lemma 1.

Lemma 1. *For a CTMG $\mathcal{M}$ with only continuous locations, the inequations $f_{\min}(l,t) \leq Pr^{\mathcal{M}}_{\mathcal{S}}(l,t) \leq f_{\max}(l,t)$ hold for every scheduler $\mathcal{S}$, every location l, and every $t \in [0,t_{\max}]$.*

In the proof, we assume a scheduler that provides a 3ε better result, and then sacrifice one ε to transfer to cylindrical schedulers (going back to the simpler incomplete space of cylindrical schedulers, but with completed reachability measure), and then sacrificing a second ε to discard long histories from consideration (going back to the simple space of finite sums over cylindrical sets).

As a result, we can do the comparison in a simple finite structure.

Proof. We start with the comparison of $Pr^{\mathcal{M}}_{\mathcal{S}}(l,t)$ and $f_{\max}(l,t)$.

Let us assume that the claim is incorrect. Then, there is a CTMG $\mathcal{M}$ (with only continuous locations) with location l_0 and a scheduler $\mathcal{S}_{3\varepsilon}$ for $\mathcal{M}$ such that the time-bounded reachability probability is at least 3ε higher for some $\varepsilon > 0$ and $t_0 \in [0,t_{\max}]$. That is, $Pr^{\mathcal{M}}_{\mathcal{S}_{3\varepsilon}}(l_0,t_0) - f_{\max}(l_0,t_0) > 3\varepsilon$

Let us fix appropriate $\mathcal{M}$, l_0, and $\mathcal{S}_{3\varepsilon}$. (Note that $\mathcal{S}_{3\varepsilon}$ does not have to be timed positional or deterministic.)

Recall that $\mathcal{S}_{3\varepsilon}$ is the limit point of a Cauchy sequence of cylindrical schedulers. Hence, almost all of these cylindrical schedulers have distance $< \varepsilon$ to $\mathcal{S}_{3\varepsilon}$.

Let us fix such a cylindrical scheduler $\mathcal{S}_{2\varepsilon}$ with distance $< \varepsilon$ to $\mathcal{S}_{3\varepsilon}$. The time-bounded reachability probability of $\mathcal{S}_{2\varepsilon}$ is still at least 2ε higher compared to $f_{\max}$. That is, $Pr^{\mathcal{M}}_{\mathcal{S}_{2\varepsilon}}(l_0,t_0) - f_{\max}(l_0,t_0) > 2\varepsilon$ holds.

For $\mathcal{S}_{2\varepsilon}$, we now consider a tightened form of time-bounded reachability, where we additionally require that the goal region is to be reached within n_ε steps. We choose n_ε big enough that the likelihood of seeing more than n_ε discrete events is less than ε. We call this time-bounded n_ε reachability.

Remark: We can estimate n_ε by taking the maximal transition rate $\lambda_{\max} = \max\{\mathbf{R}(l,a,L) \mid l \in L_c, a \in Act\}$, and choose n_ε big enough that the likelihood of having more than n_ε transitions was smaller than ε even if all transitions had transition rate $\lambda_{\max}$. As the number of steps is Poisson distributed in this case, a suitable n_ε is easy to find.

The adjustment to time-bounded n_ε reachability leads to a small change in the initialisation of the differential equations: for timed histories π of length $> n_\varepsilon$ that do not contain a location $l \in G$ in the goal region within the first n_ε steps, we use $f(\pi, t) = 0$ (even if it contains a goal region after more than n_ε steps) for all $t \in [0, t_{\max}]$. As the probability measure of all timed histories π of length $> n_\varepsilon$ is $< \varepsilon$, time-bounded n_ε reachability for $\mathcal{S}_{2\varepsilon}$ is still at least ε higher than the value for $f_{\max}$.

Let us use $f(\pi, t)$ to express the time-bounded n_ε reachability for $\mathcal{S}_{2\varepsilon}$ on a path π at time t. Then this claim can be phrased as $f(l_0, t_0) - f_{\max}(l_0, t_0) > \varepsilon$.

We have now reached a finite structure, and can easily show that this leads to a contradiction: we provide an inductive argument which demonstrates that $f_{\max}(l, t) \geq f(\pi, t)$ holds for all π that end in l and all $t \in [0, t_{\max}]$.

As a basis for our induction, this obviously holds for all timed histories longer than n_ε: in this case, $f_{\max}(l, t) = 1$ or $f(\pi, t) = 0$ holds (where the or is not exclusive).

For our induction step, let us assume we have demonstrated the claim for all histories of length $> n$. Let us, for a timed history π of length n that ends in l and some point $t \in [0, t_{\max}]$ assume that $f_{\max}(l, t) \leq f(\pi, t)$.

For $l \in G$ the initialisation conditions immediately lead to the contradiction $1 < f(\pi, t)$. For $l \notin G$, we can stepwise infer

$$\begin{aligned}
-\dot{f}_{\max}(l,t) &= \max_{a \in Act(l)} \textstyle\sum_{l' \in L} \mathbf{R}(l,a,l') \cdot (f_{\max}(l',t) - f_{\max}(l,t)) \\
&\geq \textstyle\sum_{a \in Act} \mathcal{S}(\pi,t)(a) \sum_{l' \in L} \mathbf{R}(l,a,l') \cdot (f_{\max}(l',t) - f_{\max}(l,t)) \\
&\geq \textstyle\sum_{a \in Act} \mathcal{S}(\pi,t)(a) \sum_{l' \in L} \mathbf{R}(l,a,l') \cdot (f_{\max}(l',t) - f(\pi,t)) \\
& \qquad\qquad \text{(with } f_{\max}(l,t) \leq f(\pi,t)\text{)} \\
&\geq \textstyle\sum_{a \in Act} \mathcal{S}(\pi,t)(a) \sum_{l' \in L} \mathbf{R}(l,a,l') \cdot \Big(f(\pi \xrightarrow{a,t} l',t) - f(\pi,t)\Big) \\
& \qquad\qquad \text{(with I.H.)} \\
&= -\dot{f}(\pi,t).
\end{aligned}$$

Taking into account that $f_{\max}(l, t_{\max})$ and $f(\pi, t_{\max})$ are both initialised to 0 for $l \notin G$, we can, using the just demonstrated $f_{\max}(l, t) \leq f(\pi, t) \Rightarrow \dot{f}_{\max}(l, t) \leq \dot{f}(\pi, t)$, infer $f_{\max}(l, t) \geq f(\pi, t)$ for all $t \in [0, t_{\max}]$: this inequation holds on the right fringe of the interval (initialisation), and when we follow the curves of $f(l, t)$ and $f_{\max}(l, t)$ to the left along $[0, t_{\max}]$, then every time f would catch up with $f_{\max}$, f cannot fall steeper than $f_{\max}$ (where 'fall' takes the usual left-to-right view, in the right-to-left direction we consider one should maybe say 'cannot have a steeper ascent') at such a position, and hence cannot not get above $f_{\max}$.

In particular, $f(l_0, t_0) \leq f_{\max}(l_0, t_0)$, which contradicts the initial assumption. The min-case can be proven accordingly. □

B Reproof of Theorem 5

To lift Theorem 6 to the full class of CTMGs, we reprove Theorem 5 in the style of the proof of Theorem 3. Recall that Theorem 3 establishes the existence of

an optimal cylindrical scheduler *using* the existence of an optimal measurable scheduler, and the form of the (differential) equations defining the time-bounded reachability probability for it. The proof given in this appendix can therefore not been used to supersede the proof in the article.

First we observe from the proof of Theorem 5 that, for discrete locations $l \in L_d$, the equations

$$f_{\mathsf{opt}}(l,t) = \underset{a \,\in\, Act(l)}{\mathsf{opt}} \sum_{l' \in L} \mathbf{P}(l, a_l, l') \cdot f_{\mathsf{opt}}(l', t) \text{ for } t \in [0, t_{\max}] \ ,$$

hold for $\mathsf{opt} \in \{\min, \max\}$, and that they together with the differential equations for the continuous locations (the differential equations remain unchanged), define f_{opt}.

The difference in the proof of Theorem 8 compared to the proof of Theorem 3 are marginal, but for the sake to readability we give the complete proof here.

Theorem 8. *For a single player continuous-time Markov game with only a reachability player, there is an optimal deterministic scheduler with finitely many switching points.*

Proof. We have seen that the true optimal reachability probability is defined by a system of equations and differential equations. In this proof we consider the effect of starting with the 'correct' values for a time $t \in [0, t_{\max}]$, but *locally fix a positional strategy* for a small left or right ε-environment of t. That is, we consider only schedulers that keep their decision constant for a (sufficiently) small time ε before or after t.

Given a CTMG $\mathcal{M}$, we consider the equations and differential equations that describe the development of the reachability probability for each location l under a positional deterministic strategy D:

$$-\dot{Pr}_l^D(\tau) = \sum_{l' \in L} \mathbf{R}(l, a_l, l') \cdot \left(Pr_{l'}^D(\tau) - Pr_l^D(\tau)\right) \quad \text{for } l \in L_c,$$

$$Pr_l^D(\tau) = \sum_{l' \in L} \mathbf{P}(l, a_l, l') \cdot Pr_{l'}^D(\tau) \quad \text{for } l \in L_d,$$

where a_l is the action chosen at l by D, starting at the support point $f_{\max}(l,t)$.

Different to the development of the true probability, the development of these linear differential equations provides us with smooth functions. This provides us with more powerful techniques when comparing two locally positional strategies: each deterministic scheduler defines a system $\dot{y} = Ay$ of ordinary homogeneous linear differential equations with constant coefficients.

As a result, the solutions $Pr_l^D(\tau)$ of these differential equations—and hence their differences $Pr_l^{D'}(\tau) - Pr_l^D(\tau)$—can be written as finite sums $\sum_{i=1}^{n} P_i(\tau) e^{\lambda_i \tau}$, where P_i is a polynomial and the λ_i may be complex. Consequently, these functions are holomorphic.

Using the identity theorem for holomorphic functions, t can only be a limit point of the set of 0 points of $Pr_l^{D'}(\tau) - Pr_l^D(\tau)$ if $Pr_l^{D'}(\tau)$ and $Pr_l^D(\tau)$ are

identical on an ε-environment of t. The same applies to their derivations: $\dot{Pr}_l^{D'}(\tau) - \dot{Pr}_l^{D}(\tau)$ either has no limit point in t, or $\dot{Pr}_l^{D'}(\tau)$ and $\dot{Pr}_l^{D}(\tau)$ are identical on an ε-environment of t.

For the remainder of the proof, we fix, for a given time t, a sufficiently small $\varepsilon > 0$ such that, for each pair of schedulers D and D' the following holds: for every location $l \in L_c$, $\dot{Pr}_l^{D'}(\tau) - \dot{Pr}_l^{D}(\tau)$ is either < 0, $= 0$, or > 0 on the complete interval $L_\varepsilon^t = (t-\varepsilon, t) \cap [0, t_{\max}] \ni \tau$, and, possibly with different sign, for the complete interval $R_\varepsilon^t = (t, t+\varepsilon) \cap [0, t_{\max}] \ni \tau$; and for every location $l \in L_d$, $Pr_l^{D}(\tau) - Pr_l^{D'}(\tau)$ is either < 0, $= 0$, or > 0 on the complete interval $L_\varepsilon^t = (t-\varepsilon, t) \cap [0, t_{\max}] \ni \tau$, and, possibly with different sign, for the complete interval $R_\varepsilon^t = (t, t+\varepsilon) \cap [0, t_{\max}] \ni \tau$.

We argue the case for the left ε-environment L_ε^t. In the '$>$' case for a location l, we say that D is *l-better* than D'. We call D *preferable* over D' if D' is not l-better than D for any location l, and *better* than D' if D is preferable over D' and l-better for some $l \in L$.

If D' is l-better than D in exactly a non-empty set $L_b \subset L$ of locations, then we can obviously use D' to construct a strategy D'' that is better than D by switching to the strategies of D' in exactly the locations L_b.

Since we choose our strategies from a finite domain—the deterministic positional schedulers—this can happen only finitely many times. Hence we can stepwise *strictly* improve a strategy, until we have constructed a strategy $D_{\max}$ that is preferable over all others.

By the definition of being preferable over all other strategies, $D_{\max}$ satisfies for all $\tau \in L_\varepsilon^t$

$$-\dot{Pr}_l^{D_{\max}}(\tau) = \max_{a \in Act(l)} \sum_{l' \in L} \mathbf{R}(l,a,l') \cdot \left(Pr_{l'}^{D_{\max}}(\tau) - Pr_l^{D_{\max}}(\tau) \right) \quad l \in L_c \ ,$$

$$Pr_l^{D_{\max}}(\tau) = \max_{a \in Act(l)} \sum_{l' \in L} \mathbf{P}(l,a,l') \cdot Pr_{l'}^{D_{\max}}(\tau) \quad l \in L_d \ .$$

We can use the same method for the right ε-environment R_ε^t, and pick the decision for t arbitrarily; we use the decision from the respective left ε environment.

Now we have fixed, for an ε-environment of an arbitrary $t \in [0, t_{\max}]$, an optimal scheduler with at most one switching point. As this is possible for all points in $[0, t_{\max}]$, the sets $I_\varepsilon^t = L_\varepsilon^t \cup R_\varepsilon^t$ define an open cover of $[0, t_{\max}]$. Using the compactness of $[0, t_{\max}]$, we infer a finite sub-cover, which establishes the existence of a strategy with a finite number of switching points. □

Again, the proof for single player safety games runs accordingly.

Theorem 9. *For a single player continuous-time Markov game with only a safety player, there is an optimal deterministic scheduler with finitely many switching points.*