Modelling and Analysis of Markov Reward Automata (extended version)

Dennis Guck[1], Mark Timmer[1], Hassan Hatefi[2],
Enno Ruijters[1] and Mariëlle Stoelinga[1]

[1] Formal Methods and Tools, University of Twente, The Netherlands
[2] Dependable Systems and Software, Saarland University, Germany

Abstract. Costs and rewards are important ingredients for many types of systems, modelling critical aspects like energy consumption, task completion, repair costs, and memory usage. This paper introduces Markov reward automata, an extension of Markov automata that allows the modelling of systems incorporating *rewards* (or *costs*) in addition to nondeterminism, discrete probabilistic choice and continuous stochastic timing. Rewards come in two flavours: action rewards, acquired instantaneously when taking a transition, and state rewards, acquired while residing in a state. We present algorithms to optimise three reward functions: the expected cumulative reward until a goal is reached, the expected cumulative reward until a certain time bound, and the long-run average reward. We have implemented these algorithms in the SCOOP/IMCA tool chain and show their feasibility via several case studies.

1 Introduction

The design of computer systems involves many trade offs: Is it cost-effective to use multiple processors to increase availability and performance? Should we carry out preventive maintenance to save future repair costs? Can we reduce the clock speed to save energy, while still meeting the required performance bounds? How can we best schedule a task set so that the operational costs are minimised? Such optimisation questions typically involve the following ingredients:

(1) rewards or costs, to measure the quality of the solution;
(2) (stochastic) timing to model speed or delay;
(3) discrete probability to model random phenomena like failures; and
(4) nondeterminism to model the choices in the optimisation process.

This paper introduces Markov reward automata (MRAs), a novel model that combines the ingredients mentioned above. It is obtained by adding rewards to the formalism of Markov automata (MAs) [15]. We support two types of rewards: *Action rewards* are obtained directly when taking a transition, and *state rewards* model the reward per time unit while residing in a state. Such reward extensions have shown valuable in the past for less expressive models, for instance leading to the tool MRMC [25] for model checking reward-based properties over

CTMCs [22] and DTMCs [1] with rewards. With our MRA model we provide a natural combination of the EMPA [3] and PEPA [9] reward formalisms.

By generalising MAs, MRAs provide a well-defined semantics for generalised stochastic Petri nets (GSPNs) [13], dynamic fault trees [4] and the domain-specific language AADL [5]. Recent work also demonstrated that MAs (and hence MRAs as well) are suitable for modelling and analysing distributed algorithms such as a leader election protocol, performance models such as a polling system and hardware models such as a processor grid [32].

Model checking algorithms for MA against Continuous Stochastic Logic (CSL) properties were discussed in [21]. Notions of strong, weak and branching bisimulation have been defined to equate behaviourally equivalent MAs [15,29,12,32], and the process-algebraic language MAPA was introduced for easily specifying large MAs in a concise manner [33]. Several types of reduction techniques [35,34] have been defined for the MAPA language and implemented in the tool SCOOP, optimising specifications to decrease the state space of the corresponding MAs while staying bisimilar [31,18]. This way, MAs can be generated efficiently in a direct way (as opposed to first generating a large model and then reducing), thus partly circumventing the omnipresent state space explosion. Additionally, the game-based abstraction refinement technique developed in [6] provides a sound approximation of time-bounded reachability over a substantially reduced abstract model. The tool IMCA [17,18] was developed to analyse the concrete MAs that are generated by SCOOP. It includes algorithms for computing time-bounded reachability probabilities, expected times and long-run averages for sets of goal states within an MA.

While the framework in place already works well for computing probabilities and expected durations, it did not yet support rewards or costs. Therefore, we extend the MAPA language from MAs to the realm of MRAs and extend most of SCOOP's reduction techniques to efficiently generate them. Further, we present algorithms for three optimisation problems over MRAs. That is, we resolve the nondeterministic choices in the MRA such that one of three optimisation criteria is minimised (or maximised):

(1) the expected cumulative reward to reach a set of goal states,
(2) the expected cumulative reward until a given time bound, and
(3) the long-run average reward.

The current paper is a first step towards a fully quantitative system design formalism. As such, we focus on positive rewards. Negative rewards, more complex optimisation criteria, as well as the handling of several rewards as multi-optimisation problem are important topics for future research. We provide detailed proofs of our main results in the appendix.

2 Markov reward automata

MAs were introduced as the union of Interactive Markov Chains (IMCs) [24] and Probabilistic Automata (PAs) [28]. Hence, they feature nondeterminism, as

well as Markovian rates and discrete probabilistic choice. We extend this model with reward functions for both the states and the transitions.

Definition 1 (Background). *A* probability distribution *over a countable set S is a function $\mu\colon S \to [0,1]$ such that $\sum_{s\in S} \mu(s) = 1$. For $S' \subseteq S$, let $\mu(S') = \sum_{s\in S'} \mu(s)$. We write $\mathbb{1}_s$ for the* Dirac distribution *for s determined by $\mathbb{1}_s(s) = 1$. We use $\mathsf{Distr}(S)$ to denote the set of all probability distributions over S.*

Given an equivalence relation $R \subseteq S \times S$, we write $[s]_R$ for the equivalence class of s induced by R, i.e., $[s]_R = \{s' \in S \mid (s,s') \in R\}$. Given two probability distributions $\mu, \mu' \in \mathsf{Distr}(S)$ and an equivalence relation R, we write $\mu \equiv_R \mu'$ to denote that $\mu([s]_R) = \mu'([s]_R)$ for every $s \in S$.

2.1 Markov reward automata

Before defining MRAs, we recall the definition of MAs. It assumes a countable universe of actions *Act*, with $\tau \in Act$ the invisible internal action.

Definition 2 (Markov Automata). *A* Markov automaton (MA) *is a tuple $\mathcal{A} = \langle S, s^0, A, \hookrightarrow, \rightsquigarrow \rangle$, where*

- *S is a countable set of* states*, of which $s^0 \in S$ is the* initial state*;*
- *$A \subseteq Act$ is a countable set of* actions*, including τ;*
- *$\hookrightarrow \subseteq S \times A \times \mathsf{Distr}(S)$ is the* probabilistic transition relation*;*
- *$\rightsquigarrow \subseteq S \times \mathbb{R}_{>0} \times S$ is the* Markovian transition relation*;*

If $(s,\alpha,\mu) \in \hookrightarrow$, we write $s \overset{\alpha}{\hookrightarrow} \mu$ and say that action α can be executed *from state s, after which the probability to go to each $s' \in S$ is $\mu(s')$. If $(s,\lambda,s') \in \rightsquigarrow$, we write $s \overset{\lambda}{\rightsquigarrow} s'$ and say that s moves to s' with rate λ.*

A state $s \in S$ that has at least one transition $s \overset{a}{\hookrightarrow} \mu$ is called *probabilistic*. A state that has at least one transition $s \overset{\lambda}{\rightsquigarrow} s'$ is called *Markovian*. Note that a state could be both probabilistic and Markovian.

The *rate* between two states $s, s' \in S$ is $\mathbf{R}(s,s') = \sum_{(s,\lambda,s')\in\rightsquigarrow} \lambda$, and the *outgoing rate of s* is $E(s) = \sum_{s'\in S} \mathbf{R}(s,s')$. We require $E(s) < \infty$ for every state $s \in S$. If $E(s) > 0$, the *branching probability distribution* after this delay is denoted by $\mathbb{P}_s$ and defined by $\mathbb{P}_s(s') = \frac{\mathbf{R}(s,s')}{E(s)}$ for every $s' \in S$. By definition of the exponential distribution, the probability of leaving a state s within t time units is given by $1 - e^{-E(s)\cdot t}$ (given $E(s) > 0$), after which the next state is chosen according to $\mathbb{P}_s$. Further, we denote by $A(s)$ the set of all enabled actions in state s.

MAs adhere to the *maximal progress assumption*, prescribing τ-transitions to never be delayed. Hence, a state that has at least one outgoing τ-transition can never take a Markovian transition. This fact is captured below in the definition of extended transitions, which is used to provide a uniform manner for dealing with both probabilistic and Markovian transitions.

Definition 3 (Extended action set). *Let $\mathcal{A} = \langle S, s^0, A, \hookrightarrow, \rightsquigarrow \rangle$ be an MA, then the* extended action set of $\mathcal{M}$ *is given by $A^\chi = A \cup \{\chi(r) \mid r \in \mathbb{R}_{>0}\}$. The actions $\chi(r)$ represent exit rates and are used to distinguish probabilistic and Markovian transitions. For $\alpha = \chi(\lambda)$, we define $E(\alpha) = \lambda$. If $\alpha \in A$, we set $E(\alpha) = 0$. Given a state $s \in S$ and an action $\alpha \in A^\chi$, we write $s \xrightarrow{\alpha} \mu$ if either*

- *$\alpha \in A$ and $s \overset{\alpha}{\hookrightarrow} \mu$, or*
- *$\alpha = \chi(E(s))$, $E(s) > 0$, $\mu = \mathbb{P}_s$ and there is no μ' such that $s \overset{\tau}{\hookrightarrow} \mu'$.*

A transition $s \xrightarrow{\alpha} \mu$ is called an extended transition. *We use $s \xrightarrow{\alpha} t$ to denote $s \xrightarrow{\alpha} \mathbb{1}_t$, and write $s \to t$ if there is at least one action α such that $s \xrightarrow{\alpha} t$. We write $s \xrightarrow{\alpha,\mu} s'$ if there is an extended transition $s \xrightarrow{\alpha} \mu$ such that $\mu(s') > 0$.*

Note that each state has an extended transition per probabilistic transition, while it has only one for all its Markovian transitions together (if there are any).

We now formally introduce the MRA. For simplicity, we chose to define MRAs in terms of two seperate reward functions. Hence, instead of integrating rewards into the transition relations, there is a separate reward function over extended transitions. This also simplifies the compatibility to the notion of MAs.

Definition 4 (Markov Reward Automata[1]). *A* Markov Reward Automaton (MRA) *is a tuple $\mathcal{M} = \langle \mathcal{A}, \rho, r \rangle$, where*

- *$\mathcal{A}$ is a Markov automaton;*
- *$\rho \colon S \to \mathbb{R}_{\geq 0}$ is the* state-reward function*;*
- *$r \colon S \times A^\chi \times \mathsf{Distr}(S) \to \mathbb{R}_{\geq 0}$ is the* transition-reward function*.*

The function ρ associates a real number to each state. This number may be zero, indicating the absence of a reward. The state-based rewards are gained *while being in a state*, and are proportional to the duration of this stay. The function r associates a real number to a transition. This number may be zero, indicating the absence of a reward. The transition-based rewards are gained when *taking the transition*.

2.2 Paths, policies and rewards

As for traditional labelled transition systems (LTSs), the behaviour of MAs and MRAs can also be expressed by means of paths[2]. A *path* in $\mathcal{M}$ is a finite sequence $\pi^{\text{fin}} = s_0 \xrightarrow{a_0,\mu_0,t_0} s_1 \xrightarrow{a_1,\mu_1,t_1} \dots \xrightarrow{a_{n-1},\mu_{n-1},t_{n-1}} s_n$ from some state s_0 to a state s_n ($n \geq 0$), or an infinite sequence $\pi^{\text{inf}} = s_0 \xrightarrow{a_0,\mu_0,t_0} s_1 \xrightarrow{a_1,\mu_1,t_1} s_2 \xrightarrow{a_2,\mu_2,t_2} \dots$, with $s_i \in S$ for all $0 \leq i \leq n$ and all $0 \leq i$, respectively. The step $s_i \xrightarrow{a_i,\mu_i,t_i} s_{i+1}$ denotes that after residing t_i time units in s_i, the MRA has moved via action a_i and probability distribution μ_i to s_{i+1}. We use

[1] Note that we introduce a separate transition-reward function instead of encoding the reward in the transition relation as in the ATVA paper [20].

[2] Note that we removed the reward from the path expression and decremented the indices compared to [20].

prefix(π, t) to denote the prefix of path π up to and including time t, formally *prefix*$(\pi, t) = s_0 \xrightarrow{a_0,\mu_0,t_0} \dots \xrightarrow{a_{i-1},\mu_{i-1},t_{i-1}} s_i$ such that $t_0 + \dots + t_{i-1} \leq t$ and $t_0 + \dots + t_{i-1} + t_i > t$. We use *step*$(\pi, i)$ to denote the transition $s_{i-1} \xrightarrow{a_{i-1}} \mu_i$. When π is finite we define $|\pi| = n$, *last*$(\pi) = s_n$, and for every path $\pi[i] = s_i$. Further, we denote by π^j the path π up to and including state s_j and with $\pi[j] = s_j$ the state on path π on position j. Let *paths** and *paths* denote the set of finite and infinite paths, respectively. We define the *total reward* of a finite path π by

$$\mathit{reward}(\pi) = \sum_{i=0}^{|\pi|-1} \rho(\pi[i]) \cdot t_i + r(\pi[i], a_i, \mu_i) \tag{1}$$

Rewards can be used to model many quantitative aspects of systems, like energy consumption, memory usage, deployment or maintenance costs, etc. The total reward of a path (e.g., total amount of energy consumed) is obtained by adding all rewards along that path, that is, all state rewards multiplied by the sojourn times of the corresponding states plus all action rewards on the path.

Policies. Policies resolve the nondeterministic choices in an MRA, i.e., make a choice over the outgoing probabilistic transitions in a state. Given a policy, the behaviour of an MRA is fully probabilistic. Formally, a *policy*, ranged over by D, is a measurable function such that $D\colon \mathit{paths}^* \to \mathsf{Distr}(A^\chi \times \mathsf{Distr}(S))$ such that for each path π, where $s_n = \mathit{last}(\pi)$, for all $A(s_n) = \{\alpha \in A^\chi \mid \exists \mu \in \mathsf{Distr}(S) \,.\, s_n \xrightarrow{\alpha} \mu\}$ and $\mu \in \mathsf{Distr}(S)$, $D(\pi)(\alpha, \mu) > 0$ implies $s \xrightarrow{\alpha} \mu$. We denote by *GM* (generic measurable) the most general class of such policies which are measurable; for more details on measurability see [26]. Policies are classified based on the level of information they used to resolve nondeterminism. A stationary deterministic policy is a mapping $D\colon S \to A^\chi \times \mathsf{Distr}(S)$ such that $D(s)$ chooses only from transitions that emanate from s; such policies always take the same transition in a state s. A time-dependent policy may decide on the basis of the states visited so far and their timings. For more details about different classes of policies and their relations we refer to [27]. Given a policy D and an initial state s, a measurable set of paths is equipped with the probability measure $\Pr_{s,D}$.

2.3 Strong bisimulation

We define a notion of strong bisimulation for MRAs. As for LTSs, PAs, IMCs and MAs, it equates systems the are equivalent in the sense that every step of one system can be mimicked by the other, and vice versa.

Definition 5 (Strong bisimulation). *Given an MRA $\mathcal{M} = \langle \mathcal{A}, \rho, r \rangle$, an equivalence relation $R \subseteq S \times S$ is a* strong bisimulation *for $\mathcal{M}$ if for every $(s, s') \in R$ and all $\alpha \in A^\chi, \mu \in \mathsf{Distr}(S), r \in \mathbb{R}_{\geq 0}$, it holds that $\rho(s) = \rho(s')$ and*

$$s \xrightarrow{\alpha} \mu \implies \exists \mu' \in \mathsf{Distr}(S) \,.\, s' \xrightarrow{\alpha} \mu' \wedge \mu \equiv_R \mu' \wedge r(s, a, \mu) = r(s', a, \mu')$$

Two states $s, s' \in S$ are strongly bisimilar *(denoted by $s \approx s'$) if there exists a strong bisimulation R for $\mathcal{M}$ such that $(s, s') \in R$. Two MRAs $\mathcal{M}, \mathcal{M}'$ are strongly bisimilar (denoted by $\mathcal{M} \approx \mathcal{M}'$) if their initial states are strongly bisimilar in their disjoint union.*

Clearly, when setting all state-based and action-based rewards to 0, MRAs coincide with MAs. Additionally, our definition of strong bisimulation then reduces to the definition of strong bisimulation for MAs. Since it was already shown in [14] that strong bisimulation for MAs coincides with the corresponding notions for all subclasses of MAs, this also holds for our definition. Hence, it safely generalises the existing notions of strong bisimulation.

2.4 Parallel composition

We can easily generalise the definition of parallel composition from MAs to MRAs, using the notations from [32] and synchronising on mutual actions as in [15]. In addition to the original construction, we now also add up the state-based rewards for each pair (s, t) and add up the action-based rewards in synchronised transitions.

Remark 1. For simplification of the parallel composition and the MAPA specification we assume, without loss of generality[3], that only probabilistic transitions are assigned rewards. This can easily be achieved by transforming each transition $s \xrightarrow{\chi(\lambda)} \mu$ with $m = r(s, \chi(\lambda), \mu) > 0$ to a pair of transitions $s \xrightarrow{\chi(\lambda)} \mathbb{1}_t$ and $t \xrightarrow{\tau} \mu$ with $r(s, \chi(\lambda), \mathbb{1}_t) = 0$ and $r(t, \tau, \mu) = m$.

Remark 1 is vital for the current definition of parallel composition. Without it, we need extra cases dealing with the parallel composition of two self-loops having identical rates and rewards, additionally complicating the conditions for the existing rules to be applicable. Now, we can handle self-loops as before, not worrying about their rewards as these are always 0 anyway.

Definition 6 (Parallel composition). *Given MRAs $\mathcal{M}_1 = \langle \mathcal{A}_1, \rho_1, r_1 \rangle$ and $\mathcal{M}_2 = \langle \mathcal{A}_2, \rho_2, r_2 \rangle$, their parallel composition is the system $\mathcal{M}_1 \,||\, \mathcal{M}_2 = \langle \mathcal{A}, \rho, r \rangle$, where $\mathcal{A} = \mathcal{A}_1 \,||\, \mathcal{A}_2$ with*

- $S = S_1 \times S_2$;
- $A = A_1 \cup A_2$;
- $s^0 = (s_1^0, s_2^0)$;
- $\rho(s_1, s_2) = \rho_1(s_1) + \rho_2(s_2)$;
- $$r(s, a, \mu) = \begin{cases} r(s_1, a, \mu_1) & \text{if } s = (s_1, s_2) \wedge a \in A_1 \setminus A_2 \\ r(s_2, a, \mu_2) & \text{if } s = (s_1, s_2) \wedge a \in A_2 \setminus A_1 \\ r(s_1, a, \mu_1) + r(s_2, a, \mu_2) & \text{if } s = (s_1, s_2) \wedge a \in A_1 \cap A_2. \end{cases}$$

[3] We note that this transformation does *not* preserve the notion of strong bisimulation that we define in Section 2.3. However, it does not influence any imaginable property over the model that does not take into account path length.

$$\frac{s_1 \overset{a}{\hookrightarrow} \mu_1}{(s_1, s_2) \overset{a}{\hookrightarrow} \mu_1 \times \mathbb{1}_{s_2}}\ a \in A_1 \setminus A_2 \qquad \frac{s_2 \overset{a}{\hookrightarrow} \mu_2}{(s_1, s_2) \overset{a}{\hookrightarrow} \mathbb{1}_{s_1} \times \mu_2}\ a \in A_2 \setminus A_1$$

$$\frac{s_1 \overset{a}{\hookrightarrow} \mu_1 \qquad s_2 \overset{a}{\hookrightarrow} \mu_2}{(s_1, s_2) \overset{a}{\hookrightarrow} \mu_1 \times \mu_2}\ a \in A_1 \cap A_2$$

$$\frac{s_1 \overset{\lambda}{\rightsquigarrow} s_1' \qquad s_1 \neq s_1'}{(s_1, s_2) \overset{\lambda}{\rightsquigarrow} (s_1', s_2)} \qquad \frac{s_2 \overset{\lambda}{\rightsquigarrow} s_2' \qquad s_2 \neq s_2'}{(s_1, s_2) \overset{\lambda}{\rightsquigarrow} (s_1, s_2')} \qquad \frac{\lambda(s_1, s_2) > 0}{(s_1, s_2) \overset{\lambda(s_1, s_2)}{\rightsquigarrow} (s_1, s_2)}$$

Table 1. Inference rules for the transitions of a parallel composition, where $\lambda(s_1, s_2) = \mathbf{R}(s_1, s_1) + \mathbf{R}(s_2, s_2)$.

and $\hookrightarrow$ and $\rightsquigarrow$ the smallest relation fulfilling the inference rules in Table 1 (i.e., if all conditions above the line of a rule hold, then so should the condition below).

3 Quantitative analysis

This section shows how to perform quantitative analyses on MRAs. We will focus on three common reward measures: (1) The expected cumulative reward until reaching a set of goal states, (2) the expected cumulative reward until a given time-bound, and (3) the long-run average reward. Typical examples where these algorithms can be used are respectively: to minimise the average energy consumption needed to download and install a medium-size software update; to minimise the average maintenance cost of a railroad line over the first year of deployment; and to maximise the yearly revenues of a data center over a long time horizon. In the following we lift the algorithms from [18] to the realm of rewards. We focus on maximising the properties. The minimisation problem can be solved similarly — namely, by replacing max by min and sup by inf below.

3.1 Notation and preprocessing

Throughout this section, we consider a fixed MRA $\mathcal{M}$ with state space S and a set of goal states $G \subseteq S$. To facilitate the algorithms, we first perform three preprocessing steps.

(1) We consider only closed MRAs, which are not subject to further interaction. Therefore, we hide all actions (renaming them to τ), focussing on their induced rewards.
(2) Due to the maximal progress assumption, a Markovian transition will never be executed from a state with outgoing τ-transitions. Hence, we remove such Markovian transitions. Thus, each state either has one outgoing Markovian transition or only probabilistic outgoing transitions. We call these states Markovian and probabilistic respectively, and use *MS* and *PS* to denote the sets of Markovian and probabilistic states.

(3) To distinguish the different τ-transitions emerging from a state $s \in PS$, we assume w.l.o.g. that these are numbered from 1 to n_s, where n_s is the number of outgoing transitions. We write $\mu_s^{\tau_i}$ for the distribution induced by taking τ_i in state s and we write $r_s^{\tau_i}$ for the reward, instead of $r(s, \tau_i, \mu_s^{\tau_i})$. For Markovian transitions we write $\mathbb{P}_s$ and r_s, respectively.

3.2 Goal-bounded expected cumulative reward

We are interested in the minimal and maximal expected cumulative reward until reaching a set of goal states $G \subseteq S$. That is, we accumulate the state and transition rewards until a state in G is reached; if no state in G is reached, we keep on accumulating rewards.

The random variable $V_G \colon \mathit{paths} \to \mathbb{R}_{\geq 0}^{\infty}$ yields the accumulated reward before first visiting some state in G. For an infinite path π, we define

$$V_G(\pi) = \begin{cases} \mathit{reward}(\pi^j) & \text{if } \pi[j] \in G \ \wedge \ \forall i < j.\ \pi[i] \notin G \\ \mathit{reward}(\pi) & \text{if } \forall i.\ \pi[i] \notin G \end{cases}$$

The maximal expected reward to reach G from $s \in S$ is then defined as

$$\mathsf{eR}^{\max}(s, G) = \sup_{D \in GM} \mathbb{E}_{s,D}(V_G) = \sup_{D \in GM} \int_{paths} V_G(\pi) \Pr_{s,D}(d\pi) \tag{2}$$

where D is an arbitrary policy on $\mathcal{M}$.

To compute $\mathsf{eR}^{\max}$ we turn it into a classical Bellman equation: For all goal states, no more reward is accumulated, so their expected reward is zero. For Markovian states $s \notin G$, the state reward of s is weighted with the expected sojourn time in s plus the expected reward accumulated via its successor states plus the transition reward to them. For a probabilistic state $s \notin G$, we select the action that maximises the expected cumulative reward. Note that, since the accumulated reward is only relevant until reaching a state in G, we may turn all states in G into absorbing Markovian states.

Theorem 1 (Bellman equation) *The function* $\mathsf{eR}^{\max} \colon S \to \mathbb{R}_{\geq 0}^{\infty}$ *is the unique fixed point of the Bellman equation*

$$v(s) = \begin{cases} \frac{\rho(s)}{E(s)} + \sum\limits_{s' \in S} \mathbb{P}_s(s') \cdot (v(s') + r_s) & \text{if } s \in MS \setminus G \\ \max\limits_{\alpha \in A(s)} \sum\limits_{s' \in S} \mu_s^{\alpha}(s') \cdot (v(s') + r_s^{\alpha}) & \text{if } s \in PS \setminus G \\ 0 & \text{if } s \in G. \end{cases}$$

A direct consequence of Theorem 1 is that the supremum in (2) is attained by a stationary deterministic policy. Moreover, this result enables us to use standard solution techniques such as value iteration and linear programming to compute $\mathsf{eR}^{\max}(s, G)$. Note that by assigning $\rho(s) = 1$ to all $s \in MS$ and setting all other rewards to 0, we compute the expected time to reach a set of goal states.

3.3 Time-bounded expected cumulative reward

A time-bounded reward is the reward gained until a time bound t is reached and is denoted by the random variable $\mathit{reward}(\cdot, t)$. For an infinite path π, we first find the prefix of π up to t and then compute the reward using (1), i. e.

$$\mathit{reward}(\pi, t) = \mathit{reward}(\mathit{prefix}(\pi, t)) \tag{3}$$

The maximum time-bounded reward then is the maximum expected reward gained within some interval $I = [0, b]$, starting from some initial state s:

$$\mathcal{R}^{\max}(s, b) = \sup_{D \in GM} \int_{paths} \mathit{reward}(\pi, b) \Pr_{s,D}(d\pi) \tag{4}$$

Similar to time-bounded reachability there is a fixed point characterisation (FPC) for computing the optimal reward within some interval of time. Here we focus on the maximum case; the minimum can be extracted similarly.

Lemma 2 (Fixed Point Characterisation) *Given a Markov reward automaton $\mathcal{M}$ and a time bound $b \geq 0$. The maximum expected cumulative reward from state $s \in S$ until time bound b is the least fixed point of higher order operator $\Omega \colon (S \times \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}) \mapsto (S \times \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0})$, such that*

$$\Omega(F)(s,b) = \begin{cases} \left(r_s + \frac{\rho(s)}{E(s)}\right)\left(1 - e^{-E(s)b}\right) & \\ \quad + \int_0^b E(s) e^{-E(s)t} \sum_{s' \in S} \mathbb{P}_s(s') F(s', b-t)\,\mathrm{d}t & s \in MS \wedge b \neq 0 \\ \max_{\alpha \in A(s)} \left(r_s^\alpha + \sum_{s' \in S} \mu_s^\alpha(s') F(s', b)\right) & s \in PS \\ 0 & \textit{otherwise.} \end{cases}$$

This FPC is a generalisation of that for time-bounded reachability [18, Lemma 1], taking both action and state rewards into account. The proof goes along the same lines as that of [26, Theorem 6.1].

Discretisation. Similar to time-bounded reachability, the FPC is not algorithmically tractable and needs to be discretised: we have to divide the time horizon $[0, b]$ into a (generally large) number of equidistant time steps, each of length $0 < \delta \leq b$, such that $b = k\delta$ for some $k \in \mathbb{N}$. First, we express $\mathcal{R}^{\max}(s, b)$ in terms of its behaviour in the first discretisation step $[0, \delta)$. To do so, we partition the paths from s into the set $\mathcal{P}_1$ of paths that make their first Markovian jump in $[0, \delta)$ and the set $\mathcal{P}_2$ of paths that do not. We write $\mathcal{R}^{\max}(s, b)$ as the sum of

1. The expected reward obtained in $[0, \delta)$ by paths from $\mathcal{P}_1$
2. The expected reward obtained in $[\delta, b]$ by paths from $\mathcal{P}_1$
3. The expected reward obtained in $[0, \delta)$ by paths from $\mathcal{P}_2$
4. The expected reward obtained in $[\delta, b]$ by paths from $\mathcal{P}_2$

It turns out to be convenient to combine the first three items, denoted by $A(s,b)$, since the resulting term resembles the expression in Lemma 2:

$$\begin{aligned} A(s,b) &= \rho(s)\delta e^{-E(s)\delta} + \int_0^\delta E(s)e^{-E(s)t}\Big(\rho(s)t + r_s + \sum_{s'\in S}\mathbb{P}_s(s')\mathcal{R}^{\max}(s', b-t)\Big)\,\mathrm{d}t \\ &= \Big(r_s + \frac{\rho(s)}{E(s)}\Big)\Big(1-e^{-E(s)\delta}\Big) + \int_0^\delta E(s)e^{-E(s)t}\sum_{s'\in S}\mathbb{P}_s(s')\mathcal{R}^{\max}(s', b-t)\,\mathrm{d}t \end{aligned} \tag{5}$$

where the first equality follows directly from the definition of $A(s,b)$ and the second equality is along the same lines as the proof of Lemma 2. It can easily be seen that $\mathcal{R}^{\max}(s,b) = A(s,b) + e^{-E(s)\delta}\mathcal{R}^{\max}(s, b-\delta)$.

Exact computation of $A(s,b)$ is in general still intractable due to the term $\mathcal{R}^{\max}(s', b-t)$. However, if the discretisation constant δ is very small, then, with high probability, at most one Markovian jump happens in each discretisation step. Hence, the reward gained by paths having multiple Markovian jumps within at least one such interval is negligible and can be omitted from the computation, while introducing only a small error. Technically, that means that we don't have to remember the remaining time within a discretisation step after a Markovian jump has happened. We can therefore discretise $A(s,b)$ into $\tilde{A}_\delta(s,k)$ and $\mathcal{R}^{\max}(s,b)$ into $\tilde{\mathcal{R}}^{\max}_\delta(s,k)$, just counting the number of discretisation steps k that are left instead of the actual time bound b:

$$\tilde{\mathcal{R}}^{\max}_\delta(s,k) = \tilde{A}_\delta(s,k) + e^{-E(s)\delta}\tilde{\mathcal{R}}^{\max}_\delta(s,k-1), \quad s \in MS \tag{6}$$

where $\tilde{A}_\delta(s,k)$ is defined by

$$\begin{aligned} \tilde{A}_\delta(s,k) &= \Big(r_s + \frac{\rho(s)}{E(s)}\Big)\Big(1-e^{-E(s)\delta}\Big) + \int_0^\delta E(s)e^{-E(s)t}\sum_{s'\in S}\mathbb{P}_s(s')\tilde{\mathcal{R}}^{\max}_\delta(s', k-1)\,\mathrm{d}t \\ &= \Big(r_s + \frac{\rho(s)}{E(s)} + \sum_{s'\in S}\mathbb{P}_s(s')\tilde{\mathcal{R}}^{\max}_\delta(s', k-1)\Big)\Big(1-e^{-E(s)\delta}\Big) \end{aligned} \tag{7}$$

Note that we used $\tilde{\mathcal{R}}^{\max}_\delta(s,k-1)$ instead of both $\mathcal{R}^{\max}(s,b-\delta)$ and $\mathcal{R}^{\max}(s,b-t)$.

Eq. (6) and (7) help us to establish a tractable discretised version of the FPC described in Lemma 2 and to formally define the discretised maximum time-bounded reward afterwards:

Definition 7 (Discretised Maximum Time-Bounded Reward). *Let $\mathcal{M}$ be an MRA, $b \geq 0$ a time bound and $\delta > 0$ a discretisation step such that $b = k\delta$ for some $k \in \mathbb{N}$. The discretised maximum time-bounded cumulative reward, $\tilde{\mathcal{R}}^{\max}_\delta$, is defined as the least fixed point of higher order operator $\Omega_\delta \colon (S \times \mathbb{N} \mapsto \mathbb{R}_{\geq 0}) \mapsto (S \times \mathbb{N} \mapsto \mathbb{R}_{\geq 0})$, such that*

$$\Omega_\delta(F)(s,k) = \begin{cases} \big(r_s + \frac{\rho(s)}{E(s)} + \sum_{s'\in S}\mathbb{P}_s(s')F(s',k-1)\big)\big(1-\mathrm{e}^{-E(s)\delta}\big) & \\ \qquad +\mathrm{e}^{-E(s)\delta}F(s,k-1) & s \in MS \wedge k \neq 0 \\ \max_{\alpha\in A(s)}\big(r_s^\alpha + \sum_{s'\in S}\mu_s^\alpha(s')F(s',k)\big) & s \in PS \\ 0 & \textit{otherwise.} \end{cases}$$

The reason behind the tractability of $\tilde{\mathcal{R}}^{\max}_{\delta}$ is hidden in Eq. (7). It brings two simplifications to the computation. First, it implies that $\tilde{\mathcal{R}}^{\max}_{\delta}$ is the conditional expected reward given that each step carries at most one Markovian transition. Second, it neglects to compute the reward after the first Markovian jump and simply assume that it is zero. We have shown the formal specification of the simplifications in [19, Lemma C1]. With the help of these simplifications, reward computation becomes tractable but indeed inexact.

The accuracy of $\tilde{\mathcal{R}}^{\max}_{\delta}$ depends on some parameters including the step size δ. The smaller δ is, the better the quality of discretisation is. It is possible to quantify the quality of the discretisation. To this end we need first to define some parameters of MRA. For a given MRA $\mathcal{M}$, assume that λ is the maximum exit rate of any Markovian state, i. e. $\lambda = \max_{s \in MS} E(s)$, and $\rho_{\max}$ is maximum state reward of any Markovian state, i. e. $\rho_{\max} = \max_{s \in MS} \rho(s)$. Moreover we define $r_{\max}$ as the maximum action reward that can be gained between two consecutive Markovian jumps. The value can be computed via Theorem 1, where we set Markovian states as the goal states. Given that $\mathsf{eR}^{\max}(s, MS)$ has already been computed, we define $r(s) = r_s + \sum_{s' \in S} \mathsf{eR}^{\max}(s', MS)$ for $s \in MS$, and $r(s) = \mathsf{eR}^{\max}(s, MS)$ otherwise. Finally we have $r_{\max} = \max_{s \in S} r(s)$. Note that in practice we use a value iteration algorithm to compute $r_{\max}$. With all of the parameters known, the following theorem quantifies the quality of the abstraction.

Theorem 3 *Let $\mathcal{M}$ be an MRA, $b \geq 0$ be a time bound, $\delta > 0$ be a discretisation step such that $b = k\delta$ for some $k \in \mathbb{N}$. Then for all $s \in S$:*

$$\tilde{\mathcal{R}}^{\max}_{\delta}(s,k) \leq \mathcal{R}^{\max}(s,b) \leq \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + \frac{b\lambda}{2}(\rho_{\max} + r_{\max}\lambda)(1 + \frac{b\lambda}{2})\delta$$

3.4 Long-run average reward

Next, we are interested in the average cumulative reward induced by a set of goal states $G \subseteq S$ in the long-run. Hence, all state and action rewards for states $s \in S \setminus G$ are set to 0. We define the random variable $\mathcal{L}_{\mathcal{M}} \colon \mathit{paths} \to \mathbb{R}_{\geq 0}$ as the long-run reward over paths in MRA $\mathcal{M}$. For an infinite path π let

$$\mathcal{L}_{\mathcal{M}}(\pi) = \lim_{t \to \infty} \frac{1}{t} \cdot \mathit{reward}(\pi, t).$$

Then, the maximal long-run average reward on $\mathcal{M}$ starting in state $s \in S$ is:

$$\mathsf{LRR}^{\max}_{\mathcal{M}}(s) = \sup_{D \in GM} \mathbb{E}_{s,D}(\mathcal{L}_{\mathcal{M}}) = \sup_{D \in GM} \int_{paths} \mathcal{L}_{\mathcal{M}}(\pi) \Pr_{s,D}(d\pi). \tag{8}$$

The computation of the expected long-run reward can be split into three steps:

1. Determine all maximal end components of MRA $\mathcal{M}$;
2. Determine $\mathsf{LRR}^{\max}_{\mathcal{M}_i}$ for each maximal end component $\mathcal{M}_i$;
3. Reduce the computation of $\mathsf{LRR}^{\max}_{\mathcal{M}}(s)$ to an SSP problem.

A sub-MRA $\mathcal{M}$ is a pair (S', K) where $S' \in S$ and K is a function that assigns to each state $s \in S'$ a non-empty set of actions, such that for all $\alpha \in K(s)$, $s \xrightarrow{\alpha} \mu$ with $\mu(s') > 0$ implies $s' \in S'$. An *end component* is a sub-MRA whose underlying graph is strongly connected; it is maximal (a *MEC*) w.r.t. K if it is not contained in any other end component (S'', K). In this section we focus on the second step. The first step can be performed by a graph-based algorithm [8,10] and the third step is as in [18].

A MEC can be seen as a unichain MRA: an MRA that yields a strongly connected graph structure under any stationary deterministic policy.

Theorem 4 *For a unichain MRA $\mathcal{M}$, for each $s \in S$ the value of $\mathsf{LRR}^{\max}_{\mathcal{M}}(s)$ equals*

$$\mathsf{LRR}^{\max}_{\mathcal{M}} = \sup_{D} \sum_{s \in S} \Big(\rho(s) \cdot \mathsf{LRA}^{\mathsf{D}}(s) + r_s^{D(s)} \cdot \nu^D(s) \Big)$$

where ν is the frequency of passing through a state, defined by

$$\nu^D(s) = \begin{cases} \mathsf{LRA}^D(s) \cdot E(s) & \text{if } s \in MS \\ \sum\limits_{s' \in S} \nu^D(s') \cdot \mu_{s'}^{D(s')}(s) & \text{if } s \in PS \end{cases}$$

and $\mathsf{LRA}^D(s)$ is the long-run average time spent in state s under stationary deterministic policy D.

Thus, the frequency of passing through a Markovian state equals the long-run average time spent in s times the exit rate, and for a probabilistic state it is the accumulation of the frequencies of the incoming transitions. Hence, the long-run reward gathered by a state s is defined by the state reward weighted with the average time spent in s and the action reward weighted by the frequency of passing through the state. Since in a unichain MRA $\mathcal{M}$, for any two states s, s', $\mathsf{LRR}^{\max}_{\mathcal{M}}(s)$ and $\mathsf{LRR}^{\max}_{\mathcal{M}}(s')$ coincides, we omit the starting state and just write $\mathsf{LRR}^{\max}_{\mathcal{M}}$. Note that probabilistic states are left immediately, so $\mathsf{LRA}^D(s) = 0$ if $s \in PS$. Further, by assigning $\rho(s) = 1$ to all $s \in MS \cap G$ and setting all other rewards to 0, we compute the long-run average time spent in a set of goal states.

Theorem 5 *The long-run average reward of a unichain MRA coincides with the limit of the time-bounded expected cumulative reward, such that $\mathsf{LRR}^D(s) = \lim\limits_{t \to \infty} \frac{1}{t} \mathcal{R}^D(s, t)$.*

For the equation from Theorem 4 it would be too expensive to compute for all possible policies and for each state the long-run average time as well as the frequency of passing through a state and weigh those with the associated rewards. Instead, we compute $\mathsf{LRR}^{\max}_{\mathcal{M}}$ by solving a system of linear inequations following the concepts of [10]. Given a unichain MRA $\mathcal{M}$, let k denote the optimal average reward accumulated in the long-run and executing the optimal policy. Then, for all $s \in S$ there is a function $h(s)$ that describes a differential cost per visit to state s, such that a system of inequations can be constructed as follows:

Minimise k subject to:

$$\begin{cases} h(s_i) = \frac{\overline{\rho}(s_i)}{E(s_i)} - \frac{k}{E(s_i)} + \sum_{s_j \in S} \mathbb{P}_{s_i}(s_j) \cdot h(s_j) & \text{if } s_i \in MS \\ h(s_i) \geq r^{\alpha}_{s_i} + \sum_{s_j \in S} \mu^{\alpha}_{s_i}(s_j) \cdot h(s_j) & \text{if } s_i \in PS \wedge \forall \alpha \in A(s_i) \end{cases} \tag{9}$$

where the state and action reward of Markovian states are combined as $\overline{\rho}(s_i) = \rho(s_i) + (r_{s_i} \cdot E(s_i))$. Standard linear programming algorithms, e.g., the simplex method [36], can be applied to solve the above system of linear equations.

To obtain the long-run average reward in an arbitrary MRA, we have to weigh the obtained long-run rewards in each maximal end component with the probability to reach those from s. This is equivalent to the third step in the long-run average computation of [18]. Further, for the discrete time setting [7] considers multiple long-run average objectives.

4 MAPA with rewards

The Markov Automata Process Algebra (MAPA) language allows MAs to be generated in an efficient and effective manner [32]. It is based on μCRL [16], allowing the standard process-algebraic constructs such as nondeterministic choice and action prefix to be used in a data-rich context: processes are equipped with a set of variables over user-definable data types, and actions can be parameterised based on the values of these variables. Additionally, conditions can be used to restrict behaviour, and nondeterministic choices over data types are possible. MAPA adds two operators to μCRL: a probabilistic choice over data types and a Markovian delay (both possibly depending on data parameters). We extend the original MAPA language by accompanying it with rewards.

Due to the action-based approach of process algebra, there is a clear separation between the action-based and state-based rewards. *Action-based rewards* are just added as decorations to the actions in the process-algebraic specification, whereas *state-based rewards* can be assigned to conditions; each state that fulfills a reward's condition is then assigned that reward. If a state satisfies multiple conditions, the rewards are accumulated.

We refer to [32] for a detailed exposition of the syntax and semantics of MAPA; this is trivially generalised to incorporate the action-based rewards. Here we give a brief overview. MAPA specifications are built from process terms, that are given by the following grammar.

Definition 8 (Process terms). *A* process term *in MAPA is any term that can be generated by the following grammar:*

$$p ::= Y(\boldsymbol{t}) \;\mid\; c \Rightarrow p \;\mid\; p + p \;\mid\; \sum_{\boldsymbol{x}:\boldsymbol{D}} p \;\mid\; (\lambda) \cdot p \;\mid\; a(\boldsymbol{t})[r] \sum_{\boldsymbol{x}:\boldsymbol{D}} f : p$$

Here, $Y(\boldsymbol{t})$ denotes *process instantiation* of process $Y(\boldsymbol{t})$. The term $c \Rightarrow p$ behaves as p if the *condition* c holds, and cannot do anything otherwise. The + operator

denotes *nondeterministic choice*, and $\sum_{\boldsymbol{x}:\boldsymbol{D}} p$ a (possibly infinite) *nondeterministic choice over data type* $\boldsymbol{D}$. Finally, $(\lambda)\cdot p$ behaves as p after a delay, determined by a negative exponential distribution with rate λ.

The term $a(\boldsymbol{t})[r]\sum_{\boldsymbol{x}:\boldsymbol{D}} f : p$ performs the action $a(\boldsymbol{t})$ while obtaining reward r, and then has a *probabilistic choice* over $\boldsymbol{D}$. It uses the value f (with $\boldsymbol{x}$ substituted by $\boldsymbol{d}$) as the probability of choosing each $\boldsymbol{d} \in \boldsymbol{D}$. This extension to MAPA can be used to specify action-based rewards on probabilistic transitions.

Example 1. The grammar in Definition 8 provides the MAPA language with an infinite number of process terms. One of these is

$$\sum_{n:\mathbb{N}} n < 3 \Rightarrow (2 \cdot n + 1) \cdot \mathit{send}(n)[2] \sum_{x:\{1,2\}} \frac{x}{3} : (Y(n+x) + Z(n+x))$$

For the expression $t = \frac{x}{3}$ we find $t[x := 2] = \frac{2}{3}$, and for the process term $p' = Y(x) + Z(x)$ we find $p'[x := 2] = Y(2) + Z(2)$. The semantics of this process term is as follows: (1) The variable n nondeterministically gets assigned any natural number; (2) If $n < 3$, then the process continues with a delay, governed by an exponential distribution with rate $2 \cdot n + 1$; (3) The process does the action *send*, parameterised by the number n that was chosen earlier and obtains a reward of 2; (4) Probabilistically, x gets assigned a value from the set $\{1, 2\}$. Each value x has probability $\frac{x}{3}$ to be chosen, so 1 has probability $\frac{1}{3}$ and 2 has with probability $\frac{2}{3}$. Note that, as expected and also required by the formal semantics, these probabilities add up to 1; (5) Nondeterministically, the behaviour continues as either $Y(n+x)$ or $Z(n+x)$, with the value chosen nondeterministically in the first step substituted for n and the value chosen probabilistically in the previous step substituted for x.

Combining all these steps, this yields the MA given in Figure 1, where each state t_i behaves as $Y(i) + Z(i)$. The behaviour of these processes can be specified separately.

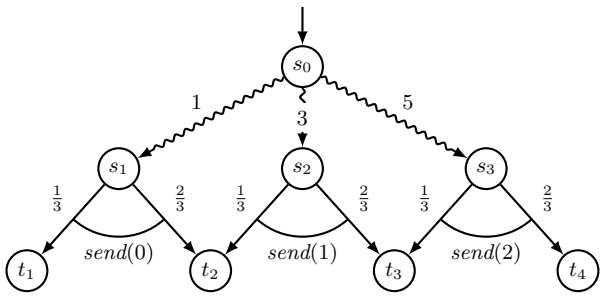

Fig. 1. Semantics of the first process term of Example 1.

4.1 MAMA extensions

Since realistic systems often consist of a very large number of states, we do not want to construct their MRA models manually. Rather, we prefer to specify them as the parallel composition of multiple components. This approach was applied earlier to generate MAs, using a tool called SCOOP [31,18,32]. It generates MAs from MAPA specifications, applying several reduction techniques in the process. The underlying philosophy is to already reduce on the specification, not having to first generate a large model before being able to minimise. The parallel composition of MRAs is described in the appendix and is equivalent to [11] for the probabilistic transitions.

We extended SCOOP to parse action-based and state-based rewards. Action-based rewards are stored as part of the transitions, while state-based rewards are represented internally by self-loops. Additionally, we generalised most of its reduction techniques to take into account the new rewards. The following reduction techniques are now also applicable to MRAs:

Dead variable reduction. This technique resets variables if their value is not needed anymore until they are overwritten. Instead of only checking whether a variable is used in conditions or actions, we generalised this technique to also check if it is used in reward expressions.

Maximal progress reduction. This technique removes Markovian transitions from states also having τ-transitions. It can be applied unchanged to MRAs.

Basic reduction techniques. The basic reduction techniques omit variables that are never changed, omit nondeterministic choices that only have one option and simplify expressions where possible. These three techniques were easily generalised by taking the reward expressions into account as well.

Confluence reduction was not yet generalised, as it is based on a much more complicated notion of bisimulation (that is not yet available for MRAs).

SCOOP takes both the action-based and state-based rewards into account when generating an input file for the IMCA toolset. This toolset implements several algorithms for computing reward-based properties, as detailed before. The connection of the tool-chain is depicted in Figure 2.

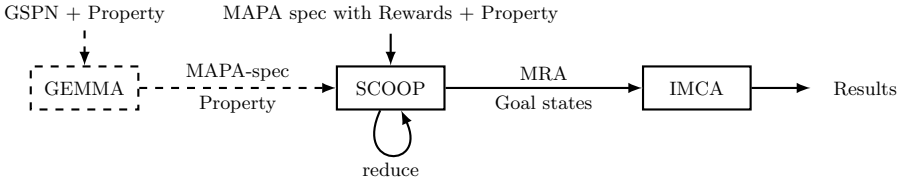

Fig. 2. Analysing Markov Reward Automata using the MAMA tool chain.

$$
\begin{array}{l}
\textbf{constant } \mathit{queueSize} = Q, \mathit{nrOfJobTypes} = N \\
\textbf{type } \mathit{Stations} = \{1,2\},\ \ \mathit{Jobs} = \{1,\ldots,\mathit{nrOfJobTypes}\} \\
\\
\mathit{Station}(i : \mathit{Stations}, q : \text{Queue}, \mathit{size} : \{0..\mathit{queueSize}\}) \\
\quad = \mathit{size} < \mathit{queueSize} \Rightarrow (2i+1) \cdot \textstyle\sum_{j:\mathit{Jobs}} \mathit{arrive}(j) \cdot \mathit{Station}(i, \text{enqueue}(q,j), \mathit{size}+1) \\
\quad + \mathit{size} > 0 \qquad\quad \Rightarrow \mathit{deliver}(i, \text{head}(q)) \cdot \sum_{k \in \{1,9\}} \tfrac{k}{10} : k = 1 \Rightarrow \mathit{Station}(i, q, \mathit{size}) \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad + k = 9 \Rightarrow \mathit{Station}(i, \text{tail}(q), \mathit{size}-1) \\
\\
\mathit{Server} = \textstyle\sum_{n:\mathit{Stations}} \sum_{j:\mathit{Jobs}} \mathit{poll}(n,j)\underline{[0.1]} \cdot (2 * j) \cdot \mathit{finish}(j) \cdot \mathit{Server} \\
\\
\gamma(\mathit{poll}, \mathit{deliver}) = \mathit{copy} \qquad \text{// actions } \mathit{poll} \text{ and } \mathit{deliver} \text{ synchronise and yield action } \mathit{copy} \\
\mathit{System} = \tau_{\{\mathit{copy}, \mathit{arrive}, \mathit{finish}\}}(\partial_{\{\mathit{poll}, \mathit{deliver}\}}(\mathit{Station}(1, \text{empty}, 0) \,||\, \mathit{Station}(2, \text{empty}, 0) \,||\, \mathit{Server})) \\
\\
\underline{\textbf{state reward } \texttt{true} \rightarrow \ \mathit{size}_1 * 0.01 + \mathit{size}_2 * 0.01}
\end{array}
$$

Fig. 3. MAPA specification of a nondeterministic polling system.

5 Case studies

To assess the performance of the algorithms and implementation, we provide two case studies: A server polling system based on [30], and a fault-tolerant workstation cluster based on [23]. Rewards were added to both examples. The experiments were conducted on a 2.2 GHz Intel® Core™ i7-2670QM processor with 8 GB RAM, running Linux.

Polling system. Figure 3 shows the MAPA specification of the polling system. It consists of two stations, each providing a job queue, and one server. When the server polls a job from a station, there is a 10% chance that it will erroneously remain in the queue. An impulse reward of 0.1 is given each time a server takes a job, and a reward of 0.01 per time unit is given for each job in the queue. The rewards are meant to be interpreted as costs in this example, for having a job processed and for taking up server memory, respectively.

Tables 2 and 4 show the results obtained by the MAMA tool-chain when analysing for different queue sizes Q and different numbers of job types N. The goal states for the expected reward are those when both queues are full. The error-bound for the time-bounded reward analysis was set to 0.1.

The tables show that the minimal reward does not depend on the number of job types, while the maximal reward does. The long-run reward computation is, for this example, considerably slower than the expected reward, and both increase more than linear with the number of states. The time-bounded reward is more affected by the time bound than the number of states, and the computation time does not significantly differ between the maximal and minimal queries.

Workstation cluster. The second case study is based on a fault-tolerant workstation cluster, described as a GSPN in [26]. Using the GEMMA [2] tool, the GSPN was converted into a MAPA specification.

The workstation cluster consists of two groups of N workstations, each group connected by one switch. The two groups are connected to each other by a backbone. Workstations, switches and the backbone experience exponentially

Q	N	T_{lim}	Time-bounded reward min	T(min)	max	T(max)
2	3	1	0.626	0.46	0.814	0.46
2	3	2	0.914	1.64	1.389	1.66
2	3	10	1.005	161.73	2.189	166.59
3	3	1	0.681	4.90	0.893	4.75
3	3	2	1.121	16.69	1.754	17.11
3	3	10	1.314	1653	4.425	1687

Table 2. Time-bounded rewards for the polling system (T in seconds).

N	Q	T_{lim}	Time-bounded reward min	T(min)	max	T(max)
4	3	10	0.0467	1.16	0.0467	1.09
4	3	20	0.0968	8.54	0.0968	8.23
4	3	50	0.2481	125.4	0.2481	123.6
4	3	100	0.5004	989.8	0.5004	991.8
4	5	10	0.0454	1.276	0.0454	1.209
4	5	20	0.0929	9.297	0.0929	9.218
4	5	50	0.2333	141.9	0.2333	143.8
4	5	100	0.4610	1123	0.4610	1132

Table 3. Time-bounded rewards for the workstation cluster (T in seconds).

Q	N	$\lvert S\rvert$	$\lvert G\rvert$	Long-run reward min	T(min)	max	T(max)	Expected reward min	T(min)	max	T(max)
2	3	1159	405	0.731	0.61	1.048	0.43	0.735	0.28	2.110	0.43
2	4	3488	1536	0.731	3.76	1.119	2.21	0.735	0.93	3.227	2.01
3	3	11122	3645	0.750	95.60	1.107	19.14	1.034	3.14	4.752	8.14
3	4	57632	24576	0.750	5154.6	1.198	705.8	1.034	31.80	8.878	95.87
4	2	5706	1024	0.769	38.03	0.968	5.73	1.330	3.12	4.199	3.12
4	3	102247	32805	Timeout(2h)				1.330	63.24	9.654	192.18

Table 4. Long-run and expected rewards for the polling system (T in seconds).

distributed failures, and can be repaired one at a time. If multiple components are eligible for repair at the same time, the choice is nondeterministic. The overall cluster is considered operational if at least Q workstations are operational and connected to each other. Rewards have been added to the system to simulate the costs of repairs and downtime. Repairing a workstation has cost 0.3, a switch costs 0.1, and the backbone costs 1 to repair. If fewer than Q workstations are operational and connected, a cost of 1 per unit time is incurred.

Tables 3 and 5 show the analysis results for this example. The goal states for the expected reward are the states where not enough operational workstations are connected. The error bound for the time-bounded reward analysis was 0.1. For this example, the long-run rewards are quicker to compute than the expected rewards. The long-run rewards do not vary much with the scheduler, since multiple simultaneous failures are rare in this system. This also explains the large expected rewards when Q is low: many repairs will occur before the cluster fails. The time-bounded rewards also show almost no dependence on the scheduler.

6 Conclusions and future work

We introduced the Markov Reward Automaton (MRA), an extension of the Markov Automaton (MA) featuring both state-based and action-based rewards (or, equivalently, costs). We defined strong bisimulation for MRAs, and validated it by stating that our notion coincides with the traditional notions of strong

				Long-run reward				Expected reward			
N	Q	$\|S\|$	$\|G\|$	min	T(min)	max	T(max)	min	T(min)	max	T(max)
4	3	1439	1008	0.00504	0.0272	0.00505	0.143	5335	337.9	5348	297.1
4	5	1439	621	0.00857	0.00787	0.00864	0.217	6.848	0.4111	6.848	0.4095
4	8	1439	1438	0.01655	0.00709	0.0166	0.182	0	0.00019	0	0.00018
8	6	4876	3584	0.00983	0.258	0.00984	1.875	16460	4502	16514	4124
8	8	4876	4415	0.00997	0.0920	0.0100	1.992	254.0	55.57	254.0	53.73
8	10	4883	4783	0.0134	0.0463	0.0134	2.064	13.70	2.941	13.70	2.904
8	16	4895	4894	0.0294	0.0351	0.0294	2.134	0	0.00059	0	0.00061

Table 5. Long-run and expected rewards for the workstation cluster (T in seconds).

bisimulation for MAs. We generalised the MAPA language to efficiently model MRAs by process-algabraic specifications, and extended the SCOOP tool to automatically generate MRAs from these specifications. Furthermore, we presented three algorithms, for computing the expecte reward until reaching a set of goal states, for computing the expected reward until reaching a time-bound, and for computing the long-run average reward while visiting a set of states. Our modelling framework and algorithms allow for a wide variety of systems—featuring nondeterminism, discrete probabilistic choice, continuous stochastic timing and action-based and state-based rewards—to be efficiently modelled, generated and analysed.

Future work will focus on developing weak notions of bisimulation for MRAs, possibly allowing the generalisation of confluence reduction. For quantitative analysis, future work will focus on considering negative rewards, optimisations with respect to time and reward-bounded reachability properties, as well as the handling of several rewards as multi-optimisation problems.

Acknowledgement. This work has been supported by the NWO project SYRUP (612.063.817), by the STW-ProRail partnership program ExploRail under the project ArRangeer (12238), by the DFG/NWO bilateral project ROCKS (DN 63-257), by the German Research Council (DFG) as part of the Transregional Collaborative Research Center "Automatic Verification and Analysis of Complex Systems" (SFB/TR 14 AVACS), and by the European Union Seventh Framework Programme under grant agreement no. 295261 (MEALS) and 318490 (SENSATION). We would like to thank Joost-Pieter Katoen for the fruitful discussions.

References

1. S. Andova, H. Hermanns, and J.-P. Katoen. Discrete-time rewards model-checked. In *FORMATS*, volume 2791 of *LNCS*, pages 88–104. Springer, 2003.
2. R. Bamberg. Non-deterministic generalised stochastic Petri nets modelling and analysis. Master's thesis, University of Twente, 2012.
3. M. Bernardo. An algebra-based method to associate rewards with EMPA terms. In *ICALP*, volume 1256 of *LNCS*, pages 358–368. Springer, 1997.

4. H. Boudali, P. Crouzen, and M. I. A. Stoelinga. A rigorous, compositional, and extensible framework for dynamic fault tree analysis. *IEEE Transactions on Dependable and Secure Computing*, 7(2):128–143, 2010.
5. M. Bozzano, A. Cimatti, J.-P. Katoen, V. Y. Nguyen, T. Noll, and M. Roveri. Safety, dependability and performance analysis of extended AADL models. *The Computer Journal*, 54(5):754–775, 2011.
6. B. Braitling, L. M. F. Fioriti, H. Hatefi, R. Wimmer, B. Becker, and H. Hermanns. MeGARA: Menu-based game abstraction and abstraction refinement of Markov automata. In *QAPL*, volume 154 of *EPTCS*, pages 48–63, 2014.
7. T. Brazdil, V. Brozek, K. Chatterjee, V. Forejt, and A. Kucera. Two views on multiple mean-payoff objectives in Markov decision processes. In *LICS*, pages 33–42. IEEE, 2011.
8. K. Chatterjee and M. Henzinger. Faster and dynamic algorithms for maximal end-component decomposition and related graph problems in probabilistic verification. In *SODA*, pages 1318–1336. SIAM, 2011.
9. G. Clark. Formalising the specification of rewards with PEPA. In *PAPM*, pages 139–160, 1996.
10. L. de Alfaro. *Formal Verification of Probabilistic Systems*. PhD thesis, Stanford University, 1997.
11. Y. Deng and M. Hennessy. Compositional reasoning for weighted Markov decision processes. *Science of Computer Programming*, 78(12):2537 – 2579, 2013. Special Section on International Software Product Line Conference 2010 and Fundamentals of Software Engineering (selected papers of FSEN 2011).
12. Y. Deng and M. Hennessy. On the semantics of Markov automata. *Information and Computation*, 222:139–168, 2013.
13. C. Eisentraut, H. Hermanns, J.-P. Katoen, and L. Zhang. A semantics for every GSPN. In *ICATPN*, volume 7927 of *LNCS*, pages 90–109. Springer, 2013.
14. C. Eisentraut, H. Hermanns, and L. Zhang. Concurrency and composition in a stochastic world. In *CONCUR*, volume 6269 of *LNCS*, pages 21–39. Springer, 2010.
15. C. Eisentraut, H. Hermanns, and L. Zhang. On probabilistic automata in continuous time. In *LICS*, pages 342–351. IEEE, 2010.
16. J. F. Groote and A. Ponse. The syntax and semantics of μCRL. In *ACP*, Workshops in Computing, pages 26–62. Springer, 1995.
17. D. Guck, T. Han, J.-P. Katoen, and M. R. Neuhäußer. Quantitative timed analysis of interactive Markov chains. In *NFM*, volume 7226 of *LNCS*, pages 8–23. Springer, 2012.
18. D. Guck, H. Hatefi, H. Hermanns, J.-P. Katoen, and M. Timmer. Modelling, reduction and analysis of Markov automata. In *QEST*, volume 8054 of *LNCS*, pages 55–71. Springer, 2013.
19. D. Guck, M. Timmer, H. Hatefi, E. Ruijters, and M. Stoelinga. Extending Markov automata with state and action rewards (extended version). Technical Report TR-CTIT-14-06, CTIT, University of Twente, Enschede, 2014.
20. D. Guck, M. Timmer, H. Hatefi, E. J. J. Ruijters, and M. I. A. Stoelinga. Modelling and analysis of Markov reward automata. In *ATVA*, to appear in LNCS. Springer, 2014.
21. H. Hatefi and H. Hermanns. Model checking algorithms for Markov automata. *Electronic Communications of the EASST*, 53, 2012.
22. B. R. Haverkort, L. Cloth, H. Hermanns, J.-P. Katoen, and C. Baier. Model checking performability properties. In *DSN*, pages 103–112. IEEE, 2002.

23. B. R. Haverkort, H. Hermanns, and J.-P. Katoen. On the use of model checking techniques for dependability evaluation. In *SRDS*, pages 228–237. IEEE, 2000.
24. H. Hermanns. *Interactive Markov Chains: The Quest for Quantified Quality*, volume 2428 of *LNCS*. Springer, 2002.
25. J.-P. Katoen, I. S. Zapreev, E. M. Hahn, H. Hermanns, and D. N. Jansen. The ins and outs of the probabilistic model checker MRMC. *Performance Evaluation*, 68(2):90–104, 2011.
26. M. R. Neuhäußer. *Model Checking Nondeterministic and Randomly Timed Systems*. PhD thesis, University of Twente, 2010.
27. M. R. Neuhäußer, M. I. A. Stoelinga, and J.-P. Katoen. Delayed nondeterminism in continuous-time Markov decision processes. In *FOSSACS*, volume 5504 of *LNCS*, pages 364–379. Springer, 2009.
28. R. Segala. *Modeling and Verification of Randomized Distributed Real-Time Systems*. PhD thesis, Massachusetts Institute of Technology, 1995.
29. L. Song, L. Zhang, and J. C. Godskesen. Late weak bisimulation for Markov automata. Technical report, ArXiv e-prints, 2012.
30. M. M. Srinivasan. Nondeterministic polling systems. *Management Science*, 37(6):667–681, 1991.
31. M. Timmer. SCOOP: A tool for symbolic optimisations of probabilistic processes. In *QEST*, pages 149–150. IEEE, 2011.
32. M. Timmer. *Efficient Modelling, Generation and Analysis of Markov Automata*. PhD thesis, University of Twente, 2013.
33. M. Timmer, J.-P. Katoen, J. C. van de Pol, and M. I. A. Stoelinga. Efficient modelling and generation of Markov automata. In *CONCUR*, volume 7454 of *LNCS*, pages 364–379. Springer, 2012.
34. M. Timmer, M. I. A. Stoelinga, and J. C. van de Pol. Confluence reduction for Markov automata. In *FORMATS*, volume 8053 of *LNCS*, pages 243–257. Springer, 2013.
35. J. C. van de Pol and M. Timmer. State space reduction of linear processes using control flow reconstruction. In *ATVA*, volume 5799 of *LNCS*, pages 54–68. Springer, 2009.
36. R. Wunderling. *Paralleler und objektorientierter Simplex-Algorithmus*. PhD thesis, Technische Universität Berlin, 1996.

A Proof of Theorem 1

Theorem 1 (Bellman equation) *The function* $\mathsf{eR}^{\max}\colon S \to \mathbb{R}^{\infty}_{\geq 0}$ *is the unique fixed point of the Bellman equation*

$$v(s) = \begin{cases} \frac{\rho(s)}{E(s)} + \sum\limits_{s' \in S} \mathbb{P}_s(s') \cdot (v(s') + r_s) & \text{if } s \in MS \setminus G \\ \max\limits_{\alpha \in A(s)} \sum\limits_{s' \in S} \mu_s^{\alpha}(s') \cdot (v(s') + r_s^{\alpha}) & \text{if } s \in PS \setminus G \\ 0 & \text{if } s \in G. \end{cases}$$

Proof. We show that Theorem 1 and Equation 2 coincide. Therefore, we will distinguish three cases: $s \in MS \setminus G$, $s \in PS \setminus G$, and $s \in G$.

– $s \in MS \setminus G$:

$$
\begin{aligned}
\mathsf{eR}^{\max}(s,G) &= \sup_{D \in GM} \mathbb{E}_{s,D}(V_G) = \sup_{D \in GM} \int_{paths} V_G(\pi) \Pr_{s,D}(\mathrm{d}\pi) \\
&= \inf_{D} \int_{paths} reward(\pi) \cdot \Pr_{s,D}(\mathrm{d}\pi) = \sup_{D \in GM} \int_{paths} \Big(\sum_{i=0}^{|\pi|-1} \rho(\pi[i]) \cdot t_i + r^{\alpha_i}_{\pi[i]} \Big) \cdot \Pr_{s,D}(\mathrm{d}\pi) \\
&= \sup_{D \in GM} \int_{paths} \left(\rho(\pi[0]) \cdot t_0 + r^{\alpha_0}_{\pi[0]} + \left(\sum_{i=1}^{|\pi|-1} \rho(\pi[i]) \cdot t_i + r^{\alpha_i}_{\pi[i]} \right) \right) \cdot \Pr_{s,D}(\mathrm{d}\pi) \\
&= \sup_{D \in GM} \int_0^\infty \rho(s) \cdot t \cdot E(s) \cdot \mathrm{e}^{-E(s)t} + r_s + \sum_{s' \in S} \mathbb{P}_s(s') \cdot \mathbb{E}_{s', D[s \xrightarrow{\perp, \mathbb{P}_s(\cdot), t} s']}(V_G)\, \mathrm{d}t \\
&= \sup_{D \in GM} \left(\int_0^\infty \rho(s) \cdot t \cdot E(s) \cdot \mathrm{e}^{-E(s)t} + r_s \, \mathrm{d}t \right. \\
&\quad \left. + \int_0^\infty \sum_{s' \in S} \mathbb{P}_s(s') \cdot \mathbb{E}_{s', D[s \xrightarrow{\perp, \mathbb{P}_s(\cdot), t} s']}(V_G)\, \mathrm{d}t \right) \\
&= \sup_{D \in GM} \left(\frac{\rho(s)}{E(s)} + r_s + \sum_{s' \in S} \mathbb{P}_s(s') \cdot \int_0^\infty \mathbb{E}_{s', D[s \xrightarrow{\perp, \mathbb{P}_s(\cdot), t} s']}(V_G)\, \mathrm{d}t \right) \\
&= \frac{\rho(s)}{E(s)} + r_s + \sup_{D \in GM} \sum_{s' \in S} \mathbb{P}_s(s') \cdot \int_0^\infty \mathbb{E}_{s', D[s \xrightarrow{\perp, \mathbb{P}_s(\cdot), t} s']}(V_G)\, \mathrm{d}t \\
&= \frac{\rho(s)}{E(s)} + r_s + \sup_{D \in GM} \sum_{s' \in S} \mathbb{P}_s(s') \cdot \mathbb{E}_{s', D}(V_G) \\
&= \frac{\rho(s)}{E(s)} + r_s + \sum_{s' \in S} \mathbb{P}_s(s') \cdot \sup_{D \in GM} \mathbb{E}_{s', D}(V_G) \\
&= \frac{\rho(s)}{E(s)} + r_s + \sum_{s' \in S} \mathbb{P}_s(s') \cdot \mathsf{eR}^{\max}(s', G) \\
&= \frac{\rho(s)}{E(s)} + \sum_{s' \in S} \mathbb{P}_s(s') \cdot \left(\mathsf{eR}^{\max}(s', G) + r_s \right) \\
&= v(s).
\end{aligned}
$$

where $D[s \xrightarrow{\perp, \mathbb{P}_s(\cdot), t} s']$ is the policy that resolves nondeterminism for path π' starting from s' as D does it for $s \xrightarrow{\perp, \mathbb{P}_s(\cdot), t} \pi'$, i.e. $D(s \xrightarrow{\perp, \mathbb{P}_s(\cdot), t} \pi') = D[s \xrightarrow{\perp, \mathbb{P}_s(\cdot), t} s'](\pi')$.

– $s \in PS \setminus G$:

$$
\begin{aligned}
\mathsf{eR}^{\max}(s,G) &= \sup_{D \in GM} \mathbb{E}_{s,D}(V_G) = \sup_{D \in GM} \int_{paths} V_G(\pi) \Pr_{s,D}(\mathrm{d}\pi) \\
&= \sup_{D \in GM} \sum_{s \xrightarrow{\alpha, \mu^\alpha_s, 0} s'} D(s)(\alpha, \mu) \cdot \mu^\alpha_s(s') \cdot \left(\mathbb{E}_{s, D[s \xrightarrow{\alpha, \mu^\alpha_s, 0} s']}(V_G) + r^\alpha_s \right)
\end{aligned}
$$

where $D[s \xrightarrow{\alpha, \mu^\alpha_s, 0} s']$ is the policy that resolves nondeterminism for path π' starting from s' as D does it for $s \xrightarrow{\alpha, \mu, 0} \pi'$, i.e. $D(s \xrightarrow{\alpha, \mu^\alpha_s, 0} \pi') = D[s \xrightarrow{\alpha, \mu^\alpha_s, 0} s'](\pi')$. Each action $\alpha \in A(s)$ uniquely determines a distribution μ^α_s, such that the successor state s', with $s \xrightarrow{\alpha, \mu^\alpha_s, 0} s'$, satisfies $\mu^\alpha_s(s') > 0$:

$$
\alpha^\star = \arg\max \left\{ \sup_{D \in GM} \sum_{s' \in S} \mu^\alpha_s(s') \cdot \mathbb{E}_{s', D}(V_G) \mid \alpha \in A(s) \right\}
$$

Hence, all optimal policies choose $\alpha^\star$ with probability 1, i.e. $D(s)(\alpha^\star, \mu_s^{\alpha^*}) = 1$ and $D(s)(\beta, \mu_s^\beta) = 0$ for all $\beta \neq \alpha^\star$. Thus, we obtain

$$\begin{aligned}
\mathsf{eR}^{\max}(s,G) &= \sup_{D\in GM} \max_{s \xrightarrow{\alpha} \mu_s^\alpha} \sum_{s'\in S} \mu_s^\alpha(s') \cdot \left(\mathbb{E}_{s,D[s\xrightarrow{\alpha,\mu_s^\alpha,0}s']}(V_G) + r_s^\alpha \right) \\
&= \max_{s \xrightarrow{\alpha} \mu_s^\alpha} \sup_{D\in GM} \sum_{s'\in S} \mu_s^\alpha(s') \cdot \left(\mathbb{E}_{s,D[s\xrightarrow{\alpha,\mu_s^\alpha,0}s']}(V_G) + r_s^\alpha \right) \\
&= \max_{s \xrightarrow{\alpha} \mu_s^\alpha} \sup_{D\in GM} \sum_{s'\in S} \mu_s^\alpha(s') \cdot \left(\mathbb{E}_{s',D}(V_G) + r_s^\alpha \right) \\
&= \max_{s \xrightarrow{\alpha} \mu_s^\alpha} \sum_{s'\in S} \mu_s^\alpha(s') \cdot \left(\sup_{D\in GM} \mathbb{E}_{s',D}(V_G) + r_s^\alpha \right) \\
&= \max_{s \xrightarrow{\alpha} \mu_s^\alpha} \sum_{s'\in S} \mu_s^\alpha(s') \cdot \left(\mathsf{eR}^{\max}(s,G) + r_s^\alpha \right) \\
&= \max_{\alpha\in A(s)} \sum_{s'\in S} \mu_s^\alpha(s') \cdot \left(\mathsf{eR}^{\max}(s,G) + r_s^\alpha \right) \\
&= v(s).
\end{aligned}$$

– $s \in G$:

$$\begin{aligned}
\mathsf{eR}^{\max}(s,G) &= \sup_{D\in GM} \mathbb{E}_{s,D}(V_G) = \sup_{D\in GM} \int_{paths} V_G(\pi) \Pr_{s,D}(d\pi) \\
&= 0 \\
&= v(s)
\end{aligned}$$

□

B Proof of Lemma 2

We prove the lemma in two steps. First we show that $\mathcal{R}^{\max}$ is the fixed point of the operator Ω described in Lemma 2. Then we show that it is the least fixed point. We recall the definition of maximum time-bounded expected reward. Given a Markov reward automaton $\mathcal{M}$, a time bound $b \geq 0$ and $s \in S$, the maximum time-bounded expected reward is define as:

$$\mathcal{R}^{\max}(s,b) = \sup_{D\in GM} \int_{paths} \mathit{reward}(\pi, b) \Pr_{s,D}(d\pi) \tag{10}$$

We distinguish between three cases. The trivial case is when $s \in MS$ and $b = 0$ since $\Omega(\mathcal{R}^{\max})(s,0) = \mathcal{R}^{\max}(s,0) = 0$. Then we consider the case $s \in MS$ and $b > 0$. We represent each path starting from s by splitting it at the point it leaves s and write it as $\pi = s \xrightarrow{\chi(E(s)),\mathbb{P}_s,t} \pi'$. We can therefore split the reward and the infinitesimal term of Eq. (10) accordingly:

$$\mathit{reward}(\pi, b) = \begin{cases} \rho(s)t + r_s + \mathit{reward}(\pi', b-t) & t \leq b \\ \rho(s)b & t > b \end{cases}$$

$$\Pr_{s,D}(\,\mathrm{d}\pi) = E(s)\mathrm{e}^{-E(s)t} \cdot \,\mathrm{d}t \cdot \sum_{s'\in S} \mathbb{P}_s(s') \Pr_{s',D_t}(\,\mathrm{d}\pi')$$

where D_t is the scheduler that resolves nondeterminisim for any prefix ζ of π' as D does it for $s \xrightarrow{\chi(E(s)),\mathbb{P}_s,t} \zeta$, i. e. $D_t(\zeta) = D(s \xrightarrow{\chi(E(s)),\mathbb{P}_s,t} \zeta)$. Plugging the above equations into Eq. (10) gives:

$$
\begin{aligned}
\mathcal{R}^{\max}(s,b) = \sup_{D\in GM} \Bigg(& \int_0^b \int_{\pi'\in paths} \big(\rho(s)t + r_s + reward(\pi', b-t)\big)\, E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \\
& + \int_b^\infty \int_{\pi'\in paths} \rho(s) b E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \Bigg) \\
= \sup_{D\in GM} \Bigg(& \int_0^b \int_{\pi'\in paths} (\rho(s)t + r_s) E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \\
& + \int_0^b \int_{\pi'\in paths} reward(\pi', b-t) E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \\
& + \int_b^\infty \int_{\pi'\in paths} \rho(s) b E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \Bigg) \\
= \sup_{D\in GM} \Bigg(& \int_0^b (\rho(s)t + r_s) E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \int_{\pi'\in paths} \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \\
& + \int_0^b E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \int_{\pi'\in paths} reward(\pi', b-t) \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \\
& + \int_b^\infty \rho(s) b E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \int_{\pi'\in paths} \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \Bigg) \\
\overset{(*)}{=} \sup_{D\in GM} \Bigg(& \int_0^b (\rho(s)t + r_s) E(s) \mathrm{e}^{-E(s)t}\,\mathrm{d}t + \int_b^\infty \rho(s) b E(s) \mathrm{e}^{-E(s)t}\,\mathrm{d}t \\
& + \int_0^b E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \int_{\pi'\in paths} reward(\pi', b-t) \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \Bigg) \\
= \sup_{D\in GM} \Bigg(& \Big(r_s + \frac{\rho(s)}{E(s)}\Big)\Big(1 - e^{-E(s)b}\Big) \\
& + \int_0^b E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \int_{\pi'\in paths} reward(\pi', b-t) \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \Bigg) \\
= & \Big(r_s + \frac{\rho(s)}{E(s)}\Big)\Big(1 - e^{-E(s)b}\Big) \\
& + \sup_{D\in GM} \Bigg(\int_0^b E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \int_{\pi'\in paths} reward(\pi', b-t) \Pr_{s',D_t}(\,\mathrm{d}\pi')\,\mathrm{d}t \Bigg) \\
= & \Big(r_s + \frac{\rho(s)}{E(s)}\Big)\Big(1 - e^{-E(s)b}\Big) \\
& + \int_0^b E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \sup_{D\in GM} \Bigg(\int_{\pi'\in paths} reward(\pi', b-t) \Pr_{s',D_t}(\,\mathrm{d}\pi') \Bigg)\,\mathrm{d}t \\
\overset{(**)}{=} & \Big(r_s + \frac{\rho(s)}{E(s)}\Big)\Big(1 - e^{-E(s)b}\Big) + \int_0^b E(s) \mathrm{e}^{-E(s)t} \sum_{s'\in S} \mathbb{P}_s(s') \mathcal{R}^{\max}(s', b-t)\,\mathrm{d}t
\end{aligned}
$$

where $(*)$ and $(**)$ respectively follow from the facts $\int_{\pi'\in paths} \Pr_{s',D_t}(\,\mathrm{d}\pi') = 1$ and $\sup_{D\in GM} \int_{\pi'\in paths} reward(\pi', b-t) \Pr_{s',D_t}(\,\mathrm{d}\pi') = \mathcal{R}^{\max}(s', b-t)$ for any fixed time point $t \leq b$.

We use similar decompositions for the remaining situation where $s \in PS$. Note that in this case closeness implies immediately leaving of s. Therefore we

describe each path π starting from s by $s \xrightarrow{\alpha,\mu_s^\alpha,0} \pi'$. It accordingly yields:

$$\begin{aligned} \mathit{reward}(\pi, b) &= r_s^\alpha + \mathit{reward}(\pi', b) \\ \Pr_{s,D}(\,\mathrm{d}\pi) &= \sum_{s' \in S} \mu_s^\alpha \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') \end{aligned}$$

where D_α is the policy that makes the same decision for any finite path ζ as D does it for $s \xrightarrow{\alpha,\mu_s^\alpha,0} \zeta$, i. e. $D_\alpha(\zeta) = D(s \xrightarrow{\alpha,\mu_s^\alpha,0} \zeta)$. Obviously the maximum reward happens when the policy make a pure decision of an action. This fact along with the above equations provides:

$$\begin{aligned} \mathcal{R}^{\max}(s,b) &= \sup_{D \in GM} \max_{\alpha \in A} \left(\int_{\pi' \in paths} (r_s^\alpha + \mathit{reward}(\pi', b)) \sum_{s' \in S} \mu_s^\alpha(s') \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') \right) \\ &= \max_{\alpha \in A} \sup_{D \in GM} \left(\int_{\pi' \in paths} (r_s^\alpha + \mathit{reward}(\pi', b)) \sum_{s' \in S} \mu_s^\alpha(s') \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') \right) \\ &= \max_{\alpha \in A} \sup_{D \in GM} \left(\int_{\pi' \in paths} r_s^\alpha \sum_{s' \in S} \mu_s^\alpha(s') \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') \right. \\ &\quad \left. + \int_{\pi' \in paths} \mathit{reward}(\pi', b) \sum_{s' \in S} \mu_s^\alpha(s') \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') \right) \\ &= \max_{\alpha \in A} \sup_{D \in GM} \left(r_s^\alpha \sum_{s' \in S} \mu_s^\alpha(s') \int_{\pi' \in paths} \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') \right. \\ &\quad \left. + \sum_{s' \in S} \mu_s^\alpha(s') \int_{\pi' \in paths} \mathit{reward}(\pi', b) \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') \right) \\ &\overset{(\dagger)}{=} \max_{\alpha \in A} \sup_{D \in GM} \left(r_s^\alpha + \sum_{s' \in S} \mu_s^\alpha(s') \int_{\pi' \in paths} \mathit{reward}(\pi', b) \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') \right) \\ &= \max_{\alpha \in A} \left(r_s^\alpha + \sum_{s' \in S} \mu_s^\alpha(s') \sup_{D \in GM} \left(\int_{\pi' \in paths} \mathit{reward}(\pi', b) \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') \right) \right) \\ &\overset{(\ddagger)}{=} \max_{\alpha \in A} \left(r_s^\alpha + \sum_{s' \in S} \mu_s^\alpha(s') \mathcal{R}^{\max}(s', b) \right) \end{aligned}$$

where (†) and (‡) respectively follow from the facts that $\int_{\pi' \in paths} \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') = 1$ and $\int_{\pi' \in paths} \mathit{reward}(\pi', b) \Pr_{s',D_\alpha}(\,\mathrm{d}\pi') = \mathcal{R}^{\max}(s', b)$ for a fixed $\alpha \in A$.

The second part of the proof shows that $\mathcal{R}^{\max}$ is the least fixed point of the characterisation given in 2. Here we employ the same technique as used in [26, Theorem 5.1]. Let F be any fixed point of the characterisation, we show that $\mathcal{R}^{\max}(s,b) \leq F(s,b)$ for all $s \in S$ and $b \geq 0$. We show by Ω^n $(n > 0)$ the n-level recursive composition of operator Ω and write $F_n = \Omega^n(F_0)$, where F_0 is the starting bottom function. For maximum time-bounded expected reward, $\mathcal{R}_n^{\max}(s,b)$ intuitively refers to the reward gained within time-bound b by taking paths up to length n, as each composition of the operator take one probabilistic or Markovian step into reward computaion. Its bottom function is thus defined as $\mathcal{R}_0^{\max}(s,b) = 0$ for all $s \in S$ and $b \geq 0$. We show by induction that $\forall n \geq 0, \forall s \in S, \forall b \geq 0. \mathcal{R}_n^{\max}(s,b) \leq F(s,b)$. For $n = 0$ it is true, since $\mathcal{R}_0^{\max}$ is minimal and thus $\mathcal{R}_0^{\max}(s,b) \leq F(s,b)$ for all $s \in S$ and $b \geq 0$. Now we assume that the hypothesis holds for $n \geq 0$, we distinguish between two cases:

– $s \in MS$: From definition we have:

$$\begin{aligned}
\mathcal{R}^{\max}_{n+1}(s,b) &= \Omega(\mathcal{R}^{\max}_n)(s,b) \\
&= \left(r_s + \frac{\rho(s)}{E(s)}\right)\left(1 - e^{-E(s)b}\right) + \int_0^b E(s) e^{-E(s)t} \sum_{s' \in S} \mathbb{P}_s(s') \mathcal{R}^{\max}_n(s', b-t)\,\mathrm{d}t \\
&\overset{\text{IH}}{\leq} \left(r_s + \frac{\rho(s)}{E(s)}\right)\left(1 - e^{-E(s)b}\right) + \int_0^b E(s) e^{-E(s)t} \sum_{s' \in S} \mathbb{P}_s(s') F(s', b-t)\,\mathrm{d}t \\
&= \Omega(F)(s,b) \\
&= F(s,b)
\end{aligned}$$

– $s \in PS$: Similarly we have:

$$\begin{aligned}
\mathcal{R}^{\max}_{n+1}(s,b) &= \Omega(\mathcal{R}^{\max}_n)(s,b) \\
&= \max_{\alpha \in A(s)} \left(r^\alpha_s + \sum_{s' \in S} \mu^\alpha_s(s') \mathcal{R}^{\max}_n(s', b)\right) \\
&\overset{\text{IH}}{\leq} \max_{\alpha \in A(s)} \left(r^\alpha_s + \sum_{s' \in S} \mu^\alpha_s(s') F(s', b)\right) \\
&= \Omega(F)(s,b) \\
&= F(s,b)
\end{aligned}$$

Finally it holds (see [26, Proposition 5.1]) that $F(s,b) \geq \lim_{n\to\infty} \mathcal{R}^{\max}_n(s,b) = \mathcal{R}^{\max}(s,b)$. □

C Proof of Theorem 3

Let $\mathcal{M}$ be an MRA, $b \geq 0$ a time bound and $\delta > 0$ a discretisation step such that $b = k\delta$ for some $k \in \mathbb{N}$. We partition the interval $[0,b]$ into $\Delta^{k,\delta} = \{[0,\delta), [\delta,2\delta), \ldots, [(k-2)\delta,(k-1)\delta), [(k-1)\delta,b]\}$, and denote the i-th sub-interval by $\Delta^{k,\delta}_i$ for $i = 1 \ldots k$. In case $b = 0$, then $k = 0$ and $\Delta^{k,\delta}$ refers to the point interval with $\Delta^{k,\delta} = \Delta^{k,\delta}_0 = \{0\}$. We then define the random vector Ξ^k_δ that counts the number of Markovian jumps in each of those sub-intervals. Formally it is defined as a function $\Xi^k_\delta : \mathit{paths} \mapsto \mathbb{N}^k$, with $\Xi^k_\delta(\pi)_i$ being the number of Markovian jumps occurred in path $\pi \in \mathit{paths}$ within sub-interval $\Delta^{k,\delta}_i$ for $i = 1 \ldots k$. Moreover let $\mathrm{sub}(\pi, I)$ denote the maximal sub-path of the infinite path π spanned by interval I. Note that it is always possible to split a path using this operator by $\pi = \mathrm{sub}(\pi, \Delta^{k,\delta}_1) \circ \cdots \circ \mathrm{sub}(\pi, \Delta^{k,\delta}_k) \circ \mathrm{sub}(\pi, (b,\infty))$. We later utilise this split for proving the theorem. We are also interested in the reward gained in each sub-path up to a certain jump. The notation $\mathit{reward}_{<j}(\pi^{\text{fin}})$ is then defined to refer to the reward gained by the finite path π^{fin} up to the j-th Markovian jump. Formally it defines as

$$\mathit{reward}_{<j}(\pi^{\text{fin}}) = \sum_{i=0}^{J^{\pi^{\text{fin}}}_j} \left(\rho(\pi^{\text{fin}}) \cdot t_i + r(\mathit{step}(\pi^{\text{fin}}, i))\right)$$

where t_i is the sojourn time at state $\pi^{\text{fin}}[i]$ and $J^{\pi^{\text{fin}}}_j$ is the index of the predecessor of the j-th Markovian state of the path. In case the path contains less

than j Markovian jumps, $J_j^{\pi^{\text{fin}}}$ is $|\pi^{\text{fin}}|$. Intuitively speaking $\mathit{reward}_{<j}(\pi^{\text{fin}})$ is the reward gained by finite path π^{fin} up to its j-th Markovian state. Afterwards we extend the notation to $\mathit{reward}^{\delta}_{<j}(\pi,k)$ to denote the reward of infinite path π up to time-bound $b = k\delta$ with the assumption that in each sub-path $\text{sub}(\pi, \Delta_i^{k,\delta})$ $(i = 1 \dots k)$, we only count the reward up to the j-th Markovian jump, i. e. $\mathit{reward}^{\delta}_{<j}(\pi,k) = \sum_{i=1}^{k} \mathit{reward}_{<j}(\text{sub}(\pi, \Delta_i^{k,\delta}))$. Note that if $k = 0$ then $\mathit{reward}^{\delta}_{<0}(\pi,0)$ is the reward gained by π at time instant zero. Using these concept the next lemma provides another representation of $\tilde{\mathcal{R}}^{\max}_{\delta}$.

Lemma C1 *Let $\mathcal{M}$ be an MRA, $b \geq 0$ a time bound and $\delta > 0$ a discretisation step such that $b = k\delta$ for some $k \in \mathbb{N}$. Then it holds that*

$$\tilde{\mathcal{R}}^{\max}_{\delta}(s,k) = \sup_{D \in GM} \int_{\pi \in paths} \mathit{reward}^{\delta}_{<1}(\pi,k) \cdot \Pr_{s,D}(d\pi \,|\, \|\Xi^k_\delta\|_\infty \leq 1). \tag{11}$$

where $\|\cdot\|_\infty$ denotes infinity norm, i. e. $\|\Xi^k_\delta\|_\infty = \max_{1 \leq i \leq k} \Xi^k_\delta(i)$.

Proof. The proof goes along the same line as that of Lemma 2. We first proof that $\tilde{\mathcal{R}}^{\max}_{\delta}$ is the fixed point of the characterisation given in Definition 7. The next step would be to show that $\tilde{\mathcal{R}}^{\max}_{\delta}$ is the least fixed point. We start with the first step and consider three cases. The trivial situation happens when $s \in MS$ and $k = 0$, then obviously $\Omega_\delta(\tilde{\mathcal{R}}^{\max}_{\delta}(s,k)) = \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) = 0$. Now we consider the case when $s \in MS$ and $k > 0$. Note that the probability measure in Eq. (11) is conditioned on $\|\Xi^k_\delta\|_\infty \leq 1$, which means that for the paths not satifying the condition the probability measure is zero. Hence we can restrict to the set of paths that satisfy the condition, i. e. $C^k_\delta = \{\pi : \|\Xi^k_\delta(\pi)\|_\infty \leq 1\}$. Moreover any path in the restricted set can be written as $\pi = \text{sub}(\pi, \Delta_1^{k,\delta}) \circ \pi'$. Given that the first jump of π happens at time point t, it is then possible to split the reward and the probability measure:

$$\mathit{reward}(\pi,k) = \begin{cases} \delta \cdot \rho(s) + \mathit{reward}^{\delta}_{<2}(\pi', k-1) & \Xi^k_\delta(1) = 0 \\ r_s + t \cdot \rho(s) + \mathit{reward}^{\delta}_{<2}(\pi'', k-1) & \Xi^k_\delta(1) = 1 \end{cases}$$

$$\Pr_{s,D}(\,\mathrm{d}\pi \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1\,) = \begin{cases} \mathrm{e}^{-E(s)\delta} \Pr_{s,D_\delta}(\,\mathrm{d}\pi' \,\big|\, \|\Xi^{k-1}_\delta\|_\infty \leq 1\,) & \Xi^k_\delta(1) = 0 \\ E(s)\mathrm{e}^{-E(s)t}\,\mathrm{d}t \sum_{s' \in S} \mathbb{P}_s(s') \Pr_{s', D[s \xrightarrow{t}]}(\,\mathrm{d}\pi'' \,\big|\, \|\Xi^{k-1}_\delta\|_\infty \leq 1\,) & \Xi^k_\delta(1) = 1 \end{cases}$$

with $\pi'' = \text{sub}(\pi, [t,t]) \circ \pi'$ and $D[s \xrightarrow{t}]$ is the scheduler that resolves nondeterminisim for any prefix ζ of π'' as follows

$$D[s \xrightarrow{t}](\zeta) = \begin{cases} D(s \xrightarrow{\chi(E(s)), \mathbb{P}_s, t} \zeta) & \text{time}(\zeta) = 0 \\ D(s \xrightarrow{\chi(E(s)), \mathbb{P}_s, t} \text{sub}(\pi, [t,t]) \circ \text{sub}(\zeta, (t, \text{time}(\zeta)]) & \text{time}(\zeta) > 0 \end{cases}$$

Therefore we can write:

$$\begin{aligned} \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) &= \sup_{D \in GM} \int_{C^k_\delta} \mathit{reward}^{\delta}_{<2}(\pi,k) \cdot \Pr_{s,D}(\,\mathrm{d}\pi \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1\,) \\ &= \sup_{D \in GM} \left(\int_0^\delta \int_{\pi'' \in paths} \Big(t\rho(s) + r_s + \mathit{reward}^{\delta}_{<2}(\pi'', k-1) \Big) E(s) \mathrm{e}^{-E(s)t} \sum_{s' \in S} \mathbb{P}_s(s') \right. \end{aligned}$$

$$
\begin{aligned}
&\cdot \Pr_{s',D[s\xrightarrow{t}]}(\,\mathrm{d}\pi'' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1)\,\mathrm{d}t \\
&+ \int_{\pi'\in paths} \Big(\delta\rho(s) + reward^\delta_{<2}(\pi', k-1)\Big)\,\mathrm{e}^{-E(s)\delta} \Pr_{s,D_\delta}(\,\mathrm{d}\pi' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1)\,\mathrm{d}t\Big) \\
= \sup_{D\in GM}\Bigg(&\int_0^\delta (t\rho(s)+r_s)E(s)\mathrm{e}^{-E(s)t}\sum_{s'\in S}\mathbb{P}_s(s')\int_{\pi''\in paths}\Pr_{s',D[s\xrightarrow{t}]}(\,\mathrm{d}\pi'' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1))\,\mathrm{d}t \\
&+ \int_0^\delta E(s)\mathrm{e}^{-E(s)t}\sum_{s'\in S}\mathbb{P}_s(s')\int_{\pi''\in paths} reward^\delta_{<2}(\pi'', k-1)\Pr_{s',D[s\xrightarrow{t}]}(\,\mathrm{d}\pi'' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1))\,\mathrm{d}t \\
&+ \delta\rho(s)\mathrm{e}^{-E(s)\delta}\int_{\pi'\in paths}\Pr_{s,D_\delta}(\,\mathrm{d}\pi' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1)\,\mathrm{d}t \\
&+ \mathrm{e}^{-E(s)\delta}\int_{\pi'\in paths} reward^\delta_{<2}(\pi', k-1)\Pr_{s,D_\delta}(\,\mathrm{d}\pi' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1)\,\mathrm{d}t\Bigg) \\
= \sup_{D\in GM}\Bigg(&\int_0^\delta (t\rho(s)+r_s)E(s)\mathrm{e}^{-E(s)t}\,\mathrm{d}t \\
&+ \int_0^\delta E(s)\mathrm{e}^{-E(s)t}\sum_{s'\in S}\mathbb{P}_s(s')\int_{\pi''\in paths} reward^\delta_{<2}(\pi'', k-1)\Pr_{s',D[s\xrightarrow{t}]}(\,\mathrm{d}\pi'' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1))\,\mathrm{d}t \\
&+ \delta\rho(s)\mathrm{e}^{-E(s)\delta} + \mathrm{e}^{-E(s)\delta}\int_{\pi'\in paths} reward^\delta_{<2}(\pi', k-1)\Pr_{s,D_\delta}(\,\mathrm{d}\pi' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1)\,\mathrm{d}t\Bigg) \\
= \sup_{D\in GM}\Bigg(&\Big(r_s + \tfrac{\rho(s)}{E(s)}\Big)\Big(1-\mathrm{e}^{-E(s)\delta}\Big) \\
&+ \int_0^\delta E(s)\mathrm{e}^{-E(s)t}\sum_{s'\in S}\mathbb{P}_s(s')\int_{\pi''\in paths} reward^\delta_{<2}(\pi'', k-1)\Pr_{s',D[s\xrightarrow{t}]}(\,\mathrm{d}\pi'' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1))\,\mathrm{d}t \\
&+ \mathrm{e}^{-E(s)\delta}\int_{\pi'\in paths} reward^\delta_{<2}(\pi', k-1)\Pr_{s,D_\delta}(\,\mathrm{d}\pi' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1)\,\mathrm{d}t\Bigg) \\
= &\Big(r_s + \tfrac{\rho(s)}{E(s)}\Big)\Big(1-\mathrm{e}^{-E(s)\delta}\Big) \\
&+ \sup_{D\in GM}\Bigg(\int_0^\delta E(s)\mathrm{e}^{-E(s)t}\sum_{s'\in S}\mathbb{P}_s(s')\int_{\pi''\in paths} reward^\delta_{<2}(\pi'', k-1)\Pr_{s',D[s\xrightarrow{t}]}(\,\mathrm{d}\pi'' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1))\,\mathrm{d}t \\
&+ \mathrm{e}^{-E(s)\delta}\int_{\pi'\in paths} reward^\delta_{<2}(\pi', k-1)\Pr_{s,D_\delta}(\,\mathrm{d}\pi' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1)\,\mathrm{d}t\Bigg) \\
\overset{(*)}{=} &\Big(r_s + \tfrac{\rho(s)}{E(s)}\Big)\Big(1-\mathrm{e}^{-E(s)\delta}\Big) \\
&+ \int_0^\delta E(s)\mathrm{e}^{-E(s)t}\sum_{s'\in S}\mathbb{P}_s(s')\sup_{D\in GM}\Big(\int_{\pi''\in paths} reward^\delta_{<2}(\pi'', k-1)\Pr_{s',D[s\xrightarrow{t}]}(\,\mathrm{d}\pi'' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1))\Big)\,\mathrm{d}t \\
&+ \mathrm{e}^{-E(s)\delta}\sup_{D\in GM}\Big(\int_{\pi'\in paths} reward^\delta_{<2}(\pi', k-1)\Pr_{s,D_\delta}(\,\mathrm{d}\pi' \,\big|\, \|\Xi_\delta^{k-1}\|_\infty \le 1)\,\mathrm{d}t\Big) \\
= &\Big(r_s + \tfrac{\rho(s)}{E(s)}\Big)\Big(1-\mathrm{e}^{-E(s)\delta}\Big) + \int_0^\delta E(s)\mathrm{e}^{-E(s)t}\sum_{s'\in S}\mathbb{P}_s(s')\tilde{\mathcal{R}}^{\max}_\delta(s', k-1) \\
&+ \mathrm{e}^{-E(s)\delta}\tilde{\mathcal{R}}^{\max}_\delta(s, k-1) \\
= &\tilde{A}_\delta(s,k) + \mathrm{e}^{-E(s)\delta}\tilde{\mathcal{R}}^{\max}_\delta(s, k-1)
\end{aligned}
$$

where $(*)$ follows from the fact that D_δ and $D[s \xrightarrow{t}]$ resolve nondeterminism for completely different paths. D_δ makes dicision for the paths that stay for at least δ time units in s whereas $D[s \xrightarrow{t}]$ does it for the paths that jump to some successor of s at time point $0 \le t \le \delta$. Therefore they can independently

maximise the corrsponding term. The proof for the next case, $s \in PS$ is similar to the proof of corresponding case in Lemma 2.

We can argue using the similar technique in the proof of Lemma 2 that $\tilde{\mathcal{R}}^{\max}_{\delta}$ is the least fixed point of the characterisation given in Definition 7.

Now we prove the error bound. We fisrt show the left hand side of the inequality.

Lemma C2 *Let $\mathcal{M}$ be an MRA, $b \geq 0$ be a time bound, $\delta > 0$ be a discretisation step such that $b = k\delta$ for some $k \in \mathbb{N}$. Then for all $s \in S$:*

$$\tilde{\mathcal{R}}^{\max}_{\delta}(s,k) \leq \mathcal{R}^{\max}(s,b)$$

Proof. We first show that for two given functions $F : S \times \mathbb{N} \mapsto \mathbb{R}_{\geq 0}$ and $G : S \times \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ the following holds

$$F(s,k) \leq G(s,k\delta) \Longrightarrow \Omega_{\delta}(F)(s,k) \leq \Omega(G)(s,k\delta), \forall s \in S, k \in \mathbb{N}$$

we consider two cases:

- $s \in MS$

$$\begin{aligned}
\Omega_{\delta}(F)(s,k) &= \Big(r_s + \frac{\rho(s)}{E(s)} + \sum_{s' \in S} \mathbb{P}_s(s')F(s',k-1)\Big)\Big(1 - \mathrm{e}^{-E(s)\delta}\Big) \\
&\quad + \mathrm{e}^{-E(s)\delta}F(s,k-1) \\
&\leq \Big(r_s + \frac{\rho(s)}{E(s)} + \sum_{s' \in S} \mathbb{P}_s(s')G(s',(k-1)\delta)\Big)\Big(1 - \mathrm{e}^{-E(s)\delta}\Big) \\
&\quad + \mathrm{e}^{-E(s)\delta}G(s,(k-1)\delta) \\
&= \int_0^{\delta} E(s)\mathrm{e}^{-E(s)t} \sum_{s' \in S} \mathbb{P}_s(s')G(s',(k-1)\delta)\,\mathrm{d}t \\
&\quad + \Big(r_s + \frac{\rho(s)}{E(s)}\Big)\Big(1 - \mathrm{e}^{-E(s)\delta}\Big) + \mathrm{e}^{-E(s)\delta}G(s,(k-1)\delta) \\
&\overset{(*)}{\leq} \int_0^{\delta} E(s)\mathrm{e}^{-E(s)t} \sum_{s' \in S} \mathbb{P}_s(s')G(s',k\delta - t)\,\mathrm{d}t \\
&\quad + \Big(r_s + \frac{\rho(s)}{E(s)}\Big)\Big(1 - \mathrm{e}^{-E(s)\delta}\Big) + \mathrm{e}^{-E(s)\delta}G(s,(k-1)t) \\
&= A(s,k) + \mathrm{e}^{-E(s)\delta}G(s,(k-1)\delta) = \Omega(G)(s,k\delta)
\end{aligned}$$

  where $(*)$ follows from the fact that $G(s,t)$ is monotonically increasing w.r.t. t.

- $s \in PS$

$$\begin{aligned}
\Omega_{\delta}(F)(s,k) &= \max_{\alpha \in A(s)} \Big(r_s^{\alpha} + \sum_{s' \in S} \mu_s^{\alpha}(s')F(s',k)\Big) \\
&\leq \max_{\alpha \in A(s)} \Big(r_s^{\alpha} + \sum_{s' \in S} \mu_s^{\alpha}(s')G(s',k\delta)\Big) = \Omega(G)(s,k\delta)
\end{aligned}$$

We know that the least fixed point of Ω is $\mathcal{R}^{\max}$ and define $H(s,k) = \mathcal{R}^{\max}(s,k\delta)$, therefore

$$\Omega_\delta(H)(s,k) \leq \Omega(\mathcal{R}^{\max})(s,k\delta) = \mathcal{R}^{\max}(s,k\delta) \Rightarrow \lim_{n\to\infty} \Omega_\delta^n(H)(s,k) \leq \mathcal{R}^{\max}(s,k\delta)$$

This means that Ω_δ has some fixed point which is less or equal than $\mathcal{R}^{\max}$. Hence its least fixed point is also less than $\mathcal{R}^{\max}$. □

For the most of the lemmas we should show that some error bound holds for both of the Markovian and probabilistic states. By the following lemma we show that if some bound holds for Markovian state, so it does for probabilistic states.

Lemma C3 *For some $k \in \mathbb{N}$:*

$$\Big(\forall s \in MS. \mathcal{R}^{\max}(s,k\delta) - \tilde{\mathcal{R}}_\delta^{\max}(s,k) \leq \epsilon\Big) \Rightarrow \Big(\forall s \in PS. \mathcal{R}^{\max}(s,k\delta) - \tilde{\mathcal{R}}_\delta^{\max}(s,k) \leq \epsilon\Big)$$

Proof. We start from an arbitrary probabilistic state s and show that it respects the bound. Consider all paths starting from state s, we can decompose them into two parts. The first part starts from s and ends up to the first Markovian state and the second part start from the first Markovian state and continue afterwards. Since in closed model all probabilistic transitions take immediately we can write $\pi = \mathrm{sub}(\pi,[0,0]) \circ \pi'$. Accordingly we can split the reward. As the models we are considering are non-zeno, some Markovian state is almost surely reached. Hence by using the law of total probability and the fact that $\mathrm{sub}(\pi,[0,0])$ does not contain any Markovian jump it holds that

$$\begin{aligned}
\Pr_{s,D}(\,\mathrm{d}\pi) &= \sum_{s'\in MS} \Pr_{s,D}(\mathrm{sub}(\pi,[0,0])) \Pr_{s',D}(\,\mathrm{d}\pi') \\
\Pr_{s,D}(\,\mathrm{d}\pi \,\big|\, \|\Xi_\delta^k\|_\infty \leq 1)) &= \sum_{s'\in MS} \Pr_{s,D}(\mathrm{sub}(\pi,[0,0])) \Pr_{s',D}(\,\mathrm{d}\pi' \,\big|\, \|\Xi_\delta^k\|_\infty \leq 1))
\end{aligned}$$

Assume that D^* is the optimal scheduler for $\mathcal{R}^{\max}$, then

$$\begin{aligned}
\mathcal{R}^{\max}(s,k\delta) &= \int_\pi \mathit{reward}(\pi,k\delta) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&= \int_\pi \Big(\mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) + \mathit{reward}(\pi',k\delta)\Big) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&= \int_\pi \mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&\quad + \int_\pi \mathit{reward}(\pi',k\delta) \sum_{s'\in MS} \Pr_{s,D^*}(\mathrm{sub}(\pi,[0,0])) \Pr_{s',D^*}(\,\mathrm{d}\pi') \\
&= \int_\pi \mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&\quad + \sum_{s'\in MS} \Pr_{s,D^*}(\Diamond^{[0,0]}\{s'\}) \cdot \int_{\pi'} \mathit{reward}(\pi',k\delta) \Pr_{s',D^*}(\,\mathrm{d}\pi')
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\pi} \mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&\quad + \sum_{s'\in MS} \Pr_{s,D^*}(\Diamond^{[0,0]}\{s'\}) \cdot \mathcal{R}^{\max}(s', k\delta) \\
&\leq \int_{\pi} \mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&\quad + \sum_{s'\in MS} \Pr_{s,D^*}(\Diamond^{[0,0]}\{s'\}) \cdot (\tilde{\mathcal{R}}^{\max}_{\delta}(s', k) + \epsilon) \\
&= \int_{\pi} \mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&\quad + \sum_{s'\in MS} \Pr_{s,D^*}(\Diamond^{[0,0]}\{s'\}) \cdot \tilde{\mathcal{R}}^{\max}_{\delta}(s', k) + \epsilon \\
&= \int_{\pi} \mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&\quad + \sum_{s'\in MS} \Pr_{s,D^*}(\Diamond^{[0,0]}\{s'\}) \cdot \int_{\pi'} \mathit{reward}^{\delta}_{<2}(\pi', k) \Pr_{s,D^*}(\,\mathrm{d}\pi' \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1) + \epsilon \\
&= \int_{\pi} \mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&\quad + \int_{\pi} \mathit{reward}^{\delta}_{<2}(\pi', k) \sum_{s'\in MS} \Pr_{s,D^*}(\mathrm{sub}(\pi,[0,0])) \Pr_{s,D^*}(\,\mathrm{d}\pi' \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1) + \epsilon \\
&= \int_{\pi} \mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) \Pr_{s,D^*}(\,\mathrm{d}\pi) \\
&\quad + \int_{\pi} \mathit{reward}^{\delta}_{<2}(\pi', k) \Pr_{s,D^*}(\,\mathrm{d}\pi \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1) + \epsilon \\
&= \int_{\pi} \mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) \Pr_{s,D^*}(\,\mathrm{d}\pi \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1) \\
&\quad + \int_{\pi} \mathit{reward}^{\delta}_{<2}(\pi', k) \Pr_{s,D^*}(\,\mathrm{d}\pi \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1) + \epsilon \\
&= \int_{\pi} \Big(\mathit{reward}\big(\mathrm{sub}(\pi,[0,0])\big) + \mathit{reward}^{\delta}_{<2}(\pi', k)\Big) \Pr_{s,D^*}(\,\mathrm{d}\pi \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1) + \epsilon \\
&= \int_{\pi} \mathit{reward}^{\delta}_{<2}(\pi, k) \Pr_{s,D^*}(\,\mathrm{d}\pi \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1) + \epsilon \\
&\leq \sup_{D\in GM} \int_{\pi} \mathit{reward}^{\delta}_{<2}(\pi, k) \Pr_{s,D}(\,\mathrm{d}\pi \,\big|\, \|\Xi^k_\delta\|_\infty \leq 1) + \epsilon \\
&= \tilde{\mathcal{R}}^{\max}_{\delta}(s', k) + \epsilon
\end{aligned}
$$

□

Now we prove the right hand side of the error bound in Theorem 3 for state and transition rewards separately. We first show the bound for state rewards and assume that the transition rewards are zero. The following lemma establish bounds that are used for the proof.

Lemma C4 *Let $\mathcal{M}$ be an MRA with $r(tr) = 0, \forall tr \in S \times A^{\chi} \times \mathsf{Distr}(S)$ and assume that λ and $\rho_{\max}$ are the maximum exit rate and state reward of any*

Markovian state in $\mathcal{M}$, respectively. Furthermore, suppose that $b \geq 0$ be a time bound, $\delta > 0$ be a discretisation step such that $b = k\delta$ for some $k \in \mathbb{N}$. Then

$$\forall s \in S.\, \mathcal{R}^{\max}(s, t+t') \leq (t+t')\rho_{\max} + \mathrm{e}^{-\lambda t'}(\mathcal{R}^{\max}(s,t) - t\rho_{\max}), \tag{12}$$

$$\begin{aligned}\forall s \in MS.\, A(s,k\delta) - \tilde{A}_\delta(s,k) &\leq k\delta\rho_{\max} - (k-1)\delta\rho_{\max}\mathrm{e}^{-E(s)\delta} - \frac{\rho_{\max}}{E(s)}(1-\mathrm{e}^{-E(s)\delta}) \\ &+ \frac{E(s)}{\lambda - E(s)}(\mathrm{e}^{-E(s)\delta} - \mathrm{e}^{-\lambda\delta})(\mathcal{R}^{\max}(s,(k-1)\delta) - \tilde{\mathcal{R}}^{\max}_\delta(s,k-1) - (k-1)\delta\rho_{\max})\end{aligned} \tag{13}$$

Proof. For the proof of Eq. (12) note that the maximum reward from any state s is $(t+t')\rho_{\max}$ when $\forall s \in S.\, \rho(s) = \rho_{\max}$. The condition may not always hold. In particular we consider the case that $\mathcal{R}^{\max}(s,t) \leq t\rho_{\max}$ which allows to compensate the bound by term $\mathcal{R}^{\max}(s,t) - t\rho_{\max}$. Therefore,

$$\begin{aligned}\mathcal{R}^{\max}(s,t+t') &\leq (t+t')\rho_{\max} \\ &\quad + \sup_{D \in GM} \Pr^{s,D}(\text{no Markovian jump in } [0,t'])(\mathcal{R}^{\max}(s,t) - t\rho_{\max}) \\ &\leq (t+t')\rho_{\max} + \mathrm{e}^{-\lambda t'}(\mathcal{R}^{\max}(s,t) - t\rho_{\max})\end{aligned}$$

Using Eq. (12) it is possible to show correctness of (13).

$$\begin{aligned}A(s,k\delta) - \tilde{A}_\delta(s,k) &= \int_0^\delta E(s)\mathrm{e}^{-E(s)t} \sum_{s' \in S} \mathbb{P}_s(s) \left(\mathcal{R}^{\max}(s, k\delta - t) - \tilde{\mathcal{R}}^{\max}_\delta(s,k-1)\right) \mathrm{d}t \\ &\leq \int_0^\delta E(s)\mathrm{e}^{-E(s)t} \sum_{s' \in S} \mathbb{P}_s(s) \Big((k\delta - t)\rho_{\max} + \mathrm{e}^{-\lambda(\delta - t)} \\ &\quad \times \left(\mathcal{R}^{\max}(s,(k-1)\delta) - (k-1)\delta\rho_{\max}\right) - \tilde{\mathcal{R}}^{\max}_\delta(s,k-1)\Big) \mathrm{d}t \qquad (^* \text{ by Eq. (12) } ^*) \\ &\leq \int_0^\delta E(s)\mathrm{e}^{-E(s)t} \sum_{s' \in S} \mathbb{P}_s(s) \Big((k\delta - t)\rho_{\max} + \mathrm{e}^{-\lambda(\delta - t)} \\ &\quad \times \left(\mathcal{R}^{\max}(s,(k-1)\delta) - (k-1)\delta\rho_{\max} - \tilde{\mathcal{R}}^{\max}_\delta(s,k-1)\right)\Big) \mathrm{d}t \\ &= k\delta\rho_{\max}(1 - \mathrm{e}^{-E(s)\delta}) - \frac{\rho_{\max}}{E(s)}\left(1 - \mathrm{e}^{-E(s)\delta}(1 + E(s)\delta)\right) \\ &\quad + \frac{E(s)}{\lambda - E(s)}(\mathrm{e}^{-E(s)\delta} - \mathrm{e}^{-\lambda\delta})\left(\mathcal{R}^{\max}(s,(k-1)\delta) - (k-1)\delta\rho_{\max} - \tilde{\mathcal{R}}^{\max}_\delta(s,k-1)\right) \\ &= k\delta\rho_{\max} - (k-1)\delta\rho_{\max}\mathrm{e}^{-E(s)\delta} - \frac{\rho_{\max}}{E(s)}(1 - \mathrm{e}^{-E(s)\delta}) \\ &\quad + \frac{E(s)}{\lambda - E(s)}(\mathrm{e}^{-E(s)\delta} - \mathrm{e}^{-\lambda\delta})(\mathcal{R}^{\max}(s,(k-1)\delta) - \tilde{\mathcal{R}}^{\max}_\delta(s,k-1) - (k-1)\delta\rho_{\max})\end{aligned}$$

□

The next lemma provides the upper bound for maximum expected state rewards.

Lemma C5 *Let $\mathcal{M}$ be an MRA with $r(tr) = 0, \forall tr \in S \times A^\chi \times \mathsf{Distr}(S)$ and assume that λ and $\rho_{\max}$ are the maximum exit rate and state reward of any*

Markovian state in $\mathcal{M}$, respectively. Furthermore, suppose that $b \geq 0$ be a time bound, $\delta > 0$ be a discretisation step such that $b = k\delta$ for some $k \in \mathbb{N}$. Then for all $s \in S$:

$$\mathcal{R}^{\max}(s,b) \leq \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + b\rho_{\max} - \frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-1}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i$$

Proof. The claim holds for $k = 0$, since

$$\mathcal{R}^{\max}(s,0) = \tilde{\mathcal{R}}^{\max}_{\delta}(s,0) \qquad \forall s \in S. \tag{14}$$

We assume $k > 0$ and prove the claim by induction.

a. Induction base: $k = 1$. We distinguish between two cases:
   (i) $s \in MS$:

$$\begin{aligned}
\mathcal{R}^{\max}(s,\delta) &= A(s,\delta) + \mathrm{e}^{-E(s)\delta}\mathcal{R}^{\max}(s,0)\\
&= A(s,\delta) + \mathrm{e}^{-E(s)\delta}\tilde{\mathcal{R}}^{\max}_{\delta}(s,0) && (^*\text{ by Eq. (14) }^*)\\
&\leq \tilde{A}_{\delta}(s,1) + \delta\rho_{\max} - \frac{\rho_{\max}}{E(s)}(1-\mathrm{e}^{-E(s)\delta}) + \mathrm{e}^{-E(s)\delta}\tilde{\mathcal{R}}^{\max}_{\delta}(s,0)\\
& && (^*\text{ by Eq. (13) and (14) }^*)\\
&\leq \tilde{\mathcal{R}}^{\max}_{\delta}(s,1) + \delta\rho_{\max} - \frac{\rho_{\max}}{E(s)}(1-\mathrm{e}^{-E(s)\delta})\\
&\overset{(*)}{\leq} \tilde{\mathcal{R}}^{\max}_{\delta}(s,1) + \delta\rho_{\max} - \frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta}) && (15)
\end{aligned}$$

   where $(*)$ follows from the fact that $f(\gamma) = \delta\rho_{\max} - \frac{\rho_{\max}}{\gamma}(1-\mathrm{e}^{-\gamma\delta})$ is monotonically increasing w.r.t γ when $\gamma \in [0,\lambda]$.
   (ii) $s \in PS$: It directly follows from case a.(i) and Lemma C3.

b. Induction step: $k-1 \rightsquigarrow k$. We distinguish between two cases:
   (i) $s \in MS$:

$$\begin{aligned}
&\mathcal{R}^{\max}(s,k\delta) = A(s,k\delta) + \mathrm{e}^{-E(s)\delta}\mathcal{R}^{\max}(s,(k-1)\delta)\\
&\leq \tilde{A}_{\delta}(s,k) + k\delta\rho_{\max} - (k-1)\delta\rho_{\max}\mathrm{e}^{-E(s)\delta} - \frac{\rho_{\max}}{E(s)}(1-\mathrm{e}^{-E(s)\delta})\\
&\quad + \frac{E(s)}{\lambda - E(s)}(\mathrm{e}^{-E(s)\delta} - \mathrm{e}^{-\lambda\delta})(\mathcal{R}^{\max}(s,(k-1)\delta) - \tilde{\mathcal{R}}^{\max}_{\delta}(s,k-1) - (k-1)\delta\rho_{\max})\\
&\quad + \mathrm{e}^{-E(s)\delta}\mathcal{R}^{\max}(s,(k-1)\delta) && (^*\text{ by Eq. (14) }^*)\\
&\leq \tilde{A}_{\delta}(s,k) + k\delta\rho_{\max} - (k-1)\delta\rho_{\max}\mathrm{e}^{-E(s)\delta} - \frac{\rho_{\max}}{E(s)}(1-\mathrm{e}^{-E(s)\delta})\\
&\quad - \frac{E(s)}{\lambda - E(s)}(\mathrm{e}^{-E(s)\delta} - \mathrm{e}^{-\lambda\delta})\frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-2}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i\\
&\quad + \mathrm{e}^{-E(s)\delta}\Big(\tilde{\mathcal{R}}^{\max}_{\delta}(s,(k-1)\delta) + (k-1)\delta\rho_{\max} - \frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-2}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i\Big)\\
& && (^*\text{ by Ind. Hyp. }^*)\\
&= \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + k\delta\rho_{\max} - (k-1)\delta\rho_{\max}\mathrm{e}^{-E(s)\delta} - \frac{\rho_{\max}}{E(s)}(1-\mathrm{e}^{-E(s)\delta})\\
&\quad - \frac{E(s)}{\lambda - E(s)}(\mathrm{e}^{-E(s)\delta} - \mathrm{e}^{-\lambda\delta})\frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-2}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i
\end{aligned}$$

$$
\begin{aligned}
&+ \mathrm{e}^{-E(s)\delta}\Big((k-1)\delta\rho_{\max} - \frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-2}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i\Big)\\
&= \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + k\delta\rho_{\max} - \frac{\rho_{\max}}{E(s)}(1-\mathrm{e}^{-E(s)\delta})\\
&\quad - \frac{E(s)}{\lambda - E(s)}(\mathrm{e}^{-E(s)\delta} - \mathrm{e}^{-\lambda\delta})\frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-2}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i\\
&\quad - \mathrm{e}^{-E(s)\delta}\frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-2}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i\\
&= \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + k\delta\rho_{\max} - \frac{\rho_{\max}}{E(s)}(1-\mathrm{e}^{-E(s)\delta})\\
&\quad - \Big(\frac{E(s)}{\lambda - E(s)}(\mathrm{e}^{-E(s)\delta} - \mathrm{e}^{-\lambda\delta}) + \mathrm{e}^{-E(s)\delta}\Big)\frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-2}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i\\
&\overset{(*)}{\leq} \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + k\delta\rho_{\max} - \frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\\
&\quad - \mathrm{e}^{-\lambda\delta}(1+\lambda\delta)\frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-2}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i\\
&= \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + k\delta\rho_{\max} - \frac{\rho_{\max}}{\lambda}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-1}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i
\end{aligned}
$$

where $(*)$ follows from the fact that both of the functions $f(\gamma) = \delta\rho_{\max} - \frac{\rho_{\max}}{\gamma}(1-\mathrm{e}^{-\gamma\delta})$ and $g(\gamma) = -(\frac{\gamma}{\lambda-\gamma}(\mathrm{e}^{-\gamma\delta}-\mathrm{e}^{-\lambda\delta})+\mathrm{e}^{-\gamma\delta})$ are monotonically increasing w.r.t γ when $\gamma \in [0,\lambda]$ and $\lim_{\gamma\to\lambda} g(\gamma) = \mathrm{e}^{-\lambda\delta}(1+\lambda\delta)$.

(ii) $s \in PS$: It directly follows from case b.(i) and Lemma C3. □

The following lemma established the upper bound for transition rewards.

Lemma C6 *Let $\mathcal{M}$ be an MRA with $\rho(s) = 0, \forall s \in S$ and assume that λ and $r_{\max}$ are respectively the maximum exit rate of any Markovian state and the maximum transition reward can be gained by doing arbitrary many probabilistic transitions as defined in Section 3.3. Furthermore, suppose that $b \geq 0$ be a time bound, $\delta > 0$ be a discretisation step such that $b = k\delta$ for some $k \in \mathbb{N}$. Then for all $s \in S$:*

$$\mathcal{R}^{\max}(s,b) \leq \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + b\lambda r_{\max} - r_{\max}(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-1}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i$$

Proof. The proof goes along the same line as the one for Lemma C5. □

Theorem 3 *Let $\mathcal{M}$ be an MRA, $b \geq 0$ be a time bound, $\delta > 0$ be a discretisation step such that $b = k\delta$ for some $k \in \mathbb{N}$. Then for all $s \in S$:*

$$\tilde{\mathcal{R}}^{\max}_{\delta}(s,k) \leq \mathcal{R}^{\max}(s,b) \leq \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + \frac{b\lambda}{2}(\rho_{\max} + r_{\max}\lambda)(1+\frac{b\lambda}{2})\delta$$

Proof. Using the idea of separation between state and transition rewards and applying the bounds provided by Lemmas C2, C5 and C6 it holds that

$$\mathcal{R}^{\max}(s,b) \leq \tilde{\mathcal{R}}^{\max}_{\delta}(s,k) + b(\rho_{\max}+\lambda r_{\max}) - (\frac{\rho_{\max}}{\lambda}+r_{\max})(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-1}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i$$

One can show by straightforward mathematics and using appropriate inequalities that

$$b(\rho_{\max}+\lambda r_{\max})-(\frac{\rho_{\max}}{\lambda}+r_{\max})(1-\mathrm{e}^{-\lambda\delta})\sum_{i=0}^{k-1}\mathrm{e}^{-\lambda i\delta}(1+\lambda\delta)^i \leq \frac{b\lambda}{2}(\rho_{\max}+r_{\max}\lambda)(1+\frac{b\lambda}{2})\delta$$

□

D Proof of Theorem 4

Theorem 4 *For a unichain MRA $\mathcal{M}$, for each $s \in S$ the value of $\mathsf{LRR}^{\max}_{\mathcal{M}}(s)$ equals*

$$\mathsf{LRR}^{\max}_{\mathcal{M}} = \sup_D \sum_{s\in S}\Big(\rho(s)\cdot \mathsf{LRA}^{\mathsf{D}}(s) + r_s^{D(s)}\cdot \nu^D(s)\Big)$$

where ν is the frequency of passing through a state, defined by

$$\nu^D(s) = \begin{cases} \mathsf{LRA}^D(s)\cdot E(s) & \text{if } s \in MS \\ \sum\limits_{s'\in S}\nu^D(s')\cdot \mu_{s'}^{D(s')}(s) & \text{if } s \in PS \end{cases}$$

and $\mathsf{LRA}^D(s)$ is the long-run average time spent in state s under stationary deterministic policy D.

Proof. We show in a sketch proof that Theorem 5 and Equation 8 coincide. Therefore, we will distinguish within the proof two cases: $s \in MS$ and $s \in PS$. For the definition of $\pi@t$ we refer to [18]. Note that we denote by $|\pi@t|$ the index of the last state of the sequence $\pi@t$, and with $\rho(\pi@t)$ the state reward of the last state of the sequence $\pi@t$.

$$\begin{aligned}
\mathsf{LRR}^{\max}_{\mathcal{M}} &= \sup_D \int_{paths} \mathcal{L}_{\mathcal{M}}(\pi)\Pr_D(d\pi) \\
&= \sup_D \int_{paths}\left[\lim_{t\to\infty}\frac{1}{t}\left(\rho(\pi@t)\cdot(t-\sum_{j=0}^{|\pi@t|-1}t_j)+\sum_{j=0}^{|\pi@t|-1}(t_j\cdot\rho(\pi[j])+r_{j+1})\right)\right]\Pr_D(d\pi) \\
&= \sup_D \int_{paths}\left[\lim_{t\to\infty}\frac{1}{t}\left(\sum_{s\in S}\left(\mathbf{1}_s(\pi@t)\cdot\rho(\pi@t)\cdot(t-\sum_{j=0}^{|\pi@t|-1}t_j)\right.\right.\right. \\
&\quad \left.\left.\left.+\sum_{j=0}^{|\pi@t|-1}\mathbf{1}_s(\pi[j])\cdot(t_j\cdot\rho(\pi[j])+r_{j+1})\right)\right)\right]\Pr_D(d\pi) \\
&= \sup_D \int_{paths}\left[\lim_{t\to\infty}\frac{1}{t}\left(\sum_{s\in S}\left(\mathbf{1}_s(\pi@t)\cdot\rho(\pi@t)\cdot(t-\sum_{j=0}^{|\pi@t|-1}t_j)\right.\right.\right. \\
&\quad \left.\left.\left.+\sum_{j=0}^{|\pi@t|-1}\mathbf{1}_s(\pi[j])\cdot t_j\cdot\rho(\pi[j])+\sum_{j=0}^{|\pi@t|-1}\mathbf{1}_s(\pi[j])\cdot r_{j+1})\right)\right)\right]\Pr_D(d\pi) \\
&= \sup_D \int_{paths}\left[\lim_{t\to\infty}\frac{1}{t}\left(\sum_{s\in S}\left(\left(\mathbf{1}_s(\pi@t)\cdot(t-\sum_{j=0}^{|\pi@t|-1}t_j)+\sum_{j=0}^{|\pi@t|-1}\mathbf{1}_s(\pi[j])\cdot t_j\right)\cdot\rho(s)\right.\right.\right. \\
&\quad \left.\left.\left.+\sum_{j=0}^{|\pi@t|-1}\mathbf{1}_s(\pi[j])\cdot r_{j+1})\right)\right)\right]\Pr_D(d\pi)
\end{aligned}$$

$$
\begin{aligned}
&= \sup_D \int_{paths} \left[\lim_{t\to\infty} \frac{1}{t} \left(\sum_{s\in S} \left(\int_0^t \mathbf{1}_s(\pi@u)du \cdot \rho(s) + \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \cdot r_{j+1} \right) \right) \right] \Pr_D(d\pi) \\
&= \sup_D \int_{paths} \left[\sum_{s\in S} \left(\lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbf{1}_s(\pi@u)du \cdot \rho(s) + \lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \cdot r_{j+1} \right) \right] \Pr_D(d\pi) \\
&= \sup_D \sum_{s\in S} \Bigg[\rho(s) \cdot \int_{paths} \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbf{1}_s(\pi@u)du \Pr_D(d\pi) \\
&\quad + \int_{paths} \left[\lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \cdot r_{j+1} \right] \Pr_D(d\pi) \Bigg] \\
&= \sup_D \sum_{s\in S} \left[\rho(s) \cdot \mathsf{LRA}^D(s) + \int_{paths} \left[\lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \cdot r_{j+1} \right] \Pr_D(d\pi) \right] \\
&= \sup_D \sum_{s\in S} \Bigg[\rho(s) \cdot \mathsf{LRA}^D(s) + \int_{paths} \Bigg[\lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \\
&\quad \cdot r^{D(\pi[j])}_{\pi[j]} \cdot \mathbb{P}^{D(\pi[j])}_{\pi[j]}(\pi[j+1]) \Big] \Pr_D(d\pi) \Bigg] \\
&= \sup_D \sum_{s\in S} \Bigg[\rho(s) \cdot \mathsf{LRA}^D(s) + r^{D(s)}_s \cdot \int_{paths} \Bigg[\lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \\
&\quad \cdot \mathbb{P}^{D(\pi[j])}_{\pi[j]}(\pi[j+1]) \Big] \Pr_D(d\pi) \Bigg]
\end{aligned}
$$

Since π is constant within the integral it follows that $\mathbb{P}^{D(\pi[j])}_{\pi[j]}(\pi[j+1]) = 1$.

$$
= \sup_D \sum_{s\in S} \left[\rho(s) \cdot \mathsf{LRA}^D(s) + r^{D(s)}_s \cdot \int_{paths} \lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \Pr_D(d\pi) \right]
$$

Now we split the equation in two cases, for $s \in MS$ and $s \in PS$

$$
\begin{aligned}
&= \sup_D \sum_{s\in MS} \left[\rho(s) \cdot \mathsf{LRA}^D(s) + r^{D(s)}_s \cdot \int_{paths} \lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \Pr_D(d\pi) \right] \\
&\quad + \sum_{s\in PS} \left[\rho(s) \cdot \mathsf{LRA}^D(s) + r^{D(s)}_s \cdot \int_{paths} \lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \Pr_D(d\pi) \right]
\end{aligned}
$$

First we consider the case for $s \in MS$:

$$
\begin{aligned}
\mathsf{LRR}^D_{MS} &= \sup_D \sum_{s\in MS} \left[\rho(s) \cdot \mathsf{LRA}^D(s) + r^{D(s)}_s \cdot \int_{paths} \lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j]) \Pr_D(d\pi) \right] \\
&= \sup_D \sum_{s\in MS} \Bigg[\rho(s) \cdot \mathsf{LRA}^D(s) + r^{D(s)}_s \cdot \int_{paths} \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbf{1}_s(\pi@u)du \\
&\quad \cdot \lim_{t\to\infty} \frac{1}{t} \frac{\sum_{j=0}^{|\pi@t|-1} \mathbf{1}_s(\pi[j])}{\int_0^t \mathbf{1}_s(\pi@u)du} \Pr_D(d\pi) \Bigg] \\
&= \sup_D \sum_{s\in MS} \left[\rho(s) \cdot \mathsf{LRA}^D(s) + r^{D(s)}_s \cdot \int_{paths} \mathsf{LRA}(s,\pi) \cdot \mathbf{R}(s,\pi) \Pr_D(d\pi) \right]
\end{aligned}
$$

Note that $\mathsf{LRA}(s,\pi)$ denotes the long-run average of s on path π and $\mathbf{R}(s,\pi)$ the exit rate for s on path π. Now let $P = \{\pi | \pi \in \mathit{paths} \wedge \mathsf{LRA}(s,\pi) = \mathsf{LRA}^D(s)\}$.

$$= \sup_D \sum_{s \in MS} \Big[\rho(s) \cdot \mathsf{LRA}^D(s) + r_s^{D(s)} \cdot \Big(\int_P \mathsf{LRA}(s,\pi) \cdot \mathbf{R}(s,\pi) \Pr_D(d\pi) + \int_{\mathit{paths} \setminus P} \mathsf{LRA}(s,\pi) \cdot \mathbf{R}(s,\pi) \Pr_D(d\pi) \Big) \Big]$$

Let $P_\epsilon(t) = \{\pi | \pi \in \mathit{paths} \wedge |\mathsf{LRA}(s,\pi^t) - \mathsf{LRA}^D(s)| \leq \epsilon\}$, such that $P = \lim_{\epsilon \to 0} \lim_{t \to \infty} P_\epsilon(t)$. For all paths $\pi \in \mathit{paths}$ it holds that $\forall \epsilon > 0. \lim_{t \to \infty} \Pr(|\mathsf{LRA}(s,\pi^t) - \mathsf{LRA}(s)| \geq \epsilon) = 0$. Thus, $\forall \epsilon > 0. \lim_{t \to \infty} \Pr(\pi \notin P_\epsilon(t)) = 0$, and $\lim_{\epsilon \to 0} \lim_{t \to \infty} \Pr(\pi \in P_\epsilon(t)) = 1$. Hence, $\Pr(P) = 1$ and $\Pr(\mathit{paths} \setminus P) = 0$.

$$\begin{aligned}
&= \sup_D \sum_{s \in MS} \Big[\rho(s) \cdot \mathsf{LRA}^D(s) + r_s^{D(s)} \cdot \int_P \mathsf{LRA}(s,\pi) \cdot \mathbf{R}(s,\pi) \Pr_D(d\pi) + 0 \Big] \\
&= \sup_D \sum_{s \in MS} \Big[\rho(s) \cdot \mathsf{LRA}^D(s) + r_s^{D(s)} \cdot \int_P \mathsf{LRA}^D(s) \cdot \mathbf{R}(s,\pi) \Pr_D(d\pi) \Big] \\
&= \sup_D \sum_{s \in MS} \Big[\rho(s) \cdot \mathsf{LRA}^D(s) + r_s^{D(s)} \cdot \mathsf{LRA}^D(s) \cdot \int_P \mathbf{R}(s,\pi) \Pr_D(d\pi) \Big] \\
&= \sup_D \sum_{s \in MS} \Big[\rho(s) \cdot \mathsf{LRA}^D(s) + r_s^{D(s)} \cdot \mathsf{LRA}^D(s) \cdot E(s) \Big]
\end{aligned}$$

The second case is for $s \in PS$:

$$\begin{aligned}
\mathsf{LRR}^D_{PS} &= \sup_D \sum_{s \in PS} \left[\rho(s) \cdot \mathsf{LRA}^D(s) + r_s^{D(s)} \cdot \int_{paths} \lim_{t \to \infty} \frac{1}{t} \sum_{j=0}^{|\pi @ t|-1} \mathbf{1}_s(\pi[j]) \Pr_D(d\pi) \right] \\
&= \sup_D \sum_{s \in PS} \left[r_s^{D(s)} \cdot \int_{paths} \lim_{t \to \infty} \frac{1}{t} \sum_{j=0}^{|\pi @ t|-1} \mathbf{1}_s(\pi[j]) \Pr_D(d\pi) \right]
\end{aligned}$$

Let $\mathcal{F}_s : S \to \mathbb{R}_{\geq 0}$ be the random variable defining the frequency of visiting a state s such that $\mathcal{F}_s(\pi) = \lim_{t \to \infty} \frac{1}{t} \sum_{j=0}^{|\pi @ t|-1} \mathbf{1}_s(\pi[j])$ is the frequency of s on path π.

$$\begin{aligned}
&= \sup_D \sum_{s \in PS} \left[r_s^{D(s)} \cdot \int_{paths} \mathcal{F}_s(\pi) \Pr_D(d\pi) \right] \\
&= \sup_D \sum_{s \in PS} r_s^{D(s)} \cdot \sum_{s' \in S} \mathbb{P}_{s'}^{D(s')}(s) \cdot \mathbb{E}_{s',D}(\mathcal{F}_s) \\
&= \sup_D \sum_{s \in PS} r_s^{D(s)} \cdot \left(\sum_{s' \in PS} \mathbb{P}_{s'}^{D(s')}(s) \cdot \mathbb{E}_{s',D}(\mathcal{F}_s) + \sum_{s' \in MS} \mathbb{P}_{s'}^{D(s')}(s) \cdot \mathbb{E}_{s',D}(\mathcal{F}_s) \right) \\
&= \sup_D \sum_{s \in PS} r_s^{D(s)} \cdot \left(\sum_{s' \in PS} \mathbb{P}_{s'}^{D(s')}(s) \cdot \mathbb{E}_{s',D}(\mathcal{F}_s) + \sum_{s' \in MS} \frac{\mathbf{R}(s',s)}{E(s')} \cdot \mathsf{LRA}^D(s') \cdot E(s') \right) \\
&= \sup_D \sum_{s \in PS} r_s^{D(s)} \cdot \left(\sum_{s' \in PS} \mathbb{P}_{s'}^{D(s')}(s) \cdot \mathbb{E}_{s',D}(\mathcal{F}_s) + \sum_{s' \in MS} \mathbf{R}(s',s) \cdot \mathsf{LRA}^D(s') \right) \\
&= \sup_D \sum_{s \in PS} r_s^{D(s)} \cdot \nu^D(s)
\end{aligned}$$

Hence it follows:

$$\mathsf{LRR}^{\max}_{\mathcal{M}} = \sup_{D} \sum_{s \in S} \left[\rho(s) \cdot \mathsf{LRA}^{D}(s) + r_s^{D(s)} \cdot \nu^{D}(s) \right]$$

□

E Proof of Theorem 5

Theorem 5 *The long-run average reward of a unichain MRA coincides with the limit of the time-bounded expected cumulative reward, such that* $\mathsf{LRR}^{D}(s) = \lim_{t\to\infty} \frac{1}{t}\mathcal{R}^{D}(s,t)$.

Proof. We show a sketch proof that the definition of LRR coincides with the definition of the cumulative reward in the long-run.

$$\begin{aligned}\mathsf{LRR}^{\mathsf{D}}_{\mathcal{M}}(s) = \mathbb{E}_{s,D}(\mathcal{L}_{\mathcal{M}}) &= \int_{paths} \mathcal{L}_{\mathcal{M}}(\pi) \Pr_{s,D}(d\pi) \\ &= \int_{paths} \lim_{t\to\infty} \frac{1}{t} \cdot reward(\pi, t) \Pr_{s,D}(d\pi)\end{aligned}$$

Let $\sum_{k=1}^{n} g(\pi, n) = reward(\pi, n)$ with $g(\pi, 1) = reward(\pi, 1)$ and $g(\pi, k) = reward(\pi, k) - reward(\pi, k-1)$ for $k > 1$.

$$\begin{aligned}&= \int_{paths} \lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{t} g(\pi, k) \Pr_{s,D}(d\pi) \\ &= \lim_{t\to\infty} \frac{1}{t} \int_{paths} \sum_{k=1}^{t} g(\pi, k) \Pr_{s,D}(d\pi) \\ &= \lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{t} \int_{paths} g(\pi, k) \Pr_{s,D}(d\pi) \\ &= \lim_{t\to\infty} \frac{1}{t} \int_{paths} reward(\pi, t) \Pr_{s,D}(d\pi) \\ &= \lim_{t\to\infty} \frac{1}{t} \cdot \mathcal{R}^{D}(s,t).\end{aligned}$$

□