Towards Analysis of Semi-Markov Decision Processes

Taolue Chen[1,⋆] and Jian Lu[2,⋆⋆]

[1] FMT, University of Twente, The Netherlands
[2] State Key Laboratory of Novel Software Technology, Nanjing University, China

Abstract. We investigate Semi-Markov Decision Processes (SMDPs). Two problems are studied, namely, the time-bounded reachability problem and the long-run average fraction of time problem. The former aims to compute the maximal (or minimum) probability to reach a certain set of states within a given time bound. We obtain a Bellman equation to characterize the maximal time-bounded reachability probability, and suggest two approaches to solve it based on discretization and randomized techniques respectively. The latter aims to compute the maximal (or minimum) average amount of time spent in a given set of states during the long run. We exploit a graph-theoretic decomposition of the given SMDP based on maximal end components and reduce it to linear programming problems.

1 Introduction

Markov decision processes (MDPs, [9]) provide a mathematical framework for modeling decision-making in situations where outcomes are partly random and partly under the control of a decision maker. In this paper, we consider *Semi-Markov Decision Processes* (SMDPs, [7]), which generalize MDPs by (1) allowing the decision maker to choose actions whenever the system state changes; (2) modeling the system evolution in *continuous* time; and (3) allowing time spent in a particular state to follow an arbitrary probability distribution [9].

This paper deals with the analysis of SMDPs. Here, we investigate two important criteria, which unfortunately received relatively scant attentions so far. They are *time-bounded reachability probabilities* and *long-run average fraction of time*. We elaborate them as follows:

- The first problem is referred to as the *time-bounded reachability problem*. Namely, given an SMDP, the aim is to compute the maximal (or minimum) probability to reach – under a given class of schedulers – a certain set of state G within T time units. To solve this problem, we propose a Bellman equation to characterize the maximal time-bounded reachability probability, and suggest two approaches to solve it, based on the discretization and the randomized technique respectively. The dual minimum probability problem can be solved accordingly.

[⋆] This author is partially supported by the SRO DSN project of CTIT, University of Twente, and an oversea research funding from the State Key Laboratory of Novel Software Technology.

[⋆⋆] This author is partially supported by the Chinese national 973 program (2009CB320702), NSFC (60736015, 60721002), and JSNSF (BK2008017).

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 41–48, 2010.

- The second problem is referred to as the *long-run average fraction of time problem*. Namely, given an SMDP, the aim is to compute the maximal (or minimum) average amount of time – under a given class of schedulers – spent in a given set of states B during the long run of the SMDP. To solve this problem, we exploit a graph-theoretic decomposition of the given SMDP based on the notion of maximal end components and thus reduce it to linear programming problems. The dual minimum probability problem can be also solved along the same vein.

Related works. Time-bounded reachability problem has been tackled in [2] and recently in [8] for continuous-time MDPs; a very similar approach has been applied to model checking continuous stochastic logic (CSL, [1]) for a closely related model akin to CTMDP, i.e., interactive Markov chains (IMCs) in [12]. In this paper, we generalize this work to SMDPs. Long-run average fraction of time problem can be considered as a special case of average reward problem, which has been studied extensively; see e.g. [6,9]. However, most of studies focus on a unichain model, which is rather restrictive. In this paper, we extend this to a general SMDP.

2 Preliminarily

Given a set H, let $\Pr : \mathcal{F}(H) \to [0,1]$ be a probability measure on the *measurable space* $(H, \mathcal{F}(H))$, where $\mathcal{F}(H)$ is a *σ-algebra* (a.k.a. *σ-field*) over H. A semi-Markov decision process is given by:

Definition 1 (SMDP). *A* semi-Markov decision process *(SMDP) $\mathcal{M}$ is a tuple $(S, s_0, A, \mathbf{P}, \mathbf{Q})$, where*

- *S is a* finite *set of states, with $s_0 \in S$ the initial state;*
- *A is a* finite *set of actions;*
- *$\mathbf{P} : S \times A \times S \to [0,1]$ is the transition probability function, satisfying that for each $s \in S$ and $\alpha \in A$, $\sum_{s' \in S} \mathbf{P}(s, \alpha, s') \in \{0,1\}$.*
- *$\mathbf{Q} : S \times A \times S \to (\mathbb{R}_{\geq 0} \to [0,1])$ is the continuous probability distribution function. We use $\mathbf{Q}(s, a, s', dt)$ to represent a time-differential.*

Typically, $\mathbf{Q} : S \times A \times S \to (\mathbb{R}_{\geq 0} \to [0,1])$ returns a *cumulative distribution function* (cdf), such that $\mathbf{P}(s, \alpha, s') = 0$ implies that $\mathbf{Q}(s, \alpha, s')(t) = 1$. We write $\mathcal{Q} : S \times A \times S \to (\mathbb{R}_{\geq 0} \to [0,1])$ for the corresponding *probability density function* (pdf). Namely, $\mathbf{Q}(s, \alpha, s')(t) = \int_0^t \mathcal{Q}(s, \alpha, s')(\tau) d\tau$; or $\mathbf{Q}(s, \alpha, s', dt) = \mathcal{Q}(s, \alpha, s')(\tau) d\tau$. Usually A is ranged over by $\alpha, \beta, \ldots$. With a bit abuse of notations, for each state s, we write $A(s) = \{\alpha \in A \mid \mathbf{P}(s, \alpha, s') > 0 \text{ for some } s'\}$, i.e., the set of actions which are enabled from s. Clearly, $\mathbf{P}(s, \alpha, s') = 0$ for any s' if $\alpha \notin A(s)$. We assume w.l.o.g. that only internal nondeterminism is allowed, i.e., the actions enabled at each state are pairwise different.

The distribution function $\mathbf{H}$ of state s, defined by $\mathbf{H}(t \mid s, \alpha) = \sum_{s' \in S} \mathbf{P}(s, \alpha, s') \cdot \mathbf{Q}(s, \alpha, s')(t)$, denotes the total holding time distribution in s under the action α, regardless of which successor is selected. To avoid the possibility of an infinite number of decision epochs within finite time (i.e., the zeno behavior), we impose the following assumption, similar to [9]: There exists $\varepsilon > 0$ and $\delta > 0$ such that $\mathbf{H}(\delta \mid s, \alpha) \leq 1 - \varepsilon$, for all $s \in S$ and $\alpha \in A$.

2.1 Semantics

As a stochastic model, in general one is interested in certain events in SMDPs. To measure the probability of events in an SMDP, we use a *path* to represent a single outcome of the associated random experiment. In the continuous-time setting, *timed* paths capture the sojourn times in each state which describe the complete trajectory of the SMDP. They are introduced below. (In the reminder of this paper, paths refer to timed paths.)

Given an SMDP $\mathcal{M}$, $Paths^n(\mathcal{M}) = S \times (A \times \mathbb{R}_{\geq 0} \times S)^n$ is the set of paths of length n in $\mathcal{M}$; the set of *finite* paths in $\mathcal{M}$ is defined as $Paths^\star(\mathcal{M}) = \bigcup_{n \in \mathbb{N}} Paths^n$, and $Paths^\omega(\mathcal{M}) = (S \times A \times \mathbb{R}_{\geq 0})^\omega$ is the set of *infinite* paths in $\mathcal{M}$. Accordingly, $Paths(\mathcal{M}) = Paths^\star(\mathcal{M}) \cup Paths^\omega(\mathcal{M})$ denotes the set of all paths in $\mathcal{M}$. When $\mathcal{M}$ is clear from the context, we denote a path $\sigma \in Paths(s_0)$ if σ starts in state s_0. The same applies to $Paths^n(s_0)$, $Paths^\omega(s_0)$ and $Paths^\star(s_0)$, respectively. Typically, a single path (of length n) σ is denoted $s_0 \xrightarrow{\alpha_0, t_0} s_1 \xrightarrow{\alpha_1, t_1} \cdots \xrightarrow{\alpha_{n-1}, t_{n-1}} s_n$, where $|\sigma| = n$ is the length of σ and $\sigma[\downarrow] = s_n$ is the last state of σ. For $k \leq |\sigma|$, $\sigma[k]$ is the $(k+1)$-th state on σ; $\sigma\langle k \rangle := t_k$ is the time spent in state s_k. If $i < j \leq |\sigma|$, then $\sigma[i..j]$ denotes the path-infix $s_i \xrightarrow{\alpha_i, t_i} s_{i+1} \cdots \xrightarrow{\alpha_{j-1}, t_{j-1}} s_j$ of σ. Let $\sigma @ t$ be the state occupied in σ at time $t \in \mathbb{R}_{\geq 0}$, i.e. $\rho @ t := \sigma[n]$ where n is the smallest index such that $\sum_{i=0}^{n} \sigma\langle i \rangle > t$. For infinite path $\sigma = s_0 \xrightarrow{\alpha_0, t_0} s_1 \xrightarrow{\alpha_1, t_1} \cdots$, we require time-divergence, i.e., $\sum_{i \geq 0} t_i = \infty$.

Our goal is to measure the probabilities of (measurable) sets of paths. To this end, we first define a σ-algebra of sets of *combined transitions* [11] which we later use to define σ-algebras of sets of finite and infinite paths. Intuitively, a combined transition is a tuple (α, t, s') which entangles the decision for action α for the time t after which the SMDP moves to successor state s'. Formally, given an SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, let $\Omega = A \times \mathbb{R}_{\geq 0} \times S$ be the set of combined transitions in $\mathcal{M}$. To define a measurable space on Ω, note that S and A are finite; hence, the corresponding σ-algebras are defined as their power sets respectively. And we use the *Borel σ-field* $\mathcal{B}(\mathbb{R}_{\geq 0})$ to measure the corresponding subsets of $\mathbb{R}_{\geq 0}$. Recall that a Cartesian product is a measurable rectangle if its constituent sets are elements of their respective σ-algebras. We then use $2^A \otimes \mathcal{B}(\mathbb{R}_{\geq 0}) \otimes 2^S$ to denote the set of all measurable rectangles. It generates the desired σ-algebra $\Im = \sigma(2^A \otimes \mathcal{B}(\mathbb{R}_{\geq 0}) \otimes 2^S)$ of sets of combined transitions, which can be used to infer the σ-algebra $\mathcal{F}(Paths^n(\mathcal{M}))$ of sets of paths of length n: $\mathcal{F}(Paths^n(\mathcal{M}))$ is generated by the set of measurable rectangles, that is $\mathcal{F}(Paths^n(\mathcal{M})) = \sigma(\{S_0 \times M_1 \times \cdots \times M_n \mid S_0 \subseteq S, M_i \in \Im\})$. The σ-algebra of sets of *infinite* paths is obtained, in turn, by applying the standard *cylinder* set construction. A set C^n of *paths of length* n is called a cylinder base; it induces the infinite cylinder $C_n = \{\pi \in Paths^\omega(\mathcal{M}) \mid \pi[0..n] \in C^n\}$. A cylinder C_n is measurable if $C^n \in \mathcal{F}(Paths^n(\mathcal{M}))$; C_n is called an infinite rectangle if $C^n = S_0 \times A_0 \times T_0 \times \cdots \times A_{n-1} \times T_{n-1} \times S_n$ and $S_i \subseteq S$, $A_i \subseteq A$ and $T_i \subseteq \mathbb{R}_{\geq 0}$. It is a measurable infinite rectangle, if $S_i \in 2^S$, $A_i \in 2^A$ and $T_i \in \mathcal{B}(\mathbb{R}_{\geq 0})$. We obtain the desired σ-algebra of sets of infinite paths $\mathcal{F}(Paths^\omega(\mathcal{M}))$ as the minimal σ-algebra generated by the set of measurable cylinders.

Schedulers. As in MDPs, the decision making in SMDPs is specified by *schedulers* (or policies). It also defines the semantics of SMDPs, namely, an SMDP and an associated scheduler induce a unique *probability measure* on the measurable spaces $(\mathit{Paths}^{\omega}(\mathcal{M}), \mathcal{F}(\mathit{Paths}^{\omega}(\mathcal{M})))$. A scheduler quantifies the probability of the next action: If state s is reached via a finite path σ, the scheduler yields a probability distribution over $A(\sigma[\downarrow])$. In the timed setting, the notion of *measurable schedulers* [11] has to be adopted. A measurable scheduler can incorporate the complete information from the history that led into the current state when deciding which of the next actions to take; in particular, it may yield different decisions depending on the time that has passed or in single states.

Definition 2 (Measurable scheduler). *Let $\mathcal{M}$ be an SMDP. A mapping $D : \mathit{Paths}^{\star}(\mathcal{M}) \times A \to [0,1]$ is a* measurable scheduler *if the functions $D(\cdot, \alpha) : \mathit{Paths}^{\star}(\mathcal{M}) \to [0, 1]$ are measurable for any $\alpha \in A$.*

We now embark on defining a probability measure on the σ-algebra $\mathcal{F}(\mathit{Paths}^{\omega}(\mathcal{M}))$. To this aim, we first define a probability measure μ_D on the set of combined transitions, i.e., on the measurable space $(\Omega, \Im)$. For all $\sigma \in \mathit{Paths}^{\star}(\mathcal{M})$, $\mu_D(\sigma, M) = \sum_{\alpha \in A} \sum_{s' \in S} D(\sigma, \alpha) \int_{\mathbb{R}_{\geq 0}} \mathbf{1}_M(\alpha, t, s') \mathbf{Q}(s, \alpha, s', dt)$, where $\mathbf{1}_M(\alpha, t, s')$ is the indicator for the set $M \subseteq \Omega$, that is, $\mathbf{1}_M(\alpha, t, s') = 1$ if the combined transition $(\alpha, t, s') \in M$ and 0 otherwise.

With the help of μ_D, we can define the probability measure on the sets of finite paths inductively as follows: $\mathrm{Pr}^0_{s_0,D} : \mathcal{F}(\mathit{Paths}^0(\mathcal{M})) \to [0, 1]$ is given as $\mathrm{Pr}^0_{s_0,D}(\Pi) = 1$ if $s_0 \in \Pi$; 0 otherwise. (Note that $\mathit{Paths}^0(\mathcal{M}) \subseteq S$.) And $\mathrm{Pr}^{n+1}_{s_0,D} : \mathcal{F}(\mathit{Paths}^{n+1}(\mathcal{M})) \to [0, 1]$ is given as $\mathrm{Pr}^{n+1}_{s_0,D}(\Pi) = \int_{\mathit{Paths}^n(\mathcal{M})} \mathrm{Pr}^n_{s_0,D}(d\pi) \int_{\Omega} \mathbf{1}_{\Pi}(\pi \cdot m) \mu_D(\pi, dm)$. Intuitively, it derives the probability $\mathrm{Pr}^{n+1}_{s_0,D}$ on sets of paths Π of length $n + 1$ by multiplying the probability $\mathrm{Pr}^n_{s_0,D}(d\pi)$ of a pat π of length n with the probability $\mu_D(\pi, dm)$ of a combined transition m such that the concatenation $\pi \cdot m$ is a path from the set Π. Hence we obtain measures on all σ-algebras $\mathcal{F}(\mathit{Paths}^n(\mathcal{M}))$ of subsets of paths of length n. This extends to a measure on $(\mathit{Paths}^{\omega}(\mathcal{M}), \mathcal{F}(\mathit{Paths}^{\omega}(\mathcal{M})))$ as follows: First, note that any measurable cylinder can be represented by a base of finite length, i.e., $B_n = \{\sigma \in \mathit{Paths}^{\omega}(\mathcal{M}) \mid \sigma[0..n] \in B^n\}$. Now the measures $\mathrm{Pr}^n_{s_0,D}$ on $\mathcal{F}(\mathit{Paths}^n(\mathcal{M}))$ extend to a unique probability measure $\mathrm{Pr}^{\omega}_{s_0,D}$ on $\mathcal{F}(\mathit{Paths}^{\omega}(\mathcal{M}))$ by defining $\mathrm{Pr}^{\omega}_{s_0,D}(B_n) = \mathrm{Pr}^n_{s_0,D}(B^n)$. The Ionescu-Tulcea extension theorem is applicable due to the inductive definition of the measures $\mathrm{Pr}^n_{s_0,D}$ and assures the extension to be well defined and unique.

3 Time-Bounded Reachability

In this section, we focus on the problem of computing the maximal time-bounded reachability probability in SMDPs. Namely, given an SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, a set of goal states $G \subseteq S$, a time bound T, we ask: what is the maximal probability to reach G within T? Formally, Let $\Diamond^{[0,T]}G$ denote a set of infinite paths of $\mathcal{M}$ which reach G within time bound T, starting from the initial state s_0. We intend to compute $p_{\max} = \sup_D \{\mathrm{Pr}^{\omega}_{s_0,D}(\Diamond^{[0,T]}G)\}$.

The basic idea to solve this problem is to "encode" the time into the state space, and then to resort to the reachability problem in MDPs. Generally, given a (continuous-state) MDP $(\Xi, A, \mathcal{P})$ with a set of goal states $\Upsilon \in \mathcal{F}(\Xi)$, the maximal reachability probability can be computed via the following celebrated Bellman equation [3]: $V(\xi) = \max_{\alpha \in A} \left\{ \int_{\Xi} \mathcal{P}((\xi), \alpha, d\xi') \cdot V(\xi') \right\}$ if $\xi \notin \Upsilon$; and 1 otherwise. By instantiating this equation with $\Xi = S \times \mathbb{R}_{\geq 0}$ and $\Upsilon = G \times [0, T]$, we obtain that

- if $s \in G$ and $x \leq T$, $V(s, x) = 1$;
- if $s \in G$ and $x > T$, $V(s, x) = 0$; and
- if $s \notin G$ and $x \leq T$, $V(s, x) =$

$$\begin{aligned} &\max\nolimits_{\alpha \in A} \left\{ \int_{S \times \mathbb{R}_{\geq 0}} \mathcal{P}((s, x), \alpha, d\xi') \cdot V(\xi') \right\} \\ = &\max\nolimits_{\alpha \in A} \left\{ \sum\nolimits_{s' \in S} \int_0^{T-x} \mathbf{P}(s, \alpha, s') \cdot \mathbf{Q}(s, s', \alpha, dt) \cdot V(s', x + t) \right\} \end{aligned} \tag{1}$$

We note that in Eq.(1), the time is encoded in an increasing manner; however, it would be convenient to use countdown, namely, forcing the time x to decrease. In this way, one restricts the state space to $S \times [0, T]$. Moreover, as a convention, we write Eq.(1) in a functional form. We summarize the main result of this section as the following theorem.

Theorem 1. *Let $\mathcal{M} = (S, A, \mathbf{P}, \mathbf{Q})$ be an SMDP and $G \subseteq S$ be a set of goal states. Let $p_{\max} = \sup_D \{\Pr^{\omega}_{s_0, D}(\Diamond^{[0,T]} G)\}$, i.e. the maximal probability to reach G within time bound T, starting from the initial state s_0. Then $p_{\max} = \mathrm{lfp}(\Omega)(s_0, T)$ where $\Omega : (S \times \mathbb{R}_{\geq 0} \to [0, 1]) \to (S \times \mathbb{R}_{\geq 0} \to [0, 1])$ is an operator given as*

$$\Omega(F)(s, z) = \max_{\alpha \in A} \left\{ \sum_{s' \in S} \int_0^z \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \cdot F(s', z - t) \right\} \tag{2}$$

and $\mathrm{lfp}(\Omega)$ denotes the least fixpoint of the functional Ω.

Computational methods. Our next task is solve the proposed Bellman equation Eq. (2). A naïve approach is to apply Picard iteration, which is unfortunately, quite inefficient. Instead, we propose two approaches: one is based on discretization; the other one is based on randomization.

Discretization. Assume a step size $h = \frac{1}{N}$ for some $N \in \mathbb{N}$. From Eq. (2), we have

$$\begin{aligned} F(s, T) &= \max_{\alpha \in A} \left\{ \sum_{s' \in S} \int_0^T \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \cdot F(s', T - t) \right\} \\ &= \max_{\alpha \in A} \left\{ \sum_{i=0}^{N-1} \sum_{s' \in S} \int_{i \cdot h}^{(i+1) \cdot h} \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \cdot F(s', T - t) \right\} \\ &\approx \max_{\alpha \in A} \left\{ \sum_{i=0}^{N-1} \sum_{s' \in S} \left(\int_{i \cdot h}^{(i+1) \cdot h} \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \right) \cdot F(s', T - i \cdot h) \right\}. \end{aligned}$$

Hence we can introduce $\widetilde{F}$ as an approximation of F:

$$\widetilde{F}(s, T) = \max_{\alpha \in A} \left\{ \sum_{i=0}^{N-1} \sum_{s' \in S} \left(\int_{i \cdot h}^{(i+1) \cdot h} \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \right) \cdot \widetilde{F}(s', T - i \cdot h) \right\},$$

which can be solved by linear programming [3].

Unfortunately, due to the diversity of potential distribution functions used in SMDP, we are not able to derive a plausible error bound between F and $\tilde{F}$, which we left for the future work.

Randomization. Inspired by [10], we proposed a *random* Bellman operator to solve Eq.(2). Suppose $\{x_1, \cdots, x_N\}$ are IID draws with respect to Lebesgue measure λ from the interval $[0, T]$, we define

$$\tilde{\Omega}(F)(s, x_i) = \max_{\alpha \in A} \left\{ \frac{1}{N} \sum_{j=1}^{N} \mathbf{P}(s, \alpha, s') \cdot \mathcal{Q}(s, \alpha, s')(x_j) \cdot F(s', x_j) \right\}. \qquad (3)$$

It turns out that Eq.(3) constitutes an approximation of Eq.(2), which is much easier to solve [10]. Note that the operator $\widetilde{\Omega}_N$ is self-approximating: for any function F, one can evaluate $\widetilde{\Omega}_N(F)(s, x)$ at any point $s \in S$ and $x \in [0, T]$ without requiring any explicit interpolation of the values of $\widetilde{\Omega}_N(s, x)$ at the random sample points $x \in \{x_1, \cdots, x_N\}$. In particular, since $\mathcal{Q}$ is a continuous function, $\widetilde{\Gamma}_N(V)$ is a (random) continuous function of (s, x) and evaluation of this function at any particular point (s, x) involves nothing more than evaluating the simple formula on the RHS of Eq. (3).

4 Long-Run Average Fraction of Time

In this section, we turn to optimizing another measure, the average amount of time spent in a set of states. Given an SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, we fix a scheduler D, and let σ be a path taken *randomly* from the set $Paths^{\omega}(s_0)$. For $B \subseteq S$, let $\mathbf{1}_B$ denote an indicator with $\mathbf{1}_B(s) = 1$ if $s \in B$ and 0 otherwise. Then the quantity $\mathbf{1}_B(\sigma@x)$ is random variable, indicating whether states in B is occupied at time t when starting in s_0. Based on this, we then define a random variable that cumulates the time spent in some state in B up to time t, starting from state s and normalize it by the time t in order to obtain a measure of the *fraction* of time spent in states form B up to time t, namely, $\text{avg}_{B,t}^{\mathcal{M}}(\sigma) = \frac{1}{t}\int_0^t \mathbf{1}_B(\sigma@x)dx$. Since $\text{avg}_{B,t}^{\mathcal{M}}$ is still a random variable, we can derive its expectation given the scheduler D (which results in $\Pr_{s_0,D}^{\omega}$). This value corresponds to the average fraction of time spent in states from B in the time frame up to t. For the long-run average faction of time, we consider the limit $t \longrightarrow \infty$, as

$$\lim_{t \longrightarrow \infty} \mathbb{E}(\text{avg}_{B,t}^{\mathcal{M}}) = \lim_{t \longrightarrow \infty} \int_{Paths^{\omega}(s_0)} \text{avg}_{B,t}^{\mathcal{M}}(\pi) \Pr_{s_0,D}^{\omega}(d\pi). \qquad (4)$$

We want to maximize this quantity w.r.t. measurable schedulers. Namely, to compute $p_{\max} = \sup_D \lim_{t \rightarrow \infty} \mathbb{E}(\text{avg}_{B,t}^{\mathcal{M}})$. The rest of this section is devoted to solving this problem. Some standard notions, in particular, the (maximal) *end component*, can be found in [5]. Intuitively, they represent sets of state-action pairs (below, each one is denoted by (C, ∇) where $C \subseteq S$ and $\nabla \subseteq A$) that once entered, can be followed forever if the scheduler chooses the actions in an appropriate way. Namely, for any scheduler D, a behavior will end up with probability 1 in an end component. (See [5] for a formal account.) Given SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, let $\text{MEC}(\mathcal{M})$ be the set of all *maximal end components* (**MECs**) of $\mathcal{M}$. Note that any two MECs are disjoint. One can identify all of these maximal end components via graph-based analysis [5].

Lemma 1. *Given any SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$ and scheduler D,*

$$\lim_{t\to\infty} \mathbb{E}(\mathrm{avg}^{\mathcal{M}}_{B,t}) = \sum_{H\in\mathrm{MEC}(\mathcal{M})} \mathit{Prob}_D(s_0, \Diamond H) \cdot \lim_{t\to\infty} \mathbb{E}(\mathrm{avg}^{H}_{B,t}).$$

In light of Lem.1, we focus on each MEC of $\mathcal{M}$, which corresponds to a *unichain*[1]. As in [9], we define the long-run average fraction of time alternatively as

$$\lim_{n\to\infty} \frac{\mathbb{E}^D_{s_0}\left\{\sum_{i=0}^{n}\int_{\tau_i}^{\tau_{i+1}} \mathbf{1}_B(X_i)dt\right\}}{\mathbb{E}^D_{s_0}\left\{\sum_{i=0}^{n}\tau_i\right\}}, \tag{5}$$

where X_i and τ_i are random variables, defined as $X_i(\sigma) = s_i$, $\tau_i(\sigma) = t_i$, given an infinite path $\sigma = s_0 \xrightarrow{\alpha_0, t_0} s_1 \xrightarrow{\alpha_1, t_1} \cdots$.

It turns out that in our setting, these two criteria, given in Eq.(4) and Eq.(5) respectively, coincide. (A proof can be found in [9].) For each $s \in S$ and $\alpha \in A$, we proceed to introduce two measures: (1) $r(s,\alpha) = \int_0^\infty \sum_{s'\in S} \mathbf{P}(s,\alpha,s') \left(\int_0^t \mathbf{1}_B(s)du\right) \mathbf{Q}(s,a, s', dt)$, denoting the expected total "reward" between two decision epochs, given that the system occupies state s at the first decision epoch and the decision maker chooses action α; and (2) $y(s,\alpha) = \mathbb{E}[\mathbf{H}(s,\alpha)]$, denoting the expected length of time until the next decision epoch, given that action α is chosen in state s at the current decision epoch. Note that if $s \in B$, then $r(s,\alpha) = \int_0^\infty \sum_{s'\in S} \mathbf{P}(s,\alpha,s') \cdot t \cdot \mathbf{Q}(s,a,s',dt)$, and if $s \notin B$, $r(s,\alpha) = 0$.

We then introduce a Bellman equation to cope with Eq.(5). For each $s \in S$,

$$h_s = \max_{\alpha\in A}\left\{ r(s,a) - \lambda y(s,a) + \sum_{s'\in S} \mathbf{P}(s,a,s')h_{s'} \right\} . \tag{6}$$

It is known [9] that they have at least one solution, and the value of λ at the solutions is what we desire. Eq.(6) can be solved via linear programming by introducing variables $\{\lambda\} \cup \{h_s \mid s \in S\}$, as follows

$$\begin{array}{l} \text{maximize } \lambda \\ \text{s.t. } h_s \le r(s,\alpha) - \lambda y(s,\alpha) + \sum_{s'\in S} \mathbf{P}(s,\alpha,s')h_{s'} \ . \end{array} \tag{7}$$

It turns out [9] that the linear programming given in (7) admits at least one solution, and the variable λ assumes the same value at all solutions.

Now for each MEC H, we can compute the maximum λ_H thanks to the linear programming given in Eq.(7), which paves the way to solve the problem of computing the maximal long-run average fraction of time by reducing to the problem of computing the *maximal expected reachability reward* in an MDP, which is a minor variant of *stochastic shortest path problem* [4]. Let $U = \{s \in H \mid H \in \mathrm{MEC}(\mathcal{M})\}$, i.e., the set of states belonging to some MEC H of $\mathcal{M}$. Since any two MECs are disjoint, for each state

[1] Meaning that for each deterministic stationary scheduler, the resulting discrete-time Markov chain consists of a single recurrent class plus a possibly empty set of transient states.

$s \in U$, we can associate a value λ_H where H is the MEC to which s belongs. We then introduce the following linear programming with a family of variables $\{x_s \mid s \in S \backslash U\}$:

$$\begin{array}{l} \text{minimize } \sum_{s \in S \backslash U} x_s \\ \text{s.t. } x_s \geq \sum_{s' \in S \backslash U} \mathbf{P}(s, \alpha, s') \cdot x_{s'} + \sum_{s' \in U} \mathbf{P}(s, \alpha, s') \cdot \lambda_H. \end{array} \tag{8}$$

x_{s_0} is what we desire.

Algorithm. We summarize the procedure outlined before in Algo. 1.

Algorithm 1

```
Require: An SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, a set of states $B \subseteq S$
Ensure: $\sup_D \lim_{t \to \infty} \mathbb{E}(\mathrm{avg}^{\mathcal{M}}_{B,t})$
1: Compute $\mathrm{MEC}(\mathcal{M}) = \{(C_1, \nabla_1), \cdots, (C_n, \nabla_n)\}$;
2: for $1 \leq i \leq n$ do
3:     Compute $r(s, \alpha)$ and $y(s, \alpha)$ for $s \in C_i$ and $\alpha \in \nabla_i(s)$, depending on $B$;
4:     Solve linear programming (7);
5: end for
6: Solve linear programming (8);
7: return $x_{s_0}$ in (8).
```

References

1. Baier, C., Haverkort, B.R., Hermanns, H., Katoen, J.-P.: Model-checking algorithms for continuous-time Markov chains. IEEE Trans. Software Eng. 29(6), 524–541 (2003)
2. Baier, C., Hermanns, H., Katoen, J.-P., Haverkort, B.R.: Efficient computation of time-bounded reachability probabilities in uniform continuous-time markov decision processes. Theor. Comput. Sci. 345(1), 2–26 (2005)
3. Bertsekas, D.P.: Dynamic Programming and Optimal Control. Athena Scientific, Belmont (2007)
4. Bertsekas, D.P., Tsitsiklis, J.N.: An analysis of stochastic shortest path problems. Mathematics of Operations Research 16(3), 580–595 (1991)
5. de Alfaro, L.: How to specify and verify the long-run average behavior of probabilistic systems. In: LICS, pp. 454–465 (1998)
6. Guo, X., Hernández-Lerma, O.: Continuous-Time Markov Decision Processes. Springer, Heidelberg (2009)
7. Jewell, W.S.: Markov-renewal programming I: Formulation, finite returen models; markov-renewal programming II: infinite return models, example. Operations Research 11, 938–971 (1963)
8. Neuhaeusser, M.R.: Model Checking Nondeterministic and Randomly Timed Systems. PhD thesis (2010)
9. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York (1994)
10. Rust, J.: Using randomization to break the curse of dimensionality. Econometrica 65(3), 487–516 (1997)
11. Wolovick, N., Johr, S.: A characterization of meaningful schedulers for continuous-time markov decision processes. In: Asarin, E., Bouyer, P. (eds.) FORMATS 2006. LNCS, vol. 4202, pp. 352–367. Springer, Heidelberg (2006)
12. Zhang, L., Neuhaeusser, M.R.: Model checking interactive markov chains. In: Esparza, J., Majumdar, R. (eds.) TACAS 2010. LNCS, vol. 6015, pp. 53–68. Springer, Heidelberg (2010)