Stochastic Processes with Expected Stopping Time

Krishnendu Chatterjee
IST Austria

Laurent Doyen
CNRS & LMF, ENS Paris-Saclay, France

***Abstract*—Markov chains are the de facto finite-state model for stochastic dynamical systems, and Markov decision processes (MDPs) extend Markov chains by incorporating non-deterministic behaviors. Given an MDP and rewards on states, a classical optimization criterion is the maximal expected total reward where the MDP stops after T steps, which can be computed by a simple dynamic programming algorithm. We consider a natural generalization of the problem where the stopping times can be chosen according to a probability distribution, such that the expected stopping time is T, to optimize the expected total reward. Quite surprisingly we establish inter-reducibility of the expected stopping-time problem for Markov chains with the Positivity problem (which is related to the well-known Skolem problem), for which establishing either decidability or undecidability would be a major breakthrough. Given the hardness of the exact problem, we consider the approximate version of the problem: we show that it can be solved in exponential time for Markov chains and in exponential space for MDPs.**

I. Introduction

Stochastic models and optimization. The de facto model for stochastic dynamical systems is finite-state Markov chains [14], [15], [18], with several application domains [3]. In modeling optimization problems, rewards are associated with states of the Markov chain, and the optimization criterion is formalized as the expected total reward provided that the Markov chain is stopped after T steps [24], [14]. The extension of Markov chains to allow non-deterministic behavior gives rise to Markov decision processes (MDPs), and the optimization criterion is to maximize, over all non-deterministic choices, the expected total reward for T steps. This notion of optimization for fixed time is called *finite-horizon planning*, which has many applications in logic and verification [12], [5] and control problems in artificial intelligence and robotics [21, Chapter 10, Chapter 25], [22, Chapter 6].

Optimization with expected stopping time. In the most basic case the stopping time for collecting rewards in the stochastic model is a fixed constant T. A natural generalization is to consider that the stochastic model can be stopped at a random time such that the expectation of the stopping time is T. We consider the problem of optimizing (maximizing/minimizing) the expected total reward, when the stopping-time probability distribution can be chosen arbitrarily such that the expected stopping time is T. In other words, we consider stochastic models of Markov chains/MDPs with total reward, and instead of fixed stopping time T, we consider expected stopping time T.

978-1-6654-4895-6/21/$31.00 ©2021 IEEE

Example and motivation. Consider the classical example where a robot explores a region for natural resources (e.g., the well-studied RockSample problem in AI literature [29]), and the exploration of the robot is modeled as a Markov chain. The success of the exploration is characterized by the expected total reward, and the stopping time T denotes the expected duration of the exploration. The expected stopping-time problem asks to choose the probability distribution of the exploration duration to optimize the collected reward, satisfying the average exploration time. A classical stopping-time distribution is the exponential distribution where the stochastic model is stopped at every instant with probability λ, called *discount factor*, which entails that the expected stopping time is $T = 1/\lambda$ [14]. The discount-factor model makes an assumption on the shape of the stopping-time distribution, whereas in realistic scenarios the discount factor is not precisely known, or time-varying discount factors are considered [11]. When the discount factors are not known, then robust solutions require the worst-case choice of the factors. Thus in many examples realistic modeling requires complex stopping-time distributions, and if the precise parameters are unknown, then a robust analysis requires to consider the worst-case value of the discount factor. Hence, when the stopping-time distribution is important yet unknown, a conservative estimate (i.e., lower bound) of the optimal value is obtained using the worst-case choices. Thus we consider problems that represent robust extensions of the classical finite-horizon planning.

Previous and our results. For fixed stopping time T, the expected total reward for Markov chains and MDPs can be computed via a simple dynamic programming (or backward induction) approach [26, Chapter 4], [14], [17], [4]. Perhaps surprisingly the optimization problem for Markov chains and MDPs with expected stopping time has not been considered in the literature (to the best of our knowledge). Our main results are as follows:

- In contrast to the simple algorithm for fixed stopping time T, we show that quite surprisingly the expected stopping-time problem is *Positivity*-hard. The Positivity problem is known to be at least as hard as the well-known Skolem problem, whose decidability has been open for more than eight decades [23]. Moreover, we establish inter-reducibility between the expected stopping-time problem and the Positivity problem, and thus show that for a simple variant (adding expectation to stopping time) of the classical Markov chain problem, establishing either decidability or undecidability would be a major breakthrough.

- We then consider approximating the optimal expected total reward under the constraint that the expected stopping time is T, and show that for every additive absolute error $\varepsilon > 0$, the approximation can be achieved in time logarithmic in $1/\varepsilon$ and exponential in the size of the Markov chain.
- For MDPs we show that infinite-memory strategies are required. While the expected stopping-time problem is Positivity-hard for MDPs (since Markov chains are a special case), we show that the approximation problem can be solved in exponential space in the size of the MDP and logarithm of $1/\varepsilon$.

Comparison with related work. The optimization problem with fixed expected stopping time has been considered for the simple model of graphs [7], which is a model without stochastic aspects. The graph problem can be solved in polynomial time [7], while in sharp contrast, we show that the problem is Positivity-hard for Markov chains.

Remark 1. *The expected stopping-time problem for Markov chains has a similar flavor as probabilistic automata (or blind MDPs) [27]. In probabilistic automata a word (or letter sequence) must be provided without the information about how the probabilistic automaton executes. Similarly, for the expected stopping-time problem for Markov chains the probability distribution for stopping times must be chosen without knowing the execution of the Markov chain (in contrast to stopping criteria based on current state or accumulated reward, which rely on knowing the execution of the Markov chain). For probabilistic automata, even for basic reachability, all problems related to approximation are undecidable [20]. In contrast, we show that while the exact problem for expected stopping time in Markov chains is Positivity-hard, the approximation problem can be solved in exponential time.*

Detailed proofs are available in en extended version of this paper [8].

II. Preliminaries

A *stopping-time distribution* (or simply, a distribution) is a function $\delta : \mathbb{N} \to [0,1]$ such that $\sum_{t\in\mathbb{N}} \delta(t) = 1$. The support of δ is $\mathsf{Supp}(\delta) = \{t \in \mathbb{N} \mid \delta(t) \neq 0\}$. We denote by Δ the set of all stopping-time distributions, and by $\Delta^{\uparrow\uparrow}$ the set of all distributions δ with $|\mathsf{Supp}(\delta)| \leq 2$, called the *bi-Dirac* distributions.

The *expected utility* of a sequence $u = u_0, u_1, \ldots$ of real numbers under a distribution δ is $\mathbb{E}_\delta(u) = \sum_{t\in\mathbb{N}} u_t \cdot \delta(t)$. In particular, the expected utility of the sequence $0, 1, 2, 3, \ldots$ of all natural numbers is called the *expected time* (of distribution δ), denoted by $\mathbb{E}_\delta$.

We recall the definition of the Positivity problem and of the related Skolem problem. In the sequel, we denote by M^t_{ij} the (i, j) entry of the t-th power of matrix M (we should write it as $(M^t)_{i,j}$, but use this simpler notation when no ambiguity can arise).

Positivity problem [23], [1]. Given a square integer matrix M, decide whether there exists an integer $t \geq 1$ such that $M^t_{1,2} > 0$.

Skolem problem [23], [1]. Given a square integer matrix M, decide whether there exists an integer $t \geq 1$ such that $M^t_{1,2} = 0$.

The decidability of the Positivity and Skolem problems is a longstanding open question [23], and there is a reduction from the Skolem problem to the Positivity problem that increases the matrix dimension quadratically [16], [23].

III. Markov Chains

We present the basic definitions related to Markov chains and the decision problems for the optimal total reward with expected stopping time.

A. Basic definitions

A *Markov chain* is a tuple $\langle M, \mu, w\rangle$ consisting of:

- an $n \times n$ stochastic matrix M (in which all entries M_{ij} are nonnegative rationals[1], and the sum $\sum_j M_{ij}$ of the elements in each row i is 1),
- an initial distribution $\mu \in ([0,1] \cap \mathbb{Q})^n$ (viewed as $1 \times n$ row vector, and such that $\sum_i \mu_i = 1$), and
- a vector $w \in \mathbb{Q}^n$ of weights (or rewards).

We also view μ and w as functions $V \to \mathbb{Q}$ where $V = \{1, 2, \ldots, n\}$ is the set of vertices of the Markov chain. We often abbreviate Markov chains as M, when μ and w are clear from the context. We denote by $\|w\| = \max_{v\in V} |w(v)|$ the largest absolute value in w.

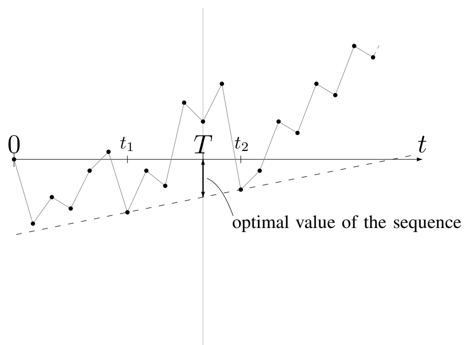
A Markov chain induces a probability measure on sequences of vertices of a fixed length, namely $\mathbb{P}(v_0 v_1 \ldots v_k) = \mu(v_0) \cdot \prod_{i=0}^{k-1} M_{v_i,v_{i+1}}$. Analogously, we denote by $\mathbb{E}(f)$ the expected value of the function $f : V^* \to \mathbb{Q}$ defined over finite sequences of vertices.

Given a stopping-time distribution $\delta : \mathbb{N} \to [0,1]$, let N_δ be a random variable whose distribution is δ. We are interested in computing the *optimal* (worst-case) expected value (or simply the value) of Markov chains with expected stopping time T, defined by:

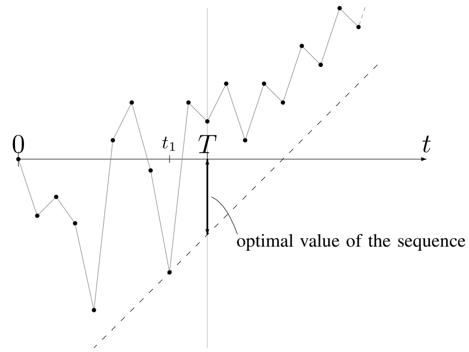
$$\begin{aligned} val(M,T) &= \inf_{\substack{\delta\in\Delta \\ \mathbb{E}_\delta = T}} \mathbb{E}\left[\sum_{i=0}^{N_\delta} w(v_i)\right] \\ &= \inf_{\substack{\delta\in\Delta \\ \mathbb{E}_\delta = T}} \mathbb{E}\left[\sum_{i=0}^{N_\delta} \mu \cdot M^i \cdot w^\intercal\right] \\ &= \inf_{\substack{\delta\in\Delta \\ \mathbb{E}_\delta = T}} \sum_{t=0}^{\infty} \delta(t) \cdot u_t, \end{aligned}$$

where $w^\intercal$ is the transpose of w, and u is the sequence of utilities defined by $u_t = \sum_{i=0}^{t} \mu \cdot M^i \cdot w^\intercal$ for all $t \geq 0$.

[1] For decidability and complexity results, we assume the numbers are rationals encoded as two binary numbers.



(a) When an optimal distribution exists



(b) When no optimal distribution exists

Fig. 1: Geometric interpretation of the value of a sequence of utilities.

With this definition in mind, we also denote the optimal expected value of a Markov chain M by $val(u, T)$. The best-case expected value, defined using $\sup$ instead of $\inf$ in the above definition, can be computed as the opposite of the worst-case expected value for the Markov chain with all weights multiplied by -1.

Exact value problem with expected stopping time. Given a Markov chain $\langle M, \mu, w\rangle$, a rational stopping time T, and a rational threshold θ, decide whether the optimal expected value of M with expected stopping time T is below θ, i.e., whether $val(M, T) < \theta$.

Approximation of the value with expected stopping time. We also consider an approximate version of the exact value problem, where the goal is to compute, given $\varepsilon > 0$, a value v_ε such that $|val(M, T) - v_\varepsilon| \leq \varepsilon$. We say that v_ε is an *approximation with additive error* ε of the optimal value.

B. Hardness of the exact value problem

This section is devoted to the proof of the following result, which establishes the inter-reducibility of the exact value problem, the Positivity problem, and the Markov Reachability problem (defined in Section III-B3).

Theorem 1. *The Positivity problem, the inequality variant of the Markov Reachability problem, and the exact value problem with expected stopping time are inter-reducible.*

The decidability status of the Positivity problem is a long-standing open question, although decidability is known for dimension $n \leq 5$ [23, Section 4]. Therefore, constructing an algorithm to compute the exact value of a Markov chain with expected stopping time T would require the significant advances in number theory that are necessary to solve the Positivity problem [23, Section 5].

We also show the converse reduction from the exact value problem to the Positivity problem. Hence proving the undecidability of the exact value problem would also be a major breakthrough, as it would entail the undecidability of the Positivity problem.

The proof of Theorem 1 is presented in the rest of this section.

1) Geometric interpretation: A geometric interpretation for (arbitrary) sequences of real numbers and expected stopping-time was developed in previous work [7]. We recall the main result in this section. The rest of our technical results is independent from [7] (see also Comparison with related work in Section I).

It is known that bi-Dirac distributions are sufficient for optimal expected value, namely for all sequences $u = u_0, u_1, \ldots$ of utilities, for all time bounds T, the following holds [7]:

$$\inf\{\mathbb{E}_\delta(u) \mid \delta \in \Delta \wedge \mathbb{E}_\delta = T\} = \inf\{\mathbb{E}_\delta(u) \mid \delta \in \Delta^{\uparrow\uparrow} \wedge \mathbb{E}_\delta = T\}.$$

Moreover the value of the expected utility of the sequence u under a bi-Dirac distribution with support $\{t_1, t_2\}$ (where $t_1 < T < t_2$) and expected time T is given by

$$u_{t_1} + \frac{T - t_1}{t_2 - t_1} \cdot (u_{t_2} - u_{t_1}). \tag{1}$$

As illustrated in Fig. 1a, this value is obtained as the intersection of the vertical axis at T and the line that connects the two points (t_1, u_{t_1}) and (t_2, u_{t_2}). Intuitively, the optimal value of a sequence of utilities is obtained by choosing the two points t_1 and t_2 such that the connecting line intersects the vertical axis at T as low as possible.

It is always possible to fix a value of t_1 such that it is sufficient to consider bi-Dirac distributions with support containing t_1 to compute the optimal value (because $t_1 \leq T$ is to be chosen among a finite set of points), but the optimal value of t_2 may not exist, as in Fig. 1b. In that case, the value of the sequence of utilities is obtained as $t_2 \to \infty$.

Given such a value of t_1, let $\nu = \inf_{t_2 \geq T} \frac{u_{t_2} - u_{t_1}}{t_2 - t_1}$, and Lemma 1 shows that $u_t \geq f_u(t)$, for all $t \geq 0$ where $f_u(t) = u_{t_1} + (t - t_1) \cdot \nu$. The optimal expected utility is

$$\begin{aligned} val(u, T) &= \min_{0 \leq t_1 \leq T} \inf_{t_2 \geq T} u_{t_1} + \frac{T - t_1}{t_2 - t_1} \cdot (u_{t_2} - u_{t_1}) \\ &= \min_{0 \leq t_1 \leq T} u_{t_1} + (T - t_1) \cdot \nu \\ &= f_u(T), \end{aligned}$$

hence $f_u(T)$ is the optimal value.

Lemma 1 (Geometric interpretation [7])**.** *For all sequences u of utilities:*

- *if $u_t \geq a \cdot t + b$ for all $t \geq 0$, then the optimal value of the sequence u is at least $a \cdot T + b$;*
- *we have $u_t \geq f_u(t)$ for all $t \geq 0$, and the optimal expected value of u is $f_u(T)$.*

It follows from Lemma 1 that the optimal value of the sequence u is the largest possible value at T of a line that lies below u: $val(u, T) = \sup\{f(T) \mid \exists a, b \cdot \forall t : f(t) = a \cdot t + b \leq u_t\}$.

2) Reduction of the Positivity problem to the exact value problem: It is known that the Positivity problem can be reduced to the inequality variant of [1, Problem A], defined below as $A^>$. A subsequent reduction of $A^>$ to the exact value problem with expected stopping time establishes one direction of Theorem 1. We present such a reduction in the proof of Lemma 2 (see also Fig. 2).

Problem $A^=$ [1]. Given a $n \times n$ aperiodic[2] stochastic matrix M with rational entries, an initial distribution $\mu = (1, 0, \dots, 0)$, and a vector $z \in \{0, 1, 2\}^n$, decide whether there exists an integer $t \geq 1$ such that $\mu \cdot M^t \cdot z^\intercal = 1$.

Problem $A^>$ [1]. Given a $n \times n$ aperiodic stochastic matrix M with rational entries, an initial distribution $\mu = (1, 0, \dots, 0)$, and a vector $z \in \{0, 1, 2\}^n$, decide whether there exists an integer $t \geq 1$ such that $\mu \cdot M^t \cdot z^\intercal > 1$.

Problems $A^=$ and $A^>$ are difficult to solve only in the case where $\mu \cdot M^t \cdot z^\intercal$ converges to 1 as $t \to \infty$. Otherwise, an argument based on the definition of convergence to a limit shows that the problems are decidable [19, Theorem 1]. Note that $\lim_{t \to \infty} \mu \cdot M^t$ exists since M is aperiodic, and the limit is the steady-state vector π, which is algorithmically computable. Hence we can assume that the instances of Problem $A^>$ are such that

$$\pi \cdot z^\intercal = 1. \tag{2}$$

Moreover, without loss of generality, we can modify M such that there is no incoming transition to the initial vertex 1 (remember that $\mu(1) = 1$) by creating a copy of the initial vertex, and redirecting the transitions to 1 towards the copy vertex. Thus we require the matrix M in Problem $A^>$ to define a Markov chain consisting of an initial vertex 1 with no incoming transition. This may however increase the dimension of the matrix by 1.

Lemma 2. *Problem $A^>$ can be reduced to the exact value problem with expected stopping time.*

Corollary 1. *The Positivity problem can be reduced to the exact value problem with expected stopping time.*

The proof of Lemma 2 is organized as follows: we first recall basic results from the theory of Markov chains, then present a reduction of Problem $A^>$ to the exact value problem with expected stopping time, and establish its correctness.

[2]Although in the original formulation of Problem A, the stochastic matrix M need not be aperiodic, the reduction of the Positivity problem to Problem A produces stochastic matrices that define aperiodic Markov chains (even ergodic unichains) [1].

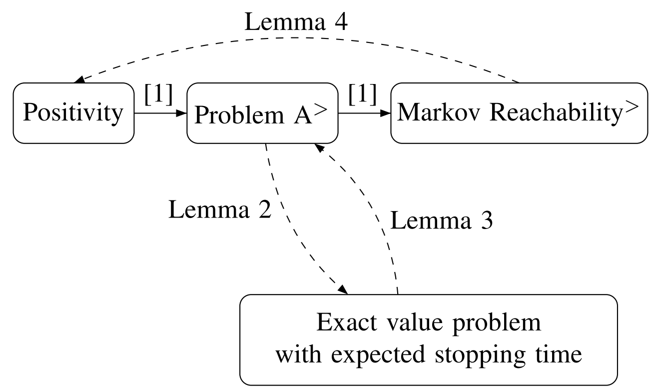


Fig. 2: Known reductions (solid lines), and reductions established in this paper (dashed lines).

Basic results. First we show that, given an aperiodic Markov chain $\langle M, \mu, w \rangle$ that has a single recurrent class, there exist vectors x, y such that the expected utility after t steps tends to $\mu \cdot x^\intercal \cdot t + \mu \cdot y^\intercal$ as $t \to \infty$, formally:

$$\lim_{t \to \infty} \left| \sum_{i=0}^{t-1} \mu \cdot M^i \cdot w^\intercal - \mu \cdot (x^\intercal \cdot t + y^\intercal) \right| = 0. \tag{3}$$

The vector x is called the *gain* per time unit, and y is the *relative-gain* vector. They can be computed by solving the following equations (following [15, Section 4.5]):

$$\begin{cases} x_i = \pi \cdot w^\intercal \text{ for all vertices } i \in V \\ y^\intercal = M \cdot (y - x)^\intercal + w^\intercal \\ \pi \cdot y^\intercal = 0 \end{cases} \tag{4}$$

The number $g = \pi \cdot w^\intercal$ is the gain per time unit. Note that $x = g \cdot e$ where $e = (1, 1, \dots, 1)$, and that $M \cdot x^\intercal = x^\intercal = g \cdot e^\intercal$ because M is stochastic and the sum of the elements in each of its row is 1. It follows that Equation (4) can be written as $y^\intercal = M \cdot y^\intercal + w^\intercal - x^\intercal$, and by $t - 1$ successive substitutions of $y^\intercal$, we get

$$\begin{aligned} y^\intercal &= M^t \cdot y^\intercal + \sum_{i=0}^{t-1} M^i \cdot w^\intercal - \sum_{i=0}^{t-1} M^i \cdot x^\intercal \\ &= M^t \cdot y^\intercal + \sum_{i=0}^{t-1} M^i \cdot w^\intercal - t \cdot x^\intercal \end{aligned}$$

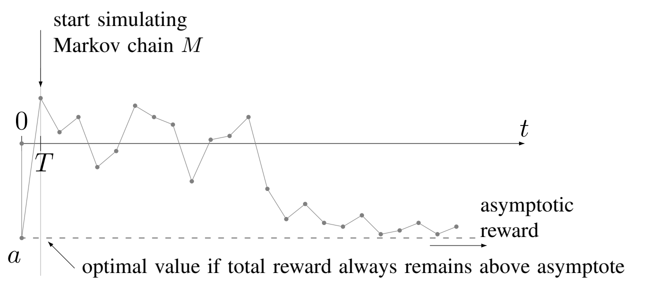


Fig. 3: Reduction of the Positivity problem to the exact value problem.

Then, the rate of convergence of the expected utility evaluates as follows, for all $t \geq 1$:

$$\begin{aligned}
&\sum_{i=0}^{t-1} \mu \cdot M^i \cdot w^\intercal - \mu \cdot (x^\intercal \cdot t + y^\intercal) \\
=&\sum_{i=0}^{t-1} \mu \cdot M^i \cdot w^\intercal - \mu \cdot x^\intercal \cdot t \\
&- \mu \cdot M^t \cdot y^\intercal - \mu \cdot \sum_{i=0}^{t-1} M^i \cdot w^\intercal + t \cdot \mu \cdot x^\intercal \\
=& - \mu \cdot M^t \cdot y^\intercal
\end{aligned} \tag{5}$$

which tends to $-\pi \cdot y^\intercal = 0$ as $t \to \infty$, establishing (3).

In the case of an aperiodic Markov chain with multiple recurrent classes, the gain and relative gain satisfying Equation (3) can be computed as the linear combination of the vectors x, y obtained for each recurrent class, where the coefficient in the linear combination is the mass of probability that reaches (in the limit) the recurrent class from the initial distribution μ.

Reduction. The reduction from Problem $A^>$ to the exact value problem with expected stopping time is as follows. Given an instance (M, μ, z) of Problem $A^>$, we construct an instance of the exact value problem in two stages. First, let $w^\intercal = z^\intercal - M \cdot z^\intercal$ be a reward vector defining a Markov chain $\langle M, \mu, w\rangle$. We explain later why w is defined in this way.

We proceed to the second stage of the construction, and define the instance of the exact value problem, namely the Markov chain $\langle M', \mu', w'\rangle$, the expected time T, and the threshold θ. The key idea of the construction is illustrated in Fig. 3. Given the Markov chain $\langle M, \mu, w\rangle$, we can compute its asymptotic expected utility, shown as the dashed line $\mu \cdot (x^\intercal \cdot t + y^\intercal)$ in Fig. 3 which also plots the sequence u_{t-1} for $t \geq 1$. Note that by Equation (3) we have $\lim_{t\to\infty} |u_{t-1} - \mu \cdot (x^\intercal \cdot t + y^\intercal)| = 0$ and by Equations (4) $x = \pi \cdot z^\intercal - \pi \cdot M \cdot z^\intercal = 0$.

We construct an instance of the exact value problem in such a way that, if the utility of M always remains above its asymptote, then the optimal value is the value of the asymptote at time T, and otherwise, the optimal value is strictly smaller. We achieve this by having an initial vertex with weight a such that, if the Markov chain $\langle M, \mu, w\rangle$ is executed (simulated) after the initial vertex, then the weight a lies exactly on the asymptote of $\langle M, \mu, w\rangle$ (see Fig. 3 and the geometric interpretation in Section III-B1). Since we simulate $\langle M, \mu, w\rangle$ after one time step, the value of a is chosen such that the point $(0, a)$ belongs to the line $\mu \cdot (x^\intercal \cdot t + y^\intercal)$. Since $\mu = (1, 0, \ldots, 0)$ in Problem $A^>$, we have $a = y(1)$. To recover the original behavior of the Markov chain $\langle M, \mu, w\rangle$, we subtract a from the weight of the initial vertex of M, thus $w'(1) = w(1) - a$. As we assumed that the initial vertex in M has no incoming transition, it is never re-visited later. We take $T = 1$ and the value of the asymptote at time T is $\mu \cdot (x^\intercal + y^\intercal) = x(1) + y(1) = a$, which we define as the threshold θ of the exact value problem, thus $\theta = a$.

Formally, the instance of the exact value problem is defined as follows:

$$w' = \begin{pmatrix} a \\ w(1) - a \\ w(2) \\ \vdots \\ w(n) \end{pmatrix}, \; M' = \begin{pmatrix} 0 & \mu \\ 0 & M \end{pmatrix}, \; \begin{array}{l} \mu' = (1, 0, \ldots, 0), \\ T = 1, \\ \theta = a \end{array}$$

where $a = y(1)$ and y is the relative-gain vector of the Markov chain $\langle M, \mu, w\rangle$. Note that the initial vertex of M has no incoming transition (in M), and thus the sequence of expected utilities in M' indeed simulates the sequence of expected utilities in M, and the asymptotic expected utilities as well as the steady-state vectors of $\langle M, \mu, w\rangle$ and $\langle M', \mu', w'\rangle$ coincide.

Correctness of the reduction. To establish the correctness of the reduction, we show the following equivalences:

1) the optimal expected value of M' with expected stopping time T is smaller than θ (i.e., the answer to the exact value problem is YES) if and only if the utility sequence of M eventually drops below its asymptote;
2) the utility sequence of M eventually drops below its asymptote if and only if $\mu \cdot M^t \cdot z^\intercal > 1$ for some $t \geq 1$ (i.e., the answer to Problem $A^>$ is YES).

To show the first equivalence, consider the first direction and assume that the value of M' is smaller than θ. Given that the line $\mu \cdot (x^\intercal \cdot t + y^\intercal)$ has value θ at $t = T$, it follows from Lemma 1 that the utility sequence of M' does not always remain above that line, and thus the utility sequence of M eventually drops below its asymptote.

Now consider the second direction of the first equivalence and assume that the utility sequence of M eventually drops below its asymptote. Then the utility sequence of M' drops below the line $\mu \cdot (x^\intercal \cdot t + y^\intercal)$, say at time $t_2 \geq 1$. We construct a distribution δ with $\mathbb{E}_\delta = T$ such that the value of the expected reward under δ is less than $\mu \cdot (x^\intercal \cdot T + y^\intercal) = \theta$ (which implies that the optimal value, obtained as the infimum over all distributions, is also below θ).

We consider two cases: (1) if $t_2 = 1$ (i.e., $t_2 = T$), consider the distribution δ such that $\delta(t_2) = 1$ (note that $\mathbb{E}_\delta = T$) and

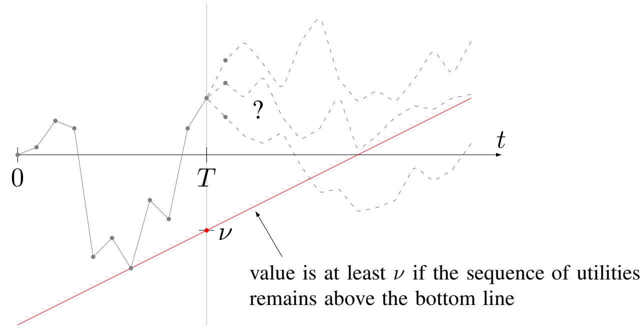


Fig. 4: Solving the exact value problem using an oracle for Problem A$^>$.

the result follows immediately; (2) otherwise, $t_2 > 1$ and consider the bi-Dirac distribution with support $\{t_1, t_2\}$ where $t_1 = 0$. Note that $t_1 < T < t_2$ and the value of the expected reward under this distribution is given by the value at time T of the line connecting the point (t_1, a) and a point below the asymptote (at t_2), see Equation (1). This value is below the value θ of the asymptote at time T since (t_1, a) is on the asymptote, and the other point (at t_2) is strictly below the asymptote.

To show the second equivalence, note that by Equation (5) the utility sequence of M eventually drops below its asymptote if and only if $-\mu \cdot M^t \cdot y^\intercal < 0$ for some $t \geq 1$. Hence we can establish the second equivalence by showing that $-\mu \cdot M^t \cdot y^\intercal < 0$ if and only if $\mu \cdot M^t \cdot z^\intercal > 1$. This is where the value of w is important. The result holds if $y = z - e$, and we just need to show that $y = z - e$ satisfies Equations (4), namely that

$$\begin{aligned} &(z-e)^\intercal = M \cdot (z-e-x)^\intercal + w^\intercal \\ &\pi \cdot (z-e)^\intercal = 0 \end{aligned}$$

that is

$$\begin{aligned} &(z-e)^\intercal = M \cdot z^\intercal - e^\intercal - x^\intercal + z^\intercal - M \cdot z^\intercal \\ &\pi \cdot z^\intercal - \pi \cdot e^\intercal = 0 \end{aligned}$$

which hold since $x = 0$ and $\pi \cdot z^\intercal = 1 = \pi \cdot e^\intercal$ (Equation (2)). This concludes the proof of Lemma 2.

Using the reduction of the Positivity problem to Problem A$^>$ [1], we obtain Corollary 1, showing that a decidability result for the exact value problem would imply the decidability of the Positivity problem, which is a longstanding open question.

3) Reduction of the exact value problem to the Positivity problem: We present the converse reduction of Section III-B2, showing that to potentially prove the exact value problem is undecidable would require such a proof for the Positivity problem as well. We sketch the reduction by showing how the exact value problem can be solved using an oracle for Problem A$^>$, illustrated in Fig. 4, and then present a reduction of Problem A$^>$ to the Positivity problem.

Lemma 3. *The exact value problem with expected stopping time can be reduced to Problem A$^>$.*

Proof sketch. Given a Markov chain $\langle M, \mu, w\rangle$ with expected stopping time T and threshold θ, we solve the exact value problem using an oracle for Problem A$^>$ as follows. First, if $u_T < \theta$ then the answer to the exact value problem is YES. Otherwise, we compute the value of utilities $u_t = \sum_{i=0}^{t} \mu \cdot M^i \cdot w^\intercal$ for all $0 \leq t < T$, and let $b = \max_{0 \leq t < T} \frac{u_t - \theta}{t - T}$. Consider the *bottom line* of equation $b \cdot (t - T) + \theta$ and observe that $u_t \geq b \cdot (t - T) + \theta$ for all $0 \leq t \leq T$ (see Fig. 4). By the geometric interpretation lemma (Lemma 1), it suffices to determine whether the sequence of utilities ever drops below the bottom line to answer the exact value problem.

This question is simple if the asymptotic behaviour of the Markov chain differs from the bottom line (e.g., if the *asymptote* given by Equations (4) does not coincide with the bottom line), as there is a time from which the sequence of utilities will remain either always above or always below the bottom line, and such a time point is computable. The difficult case is when the bottom line coincides with the asymptotic behaviour of the sequence of utilities. Then the condition for the sequence of utilities to eventually drop below the bottom line is that $-\mu \cdot M^t \cdot y^\intercal < 0$ for some $t \geq 1$ (Equation (5)) which, up to an elementary transformation can be stated as an equivalent question of the form "does $\mu \cdot M^t \cdot z^\intercal > 1$ for some $t \geq 1$?" where $z \in \{0, 1, 2\}^n$, and thus solvable using an oracle for Problem A$^>$. □

To obtain the inter-reducibility result of Theorem 1, we need to show that Problem A$^>$ can be reduced to the Positivity problem, which we establish by showing that the inequality version of the Markov reachability problem (defined below) can be reduced to the Positivity problem, as it is known that Problem A$^>$ can be reduced to the inequality variant of the Markov reachability problem [1] (see also Fig. 2). This is a straightforward result established in Lemma 4.

Markov reachability$^>$ problem [1]. Given a square stochastic matrix M with rational entries and a rational number $r > 0$, decide whether there exists an integer $t \geq 1$ such that $M^t_{1,2} > r$.

Markov reachability$^=$ problem [1]. Given a square stochastic matrix M with rational entries and a rational number $r > 0$, decide whether there exists an integer $t \geq 1$ such that $M^t_{1,2} = r$.

Lemma 4. *The Markov reachability$^>$ problem can be reduced to the Positivity problem.*

The reduction in Lemma 4 increases by 3 the dimension of the matrix (given a $n \times n$ stochastic matrix for the Markov reachability$^>$ problem, we construct a $(n+3) \times (n+3)$ matrix for the Positivity problem).

The results of Lemma 2, 3 and 4 establish Theorem 1. The reduction in Lemma 4 can easily be adapted to show that the Markov reachability$^=$ problem can be reduced to the Skolem problem and thus these problems are inter-reducible with Problem A$^=$.

Theorem 2. *The Skolem problem, Problem A$^=$, and the Markov reachability$^=$ problem are inter-reducible.*

C. Approximation of the optimal value

We can compute an approximation of the optimal value with additive error by considering an approximation u' of the exact sequence u of expected utilities of the Markov chain as follows: for a large number of time steps, let the approximate sequence u' be equal to u, and then from some point on it switches to the value of the limit (asymptotic, and possibly periodic) sequence of expected utilities at the steady-state distribution(s). By taking the switching point large enough, the approximation sequence u' can be made arbitrarily close to the exact sequence u. We show that the value of the sequences u' approximates arbitrarily closely the (exact) optimal value of u.

By the results of Section III-B1, the optimal expected value of any sequence u' of utilities is given by the expression

$$val(u', T) = \min_{0 \le t_1 \le T} \inf_{t_2 \ge T} \frac{u'_{t_1}(t_2 - T) + u'_{t_2}(T - t_1)}{t_2 - t_1}. \quad (6)$$

We can effectively compute the value of $val(u', T)$ when u' is an ultimately periodic sequence, i.e. $u' = A.C^\omega$ where A, C are finite sequences (with C nonempty): we show in Lemma 5 that the infinite range of t_2 in the expression (6) can be replaced by a finite range, because the optimal value is obtained either by taking t_2 before the first repetition of the cycle C, or by taking $t_2 \to \infty$ (i.e., if repeating the cycle once improves the value, then repeating the cycle infinitely often improves the value even more). Let S_A and S_C be the sum of the weights in A and C respectively, let $M_C = \frac{S_C}{|C|}$ be the average weight of the cycle C.

Lemma 5. *The optimal value of an ultimately periodic sequence $u = A.C^\omega$ is $val(u, T) = \min\{E_1, E_2\}$ where*

$$E_1 = \min_{0 \le t_1 \le T} \min_{T \le t_2 \le |A|+|C|} \frac{u_{t_1}(t_2 - T) + u_{t_2}(T - t_1)}{t_2 - t_1}, \text{and}$$

$$E_2 = \min_{0 \le t_1 \le T} u_{t_1} + M_C \cdot (T - t_1).$$

If $T \ge |A| + |C|$, then $val(u, T) = \min_{0 \le t_1 \le |A|+|C|} u_{t_1} + M_C \cdot (T - t_1)$.

We show that for a sequence u' of utilities that approximates the sequence u, the value of u' approximates the value of u and the error can be bounded. Precisely, if the weights in a Markov chain are shifted by at most η, then the optimal expected value of the Markov chain with expected stopping time T is shifted by at most $\eta \cdot (T+1)$. Consider w' such that $|w'(v) - w(v)| \le \eta$ for all vertices $v \in V$, and consider the sequences u and u' of utilities of a path according to w and w' respectively. Then we have $|u'_t - u_t| \le (t+1) \cdot \eta$ for all $t \ge 0$, and for all distributions δ with $\mathbb{E}_\delta = T$:

$$\begin{aligned} \Big| \sum_i \delta(i) \cdot u'_i - \sum_i \delta(i) \cdot u_i \Big| &\le \sum_i \delta(i) \cdot |u'_i - u_i| \\ &\le \sum_i \delta(i) \cdot (i+1) \cdot \eta \\ &= (T+1) \cdot \eta. \end{aligned}$$

It follows that $|val(u', T) - val(u, T)| \le (T+1) \cdot \eta$, that is the value of the sequence is shifted by at most $(T+1) \cdot \eta$ (it is easy to see that if $\forall \delta : |f(\delta) - g(\delta)| \le K$, then $|\inf_\delta f - \inf_\delta g| \le K$).

Lemma 6. *Given $\eta \ge 0$ and two sequences u and u' of utilities such that $|u'_t - u_t| \le (t+1) \cdot \eta$ for all $t \ge 0$, we have $|val(u', T) - val(u, T)| \le (T+1) \cdot \eta$. Analogously, if $u'_t = u_t + (t+1) \cdot \eta$ for all $t \ge 0$, then $val(u', T) = val(u, T) + (T+1) \cdot \eta$.*

We recall a result about Markov chains, which states that for Markov chains with only aperiodic recurrent classes, the vector $\mu \cdot M^t$ converges to a steady-state vector π, and the rate of convergence is bounded by an exponential in n [15, Theorem 4.3.7]. For all $j \in V$:

$$|(\mu \cdot M^t)_j - \pi_j| \le K_1 \cdot K_2^t$$

where K_1, K_2 are constants with $K_2 < 1$, namely $K_2 = (1 - \alpha^{n^2})^{1/3n^2}$ where α is the smallest non-zero probability in M (i.e., $\alpha = \min\{M_{ij} \mid M_{ij} > 0\}$) and n is the number of vertices of M.

For general Markov chains (with possibly periodic recurrent classes), we adapt the above result as follows. Consider the set $\mathcal{T}$ of transient vertices, each recurrent class $C_1, C_2, \dots, C_l$ with their respective period $d_1, d_2, \dots, d_l$, and let $d = \mathrm{lcm}\{d_1, \dots, d_l\}$ be their least common multiple. Note that $d_i \le n$ for all $1 \le i \le l$ and d is at most the product of all prime numbers smaller than n, thus at most exponential in n [13]. Then M^d can be viewed as the transition matrix of a Markov chain with aperiodic recurrent classes, and thus $\mu \cdot M^{d \cdot t}$ converges to a steady-state vector π as $t \to \infty$. Considering a recurrent class C_i, and the vertices $j \in C_i \cup \mathcal{T}$ the rate of convergence can be bounded as follows, where α^{d_i} is a lower bound on the smallest non-zero probability in M^{d_i}:

$$\begin{aligned} |(\mu \cdot M^{d \cdot t})_j - \pi_j| &= |(\mu \cdot (M^{d_i})^{\frac{d \cdot t}{d_i}})_j - \pi_j| \\ &\le K_1 \cdot (1 - \alpha^{d_i \cdot n^2})^{\frac{d \cdot t}{d_i \cdot 3n^2}} \\ &\le K_1 \cdot (1 - \alpha^{n^3})^{\frac{t}{3n^2}}, \end{aligned}$$

which is independent of i, and thus holds for all $j \in V$. Let $K_3 = (1 - \alpha^{n^3})^{1/3n^2}$.

It follows that $|\mu \cdot M^{d \cdot t} \cdot w^\intercal - \pi \cdot w^\intercal| \le n \cdot W \cdot K_1 \cdot K_3^t$ where $W = \|w\|$ is the largest absolute weight in w. Then for all $\varepsilon > 0$, for all $t \ge \frac{\ln(\frac{\varepsilon}{n \cdot W \cdot K_1})}{\ln(K_3)} =: B$, we have $|\mu \cdot M^{d \cdot t} \cdot w^\intercal - \pi \cdot w^\intercal| \le \varepsilon$, and by the same reasoning with initial distributions $\mu \cdot M, \mu \cdot M^2, \dots, \mu \cdot M^{d-1}$ we get $|\mu \cdot M^{d \cdot t + k} \cdot w^\intercal - \pi \cdot M^k \cdot w^\intercal| \le \varepsilon$ for all $0 \le k < d$.

Consider the sequence u' defined by

$$u'_t = \begin{cases} u_t & \text{for all } t \le d \cdot B \\ u_{d\cdot B} + \sum_{k=d\cdot B+2}^{t} \pi \cdot M^{k\%d} \cdot w^{\intercal} & \text{for all } t > d \cdot B \end{cases}$$

where $k\%d$ is the remainder of the division of k by d. Intuitively, u'_t approximates u_t after time $t = d \cdot B$ by considering the (expected) weight at time t to be given by the limit (expected) weight at the steady-state distribution.

Then $|u'_t - u_t| \le (t+1) \cdot \varepsilon$ for all $t \ge 0$, and therefore $|val(u', T) - val(u, T)| \le \varepsilon \cdot (T+1)$ (by Lemma 6). The sequence u' is an ultimately periodic sequence of the form $A.C^{\omega}$ where $|A| = d \cdot B$ and $|C| = d$. Hence the optimal value of u' is given by Lemma 5 and can be obtained by computing the first $d \cdot B + d$ terms of the sequence u', the steady-state vector π, the number d, and the average weight $M_C = \frac{S_C}{|C|}$ where $S_C = \sum_{i=0}^{d-1} \pi \cdot M^i \cdot w^{\intercal}$. This provides a way to compute an approximation with additive error ε of the optimal value of a Markov chain in time $O(P(n) \cdot T \cdot B \cdot d)$ where $P(n)$ is a polynomial in the size of the Markov chain (that accounts for matrix multiplication, steady-state vector computation, etc.). Using the fact that $(1 - \frac{1}{x})^x \in O(1)$, and that $\ln\left(1 - \frac{1}{x}\right) \in O(-1/x)$, we obtain the bounds in Theorem 3 in the special cases where α or n is constant.

Theorem 3. *The optimal expected value of a Markov chain with expected stopping time T can be computed to an arbitrary level of precision $\varepsilon > 0$, in time*

$$O\left(P(n) \cdot T \cdot \frac{\ln\left(\frac{\varepsilon}{n \cdot W}\right)}{\ln(K_3)} \cdot 2^{O(n)}\right)$$

where $K_3 = (1 - \alpha^{n^3})^{1/3n^2}$ and $P(\cdot)$ is a polynomial.

If α (the smallest non-zero probability) is constant, then the computation time is in

$$O\left(\frac{P(n) \cdot 2^{O(n)}}{\alpha^{n^3}} \cdot T \cdot \ln\left(\frac{n \cdot W}{\varepsilon}\right)\right) \ (\text{as } n \to \infty).$$

If n (the number of vertices) is constant, then the computation time is in

$$O\left(\frac{1}{\alpha^{O(1)}} \cdot T \cdot \ln\left(\frac{W}{\varepsilon}\right)\right) \ (\text{as } \alpha \to 0).$$

There is a family of Markov chains on which the approximation algorithm of Theorem 3 runs in time exponential in the number of vertices of the Markov chains, showing that the complexity analysis of our algorithm cannot be improved to eliminate the exponential dependency in the number of vertices. However, whether there exists a polynomial-time algorithm for the approximation problem is an open question.

Proposition 1. *There exists a family of aperiodic Markov chains $M(n, \alpha)$ with $2n$ vertices ($n \in \mathbb{N}$) and smallest probability α ($\alpha \le \frac{1}{2}$) such that, for the initial distribution $\mu = (1, 0, \dots, 0)$, we have*

$$\max_j |(\mu \cdot M(n, \alpha)^t)_j - \pi_j| \ge (1 - \alpha^n)^{\frac{t}{n}},$$

where π is the steady-state vector of $M(n, \alpha)$, and the computation time of the approximation algorithm (of Theorem 3) for $M(n, \alpha)$ is at least

$$\frac{n \cdot \ln(1/\varepsilon)}{\alpha^n}.$$

IV. Markov Decision Processes

Markov decision processes (MDPs) extend Markov chains with transition choices determined by control actions. We give the basic definitions of MDPs and of the optimal expected value of an MDP with expected stopping time T.

A. Basic definitions

A *Markov decision process* is a tuple $\mathcal{M} = \langle V, A, \theta, \mu, w \rangle$ consisting of:

- a finite set V of vertices and a finite set A of actions,
- a transition function $\theta : V \times A \to (V \to [0, 1])$ such that $\theta(v, a)$ is a probability distribution over V, that is $\sum_{v' \in V} \theta(v, a)(v') = 1$ for all $v \in V$ and $a \in A$.
- $\mu : V \to [0, 1]$ is an initial distribution and $w : V \to \mathbb{Q}$ is a vector of weights, as in Markov chains.

Given a vertex $v \in V$ and a set $U \subseteq V$, let $A_U(v)$ be the set of all actions $a \in A$ such that $\mathsf{Supp}(\theta(v, a)) \subseteq U$. A *closed* set in an MDP is a set $U \subseteq V$ such that $A_U(v) \neq \varnothing$ for all $v \in U$. A set $U \subseteq V$ is an *end-component* [10], [3] if (i) U is closed, and (ii) the graph (U, E_U) is strongly connected where $E_U = \{(v, v') \in U \times U \mid \theta(v, a)(v') > 0 \text{ for some } a \in A_U(v)\}$ denote the set of edges given the actions. In the sequel, end-components should be considered *maximal*, that is such that no strict superset is an end-component.

A *strategy* in $\mathcal{M}$ is a function $\sigma : V^+ \to (A \to [0, 1])$ such that $\sigma(\rho)$ is a probability distribution over A, for all sequences $\rho \in V^+$. A strategy σ is *pure* if for all $\rho \in V^+$, there exists an action $a \in A$ such that $\sigma(\rho)(a) = 1$; σ is *memoryless* if $\sigma(\rho v) = \sigma(\rho' v)$ for all $\rho, \rho' \in V^*$ and $v \in V$; σ uses *finite memory* if there exists a right congruence $\approx$ over V^+ (i.e., if $\rho \approx \rho'$, then $\rho \cdot v \approx \rho' \cdot v$ for all $\rho, \rho' \in V^+$ and $v \in V$) of finite index such that $\rho \approx \rho'$ implies $\sigma(\rho) = \sigma(\rho')$.

Given the initial distribution μ, and a strategy σ, a probability can be assigned to every finite path $\rho = v_0 \cdots v_n$ as follows:

$$\mathbb{P}^{\sigma}_{\mu}(v_0 v_1 \dots v_k) = \mu(v_0) \cdot \prod_{i=0}^{k-1} \sum_{a \in A} \sigma(v_0 \cdots v_i)(a) \cdot \theta(v_i, a)(v_{i+1}).$$

Analogously, we denote by $\mathbb{E}^{\sigma}_{\mu}(f)$ the expected value of the function $f : V^* \to \mathbb{Q}$ defined over finite sequences of vertices. Let $u_t = \mathbb{E}^{\sigma}_{\mu}(\sum_{i=0}^{t} w(v_i))$ and define the optimal expected value of $\mathcal{M}$ with expected stopping time $T \in \mathbb{Q}$ as follows:

$$val(\mathcal{M}, T) = \sup_{\sigma} \inf_{\substack{\delta \in \Delta \\ \mathbb{E}_{\delta} = T}} \sum_{t=0}^{\infty} \delta(t) \cdot u_t.$$

The strategy σ is *ε-optimal* if the sequence $u = (u_t)_{t \in \mathbb{N}}$ it induces is such that $val(u, T) \ge val(\mathcal{M}, T) - \varepsilon$. For $\varepsilon = 0$, we simply say that σ is *optimal* (instead of 0-optimal).

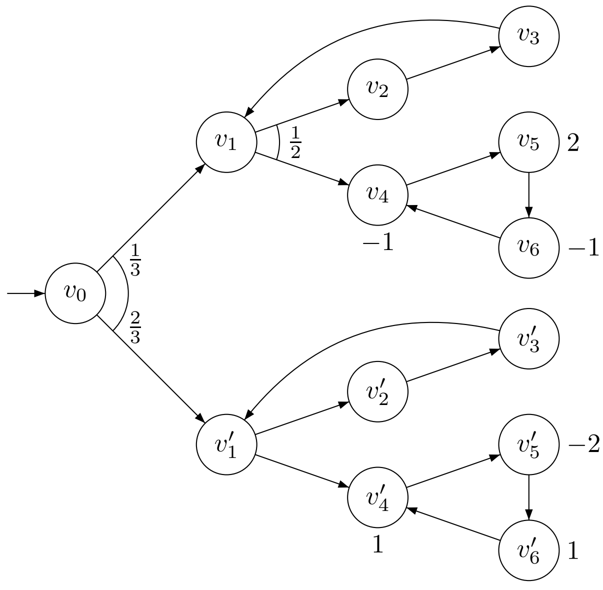


Fig. 5: An MDP where infinite memory is required for optimal expected value.

For an arbitrary strategy σ, with probability 1 the set of states visited infinitely often along an (infinite) path is an end-component [9], [10]. Let the limit-probability of a (maximal) end-component U be the probability that the set of states visited infinitely often along a path is a subset of U. A *limit distribution* under σ is a distribution δ^* such that, for every end-component U, the limit-probability of U is $\sum_{v \in U} \delta^*(v)$.

B. Infinite memory is necessary

Since MDPs are an extension of Markov chains, the problem of computing the optimal expected value $val(\mathcal{M}, T)$ is Positivity-hard (by Corollary 1). Another source of hardness for this problem is that infinite memory is required for optimal strategies, as illustrated in the following example.

Example. We show in Fig. 5 an MDP where infinite memory is required for optimal expected value. The only strategic choice is in vertex v_1' (we omit the actions in the figure, and all weights not shown are 0). In particular, the upper part $\{v_1, \dots, v_6\}$ is a Markov chain and after $3k+2$ steps, the probability mass in v_4 is $p_k = \frac{1}{3} \cdot (1 - \frac{1}{2^{k+1}})$. For instance $p_0 = \frac{1}{6}$. Note that one step before, the probability mass in v_1 is $\frac{1}{3} \cdot \frac{1}{2^k}$.

We claim that the optimal expected value of the MDP is 0, which can be obtained by a strategy $\sigma_{\rm opt}$ that ensures utility 0 at every step: let m_k be the mass of probability in v_1' after $3k+1$ steps (thus $m_0 = \frac{2}{3}$, and $m_1, m_2, \dots$ depend on the strategy). In v_1', after $3k+1$ steps, the strategy $\sigma_{\rm opt}$ chooses v_4' with probability α_k such that $m_0 \cdot \alpha_0 = p_0$, thus $\alpha_0 = \frac{1}{4}$, and $m_k \cdot \alpha_k = p_k - p_{k-1}$ for all $k \geq 1$. It is easy to see that $m_k = \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2^k}$ and $\alpha_k = \frac{1}{2+2^{k+1}}$ ensure this as well as $m_{k+1} = m_k \cdot (1 - \alpha_k)$ for all $k \geq 0$. Therefore the strategy $\sigma_{\rm opt}$ maintains always the same probability in v_4' as in v_4, and the expected total reward is 0 at every step.

It is easy to show that any other strategy (with a different value of some α_k) produces a negative total utility at some time step (either by putting too much probability into v_4', and thus too much probability for weight -2 in v_5', as compared to the weight 2 in v_5, or by putting too little probability into v_4', and thus too little probability for weight 1 in v_4', as compared to the weight -1 in v_4), and that it entails a negative expected value of the MDP.

The strategy $\sigma_{\rm opt}$ requires infinite memory, since the sequence α_k is strictly decreasing, and the vertex v_1' is reached after $3k+1$ steps along a unique path $\rho_k = v_0 v_1' (v_2' v_3' v_1')^k$. It follows that for all right congruences $\approx$ over V^+ such that $\rho \approx \rho'$ implies $\sigma_{\rm opt}(\rho) = \sigma_{\rm opt}(\rho')$, we have $\rho_k \not\approx \rho_l$ for $k \neq l$ since $\alpha_k \neq \alpha_l$ for $k \neq l$, thus $\approx$ cannot have finite index.

As the above example illustrates, infinite-memory strategies are required in MDPs. The expected stopping-time problem can be formulated as a game between a player that controls the transition choice and the opponent that chooses the stopping times. However, the game is not a perfect-information game as the opponent chooses the stopping times without knowing the execution of the MDP (in particular, the stopping-time distribution cannot be adapted according to the outcome of the probabilistic choices in the MDP). As a consequence, while finite-memory strategies are sufficient in finite-horizon planning (even in perfect-information stochastic games), in contrast we show infinite-memory strategies are required. In general, in imperfect-information probabilistic models such as probabilistic automata [25], [27], [28], infinite-memory strategies are required [2], and the basic computational problems (such as optimal reachability probability) as well as their *approximation* are undecidable [20]. However, our setting only represents limited imperfect information for the opponent, and we establish in the rest of this section that the approximation problem is decidable.

C. Approximation of the optimal value

The problem of computing $val(\mathcal{M}, T)$ up to an additive error ε can be solved as follows. We show that there exist ε-optimal strategies of a simple form: after some time t^* (that depends on ε), it is sufficient to play a (memoryless) strategy that maximizes the mean-payoff expected reward, defined as follows for a strategy σ in $\mathcal{M}$:

$$\mathsf{MP}(\mathcal{M}, \sigma) = \limsup_{t \to \infty} \frac{1}{t} \sum_{i=0}^{t-1} \mathbb{E}_\mu^\sigma(w(v_i)),$$

and the *optimal mean-payoff value* is

$$val^{\mathsf{MP}}(\mathcal{M}) = \sup_\sigma \mathsf{MP}(\mathcal{M}, \sigma).$$

Remark 2. *It is known that (see e.g. [26]):*

- *pure memoryless strategies are sufficient for mean-payoff optimality, that is there exists a pure memoryless strategy σ such that $val^{\mathsf{MP}}(\mathcal{M}) = \mathsf{MP}(\mathcal{M}, \sigma)$;*

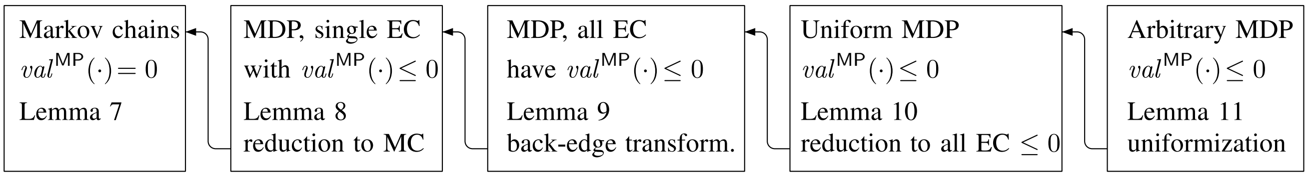


Fig. 6: Main steps towards the proof that the supremum of total expected reward is bounded in MDPs with mean-payoff value at most 0 (Theorem 4).

- *for variants of the definition of mean-payoff expected reward (using* $\liminf$ *instead of* $\limsup$*), or where the* $\limsup$ *and* $\mathbb{E}(\cdot)$ *operators are swapped (also known as the expected mean-payoff value), the same pure memoryless strategy is optimal;*
- *all vertices in an end-component have the same optimal mean-payoff value.*

Intuitively, a strategy σ that plays according to an optimal mean-payoff strategy after some time t^* has an asymptotic behaviour that is at least as good as any strategy, in particular any ε-optimal strategy; up to time t^* (thus for finitely many steps), if the strategy σ plays like an ε-optimal strategy, then the sequence of expected reward (defined above as u_t) is also good enough; the only question is whether switching to an optimal mean-payoff strategy may induce a transient loss of reward after t^* that could impede ε-optimality. In fact, we show that (1) the loss is bounded, and (2) the impact of a bounded loss on the expected value is negligible if t^* is large enough. That the loss is bounded, namely:

$$\sup_\sigma \sup_t \sum_{i=0}^{t} \mathbb{E}^\sigma_\mu(w(v_i)) \text{ is bounded if } val^{\mathsf{MP}}(\mathcal{M}) \leq 0,$$

may appear intuitively true, but is not simple to prove even in the special case where the mean-payoff value is 0. The proof has several steps, summarized in Fig. 6, leading to Theorem 4. We start by proving that the loss is bounded in the simple case of Markov chains with mean-payoff value 0, then for larger classes of MDPs, using reductions that transform an MDP $\mathcal{M}$ of a larger class into an MDP $\mathcal{M}'$ of a smaller class for which a bound on the loss is already established. The transformations may increase the total expected reward (as then, an upper bound for $\mathcal{M}'$ gives an upper bound for $\mathcal{M}$).

Lemma 7. *In aperiodic Markov chains* $\langle M, \mu, w\rangle$*, if the mean-payoff value, defined as* $\limsup_{t\to\infty} \frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}(w(v_i))$*, is 0, then*

$$\sup_t \left| \sum_{i=0}^{t} \mathbb{E}(w(v_i)) \right| \leq 4nW \cdot t_0(\alpha)$$

where α *is the smallest positive transition probability,* $t_0(\alpha) = 3 \cdot n^5 \cdot \left(\frac{1}{\alpha}\right)^{n^2}$*, and* W *is the largest absolute weight according to* w*.*

To prove a similar result for MDPs (Theorem 4), we first consider the case of MDPs that consist of a single end-component, and show by contradiction that if it has mean-payoff value 0 and a large expected total reward could be accumulated from a vertex v_0 using some strategy σ_0, then by reaching v_0 again (which is possible since the MDP is strongly connected) and repeating the same strategy σ_0, we could get a strictly positive mean-payoff value. A technical difficulty in this proof is that v_0 may be reached by paths of different lengths, but the large expected total reward that can be accumulated from v_0 is obtained in a fixed number of steps.

Lemma 8. *In an MDP* $\mathcal{M}$ *that is an end-component (i.e.,* V *is an end-component), if* $val^{\mathsf{MP}}(\mathcal{M}) \leq 0$ *and* $|V| = n$*, then*

$$\sup_\sigma \sup_t \sum_{i=0}^{t} \mathbb{E}^\sigma_\mu(w(v_i)) \leq 12 \cdot n^6 \cdot W \cdot \left(\frac{1}{\alpha}\right)^{n^3}$$

where α *is the smallest positive transition probability in* $\mathcal{M}$*, and* W *its largest absolute weight.*

We can easily extend the result to MDPs with several end-components, if all of them have mean-payoff value at most 0.

Lemma 9. *In an MDP* $\mathcal{M}$ *with* n *vertices in which all end-components have an optimal mean-payoff value at most* 0*, we have*

$$\sup_\sigma \sup_t \sum_{i=0}^{t} \mathbb{E}^\sigma_\mu(w(v_i)) \leq 12 \cdot n^8 \cdot W \cdot \left(\frac{1}{\alpha}\right)^{n^3+n}$$

where α *is the smallest positive transition probability in* $\mathcal{M}$*, and* W *its largest absolute weight.*

In an arbitrary MDP with mean-payoff value at most 0, some end-components may have positive value, and others negative value, as in the example of Fig. 7: the three end-components $\{v_0\}$, $\{v_1, v_2\}$, $\{v_3\}$ have respective mean-payoff value -1, 1, and -2. From the initial distribution μ where $\mu(v_0) = \mu(v_1) = \frac{1}{2}$, the mean-payoff value is 0. The case where the MDP has some end-components with positive mean-payoff value requires a slightly more technical proof (see also Fig. 6): we first show in Lemma 10 that the supremum of expected total reward in MDPs is bounded if all end-components are *uniform* (an end-component is uniform if all its vertices have the same weight); then we present uniformization in Lemma 11 to transform arbitrary MDPs into uniform MDPs.

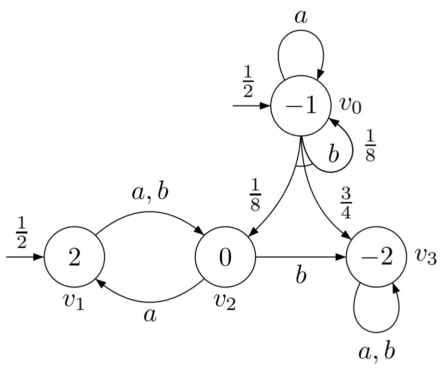


Fig. 7: An MDP with positive and negative end-components. Its mean-payoff value is 0.

Given an MDP with weight vector w, let $\mathcal{E}$ be the union of all its end-components. Define the vector w_{trans} and w_{ec} as follows:

$$w_{\mathsf{trans}}(v) = \begin{cases} w(v) & \text{if } v \in V \setminus \mathcal{E} \\ 0 & \text{if } v \in \mathcal{E} \end{cases}$$
$$w_{\mathsf{ec}}(v) = \begin{cases} 0 & \text{if } v \in V \setminus \mathcal{E} \\ w(v) & \text{if } v \in \mathcal{E} \end{cases}$$

It follows that $w = w_{\mathsf{trans}} + w_{\mathsf{ec}}$ and by the triangular inequality, we have

$$\sup_k \sum_{i=0}^{k} \mathbb{E}^\sigma_\mu(w(v_i)) \leq \sup_k \sum_{i=0}^{k} \mathbb{E}^\sigma_\mu(w_{\mathsf{trans}}(v_i)) + \sup_k \sum_{i=0}^{k} \mathbb{E}^\sigma_\mu(w_{\mathsf{ec}}(v_i)).$$

Using Lemma 9, it is easy to bound the supremum of expected total reward for w_{trans}, and we present a bound on the supremum of expected total reward for w_{ec} in uniform MDPs as follows.

Lemma 10. *Given an MDP $\mathcal{M}$ with n vertices, let w_{trans} and w_{ec} be the weight vectors of the transient vertices and of the end-components, respectively. We have*

$$\sup_\sigma \sup_t \sum_{i=0}^{t} \mathbb{E}^\sigma_\mu(w_{\mathsf{trans}}(v_i)) \leq 12 \cdot n^8 \cdot W \cdot \left(\frac{1}{\alpha}\right)^{n^3+n},$$

and if $val^{\mathsf{MP}}(\mathcal{M}) \leq 0$ and all end-components of M are uniform, then

$$\sup_\sigma \sup_t \sum_{i=0}^{t} \mathbb{E}^\sigma_\mu(w_{\mathsf{ec}}(v_i)) \leq 12 \cdot n^8 \cdot W \cdot \left(\frac{1}{\alpha}\right)^{n^3+n},$$

where α is the smallest positive transition probability in $\mathcal{M}$, and W its largest absolute weight.

We present a uniformization procedure that, given an MDP $\mathcal{M}$ with mean-payoff value at most 0, constructs an MDP $\mathcal{M}'$ with the same mean-payoff value as $\mathcal{M}$, with a larger supremum of expected total reward, and in which all end-components are *uniform* (an end-component is uniform if all its vertices have the same weight).

Lemma 11. *Given an MDP $\mathcal{M}$ with n vertices, we can construct an MDP $\mathcal{M}'$ with the following properties:*

1) *all end-components of $\mathcal{M}'$ are uniform.*
2) *$\mathcal{M}$ and $\mathcal{M}'$ have the same mean-payoff value;*
3) *the number of vertices in $\mathcal{M}'$ is at most $3n$;*
4) *the supremum of expected total reward in $\mathcal{M}$ is less than half the supremum of expected total reward in $\mathcal{M}'$;*

We finally obtain an analogue of Lemma 7 for MDPs: the expected total reward is bounded in MDPs with non-positive mean-payoff value.

Theorem 4. *Given an MDP $\mathcal{M}$ with n vertices and $val^{\mathsf{MP}}(\mathcal{M}) \leq 0$, we have:*

$$\sup_\sigma \sup_t \sum_{i=0}^{t} \mathbb{E}^\sigma_\mu(w(v_i)) \in O\left(n^{16} \cdot W \cdot \left(\frac{1}{\alpha}\right)^{O(n^3)}\right)$$

where α is the smallest positive transition probability in $\mathcal{M}$, and W its largest absolute weight.

Using Theorem 4, for all $\varepsilon > 0$ we can compute a bound t^* such that there exists an ε-optimal strategy (for expected value) that plays according to an optimal mean-payoff strategy after time t^*.

Lemma 12. *Given an MDP $\mathcal{M}$ and $\varepsilon > 0$, there exists an ε-optimal strategy that plays, after time $t^* = \frac{T\cdot(2B^*+\varepsilon)}{\varepsilon}$ (where B^* is the bound given by Theorem 4), according to a memoryless optimal strategy σ_{MP} for the mean-payoff value.*

Proof. Consider an arbitrary strategy σ in $\mathcal{M}$ (under expected stopping time T), and given $t^* \geq T$, consider a strategy σ^* that plays like σ up to time t^*, and then switches to a memoryless mean-payoff optimal strategy σ_{MP}, in the MDP $\mathcal{M}$ with initial distribution $\mu^* = \delta^\sigma_{t^*}$ (the vertex distribution of $\mathcal{M}$ after t^* steps under strategy σ). Let η^* be the optimal mean-payoff value from μ^* in $\mathcal{M}$, and let $w' = w - \eta^*$ (where $w'(v) = w(v) - \eta^*$ for all $v \in V$). With weight vector w', the optimal mean-payoff value of $\mathcal{M}$ is 0 from μ^*.

Using Lemma 7 in the Markov chain obtained by fixing the strategy σ_{MP} in $\mathcal{M}$ with initial distribution μ^*, we obtain:

$$\sup_t \left| \sum_{i=0}^{t} \mathbb{E}^{\sigma_{\mathsf{MP}}}_{\mu^*}(w'(v_i)) \right| \leq \underbrace{12 \cdot n^6 \cdot W \cdot \left(\frac{1}{\alpha}\right)^{n^2}}_{C^*}. \qquad (7)$$

Let $u_t = \sum_{i=0}^{t} \mathbb{E}^\sigma_\mu(w(v_i))$ and let $u^*_t = \sum_{i=0}^{t} \mathbb{E}^{\sigma^*}_\mu(w(v_i))$ be the sequence of expected total reward under strategy σ and σ^* respectively. To show ε-optimality of σ^*, take $t^* \geq \frac{T\cdot(2B^*+\varepsilon)}{\varepsilon}$ and show that:

$$val(u^*, T) \geq val(u, T) - \varepsilon$$

The proof is in two steps. First we bound the difference $u_t - u_t^*$ as follows, for all $t \geq 1$:

$$\begin{aligned}
u_t - u_t^* &= \sum_{i=0}^{t} \mathbb{E}_\mu^\sigma(w(v_i)) - \sum_{i=0}^{t} \mathbb{E}_\mu^{\sigma^*}(w(v_i)) \\
&= \sum_{i=0}^{t} \mathbb{E}_\mu^\sigma(w'(v_i)) - \sum_{i=0}^{t} \mathbb{E}_\mu^{\sigma^*}(w'(v_i)) \\
&\qquad \text{(since } \mathbb{E}(w(\cdot)) = \mathbb{E}(w'(\cdot)) + \eta^*\text{)} \\
&= \sum_{i=t^*}^{t} \mathbb{E}_\mu^\sigma(w'(v_i)) - \sum_{i=t^*}^{t} \mathbb{E}_\mu^{\sigma^*}(w'(v_i)) \\
&\qquad (\sigma \text{ and } \sigma^* \text{ agree in the first } t^* \text{ steps}) \\
&\leq B^* + C^* \leq 2B^* \\
&\qquad \text{(triangular inequality and bounds given by Theorem 4 and (7))}
\end{aligned}$$

In a second step, consider an arbitrary bi-Dirac distribution δ with support $\{t_1, t_2\}$ and expected stopping-time T, and consider the difference between the value of sequences u_t and u_t^* under δ, if $t_2 \geq t^*$ (the difference is 0 if $t_2 < t^*$):

$$\begin{aligned}
&\mathbb{E}_\delta(u) - \mathbb{E}_\delta(u^*) \\
=&\frac{u_{t_1}(t_2 - T) + u_{t_2}(T - t_1)}{t_2 - t_1} - \frac{u^*_{t_1}(t_2 - T) + u^*_{t_2}(T - t_1)}{t_2 - t_1} \\
=&\frac{T - t_1}{t_2 - t_1} \cdot (u_{t_2} - u^*_{t_2}) \\
&\qquad (\text{since } \sigma \text{ and } \sigma^* \text{ agree in the first } t^* \text{ steps, and thus } u_{t_1} = u^*_{t_1}) \\
\leq&\frac{T - t_1}{t_2 - t_1} \cdot 2B^* \leq \frac{T}{t^* - T} \cdot 2B^* \leq \varepsilon \\
&\qquad (\text{since } 0 \leq t_1 \leq T)
\end{aligned}$$

It follows that under all bi-Dirac distributions δ with expected stopping-time T, the expected value of the sequence u_t^* is, up to additive error ε, greater than the expected value of u_t. Therefore, since bi-Dirac distributions are sufficient for optimality (Section III-B1), we have $val(u^*, T) \geq val(u, T) - \varepsilon$. Hence σ^* is ε-optimal. □

We can express in the existential theory of the reals that the value of a strategy that eventually plays according to a memoryless strategy (as in Lemma 12) is above a given threshold, which entails decidability of computing an approximation of the optimal value up to an additive error ε.

Lemma 13. *Given an MDP $\mathcal{M}$ and a time t^*, we can compute to an arbitrary level of precision $\varepsilon > 0$ the optimal value among the strategies that play after time t^* according to a memoryless strategy.*

Proof. We describe the choices of an arbitrary strategy up to time t^* using variables $x_{v,t,a}$ for every $v \in V$, $0 \leq t \leq t^*$, and $a \in A$, where $x_{v,t,a}$ is the probability to play action a at time t in vertex v. Note that we ignore the history of vertices, which is no loss of generality since the utility achieved by a strategy at time t only depends on the probability mass in each vertex at time t, and if a sequence of distribution can be achieved by some strategy, then it can be achieved by a Markov strategy (in which the choice depends only on the time and the current vertex). It is easy to express the probability mass in v at time t (and therefore the utility u_t) as a function of the variables $x_{v,t,a}$.

After time t^*, consider a memoryless strategy and we can express its mean-payoff value η^* as a function of the vertex distribution at time t^*, thus as a function of the variables $x_{v,t,a}$. Then for $t = t^* + 1, t^* + 2, \dots, \hat{t}$, we compute the utility u_t at time t as a function of the variables $x_{v,t,a}$, and consider the utility sequence $u_0, \dots, u_{\hat{t}}, u_{\hat{t}} + \eta^*, u_{\hat{t}} + 2\eta^*, \dots$ (corresponding to an ultimately periodic path) using Lemma 5 and by an argument similar to the proof of Lemma 6 using the bound of Lemma 7 for Markov chains, we get a bound on the approximation error as follows: the value after $\hat{t}$ differ by at most $D = n \cdot W \cdot K_1 \cdot K_3^{\hat{t} - t^*}$ from the actual utility, thus the error on the value is at most

$$\frac{D \cdot (T - t_1)}{t_2 - t_1} \leq D \cdot T$$

which is at most ε for $\hat{t} \geq t^* + B$ where $B = \frac{\ln(\frac{\varepsilon}{n \cdot W \cdot T \cdot K_1})}{\ln(K_3)}$ (Lemma 6) □

By Lemma 12 and Lemma 13, we can compute up to error $\frac{\varepsilon}{2}$ the value of an $\frac{\varepsilon}{2}$-optimal strategy, and since the error is additive ($\varepsilon = \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$), it follows from the proof of Lemma 13 that, by computing (as a symbolic expression in variables $x_{v,t,a}$) the sequence of utilities up to time $\hat{t} = \frac{T \cdot (4B^* + \varepsilon)}{\varepsilon} + \frac{\ln(\frac{\varepsilon}{2n \cdot W \cdot T \cdot K_1})}{\ln(K_3)}$ and then considering an increment of η^* at every step, we can compute the value of optimal expected value of the MDP up to error ε in exponential space (since $\hat{t}$ is exponential and the existential theory of the reals can be decided in PSPACE [6]). In this way, we obtain the main result of this section: an approximation of the value with expected stopping time can be computed for MDPs up to an arbitrary additive error.

Theorem 5. *The optimal expected value of an MDP with expected stopping time T can be computed to an arbitrary level of precision $\varepsilon > 0$, in exponential space.*

V. Conclusion

We studied Markov chains and MDPs with expected stopping time, and showed the hardness of computing the exact value, as the associated decision problem for Markov chains is inter-reducible with the Positivity problem, thus at least as hard as the Skolem problem. Approximation of the value can be computed in exponential time for Markov chains, and exponential space for MDPs (thus the approximation problem is decidable although optimal strategies require infinite memory).

It is an open question to determine the exact complexity of the approximation problem, and whether approximations can be computed in polynomial time, or if any complexity-theoretic lower bound can be established. We are not aware of any complexity lower bounds for approximation of the

Positivity problem. Another direction for future work is to determine the memory requirement for pure strategies in MDPs.

Acknowledgment. We are grateful to the anonymous reviewers of LICS 2021 and of a previous version of this paper for insightful comments that helped improving the presentation. This research was partially supported by the grant ERC CoG 863818 (ForM-SMArt).

References

[1] S. Akshay, T. Antonopoulos, J. Ouaknine, and J. Worrell. Reachability problems for Markov chains. *Inf. Process. Lett.*, 115(2):155–158, 2015.

[2] C. Baier, M. Größer, and N. Bertrand. Probabilistic ω-automata. *J. ACM*, 59(1):1, 2012.

[3] C. Baier and J.-P. Katoen. *Principles of Model Checking*. MIT, 2008.

[4] N. Balaji, S. Kiefer, P. Novotný, G. A. Pérez, and M. Shirmohammadi. On the complexity of value iteration. In *Proc. of ICALP: Automata, Languages, and Programming*, volume 132 of *LIPIcs*, pages 102:1–102:15. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2019.

[5] A. Biere, A. Cimatti, E. M. Clarke, O. Strichman, and Y. Zhu. Bounded model checking. *Advances in Computers*, 58:117–148, 2003.

[6] J. F. Canny. Some algebraic and geometric computations in PSPACE. In *Proc. of STOC: Symposium on Theory of Computing*, pages 460–467. ACM, 1988.

[7] K. Chatterjee and L. Doyen. Graph planning with expected finite horizon. In *Proc. of LICS: Symposium on Logic in Computer Science*, pages 1–13. IEEE, 2019.

[8] K. Chatterjee and L. Doyen. Stochastic processes with expected stopping time. *CoRR*, abs/2104.07278, 2021.

[9] C. Courcoubetis and M. Yannakakis. The complexity of probabilistic verification. *J. ACM*, 42(4):857–907, 1995.

[10] L. de Alfaro. *Formal Verification of Probabilistic Systems*. PhD thesis, Stanford University, 1997.

[11] I. Dew-Becker. *Essays on Time-Varying Discount Rates*. PhD thesis, Harvard University, 2012.

[12] E. A. Emerson, A. K. Mok, A. P. Sistla, and J. Srinivasan. Quantitative temporal reasoning. *Real-Time Systems*, 4(4):331–352, 1992.

[13] P. Erdös. Ramanujan and I. In *Number Theory*, Lecture Notes in Mathematics 1395, pages 1–20. Springer, 1989.

[14] J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer-Verlag, 1997.

[15] R. G. Gallager. *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

[16] V. Halava, T. Harju, and M. Hirvensalo. Positivity of second order linear recurrent sequences. *Discrete Applied Mathematics*, 154(3):447–451, 2006.

[17] H. Howard. *Dynamic Programming and Markov Processes*. MIT, 1960.

[18] J. G. Kemeny, J. L. Snell, and A. W. Knapp. *Denumerable Markov chains*. D. Van Nostrand Company, 1966.

[19] Y. Kwon and G. Agha. Linear inequality LTL (iLTL): A model checker for discrete time Markov chains. In *Proc. of ICFEM: Formal Methods and Software Engineering*, LNCS 3308, pages 194–208. Springer, 2004.

[20] O. Madani, S. Hanks, and A. Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artif. Intell.*, 147(1-2):5–34, 2003.

[21] P. Norvig and S. J. Russell. *Artificial Intelligence - A Modern Approach (3rd ed.)*. Pearson Education, 2010.

[22] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[23] J. Ouaknine and J. Worrell. Positivity problems for low-order linear recurrence sequences. In *Proc. of SODA: Symposium on Discrete Algorithms*, pages 366–379. SIAM, 2014.

[24] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of Markov decision processes. *Math. Oper. Res.*, 12(3):441–450, 1987.

[25] A. Paz. *Introduction to probabilistic automata*. Academic Press, 1971.

[26] M. L. Puterman. *Markov decision processes*. Wiley and Sons, 1994.

[27] M. O. Rabin. Probabilistic automata. *Information & Control*, 6:3:230–245, 1963.

[28] R. Reisz. Decomposition theorems for probabilistic automata over infinite objects. *Informatica, Lithuanian Acad. Sci.*, 10:4:427–440, 1999.

[29] T. Smith and R. G. Simmons. Heuristic search value iteration for POMDPs. In *Proc. of UAI: Uncertainty in Artificial Intelligence*, pages 520–527. AUAI Press, 2004.