The Complexity of Graph-Based Reductions for Reachability in Markov Decision Processes

Stéphane Le Roux[1(✉)] and Guillermo A. Pérez[2]

[1] Department of Mathematics, Technische Universität Darmstadt, Darmstadt, Germany
leroux@mathematik.tu-darmstadt.de

[2] Departement d'Informatique, Université libre de Bruxelles, Brussels, Belgium
gperezme@ulb.ac.be

Abstract. We study the never-worse relation (NWR) for Markov decision processes with an infinite-horizon reachability objective. A state q is never worse than a state p if the maximal probability of reaching the target set of states from p is at most the same value from q, regardless of the probabilities labelling the transitions. Extremal-probability states, end components, and essential states are all special cases of the equivalence relation induced by the NWR. Using the NWR, states in the same equivalence class can be collapsed. Then, actions leading to sub-optimal states can be removed. We show that the natural decision problem associated to computing the NWR is coNP-complete. Finally, we extend a previously known incomplete polynomial-time iterative algorithm to under-approximate the NWR.

1 Introduction

Markov decision processes (MDPs) are a useful model for decision-making in the presence of a stochastic environment. They are used in several fields, including robotics, automated control, economics, manufacturing and in particular planning [20], model-based reinforcement learning [22], and formal verification [1]. We elaborate on the use of MDPs and the need for graph-based reductions thereof in verification and reinforcement learning applications below.

Several verification problems for MDPs reduce to reachability [1,5]. For instance, MDPs can be model checked against linear-time objectives (expressed in, say, LTL) by constructing an omega-automaton recognizing the set of runs that satisfy the objective and considering the product of the automaton with the original MDP [6]. In this product MDP, accepting end components—a generalization of strongly connected components—are identified and selected as target components. The question of maximizing the probability that the MDP behaviours satisfy the linear-time objective is thus reduced to maximizing the probability of reaching the target components.

The maximal reachability probability is computable in polynomial time by reduction to linear programming [1,6]. In practice, however, most model checkers

C. Baier and U. Dal Lago (Eds.): FOSSACS 2018, LNCS 10803, pp. 367–383, 2018.
https://doi.org/10.1007/978-3-319-89366-2_20

use value iteration to compute this value [9,17]. The worst-case time complexity of value iteration is pseudo-polynomial. Hence, when implementing model checkers it is usual for a graph-based pre-processing step to remove as many unnecessary states and transitions as possible while preserving the maximal reachability probability. Well-known reductions include the identification of extremal-probability states and maximal end components [1,5]. The intended outcome of this pre-processing step is a reduced amount of transition probability values that need to be considered when computing the number of iterations required by value iteration.

The main idea behind MDP reduction heuristics is to identify subsets of states from which the maximal probability of reaching the target set of states is the same. Such states are in fact redundant and can be "collapsed". Figure 1 depicts an MDP with actions and probabilities omitted for clarity. From p and q there are strategies to ensure that s is reached with probability 1. The same holds for t. For instance, from p, to get to t almost surely, one plays to go to the distribution directly below q; from q, to the distribution above q. Since from the state p, there is no strategy to ensure that q is reached with probability 1, p and q do not form an *end component*. In fact, to the best of our knowledge, no known MDP reduction heuristic captures this example (i.e., recognizes that p and q have the same maximal reachability probability for all possible values of the transition probabilities).

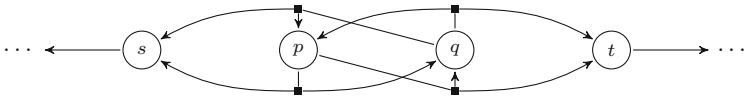

Fig. 1. An MDP with states depicted as circles and distributions as squares. The maximal reachability probability values from p and q are the same since, from both, one can enforce to reach s with probability 1, or t with probability 1, using different strategies.

In reinforcement learning the actual probabilities labelling the transitions of an MDP are not assumed to be known in advance. Thus, they have to be estimated by experimenting with different actions in different states and collecting statistics about the observed outcomes [14]. In order for the statistics to be good approximations, the number of experiments has to be high enough. In particular, when the approximations are required to be *probably approximately correct* [23], the necessary and sufficient number of experiments is pseudo-polynomial [13]. Furthermore, the expected number of steps before reaching a particular state even once may already be exponential (even if all the probabilities are fixed). The fact that an excessive amount of experiments is required is a known drawback of reinforcement learning [15,19].

A natural and key question to ask in this context is whether the maximal reachability probability does indeed depend on the actual value of the probability labelling a particular transition of the MDP. If this is not the case, then it need

not be learnt. One natural way to remove transition probabilities which do not affect the maximal reachability value is to apply model checking MDP reduction techniques.

Contributions and Structure of the Paper. We view the directed graph underlying an MDP as a directed bipartite graph. Vertices in this graph are controlled by players *Protagonist* and *Nature*. Nature is only allowed to choose full-support probability distributions for each one of her vertices, thus instantiating an MDP from the graph; Protagonist has strategies just as he would in an MDP. Hence, we consider infinite families of MDPs with the same support. In the game played between Protagonist and Nature, and for vertices u and v, we are interested in knowing whether the maximal reachability probability from u is never (in any of the MDPs with the game as its underlying directed graph) worse than the same value from v.

In Sect. 2 we give the required definitions. We formalize the *never-worse relation* in Sect. 3. We also show that we can "collapse" sets of equivalent vertices with respect to the NWR (Theorem 1) and remove sub-optimal edges according to the NWR (Theorem 2). Finally, we also argue that the NWR generalizes most known heuristics to reduce MDP size before applying linear programming or value iteration. Then, in Sect. 4 we give a graph-based characterization of the relation (Theorem 3), which in turn gives us a coNP upper bound on its complexity. A matching lower bound is presented in Sect. 5 (Theorem 4). To conclude, we recall and extend an iterative algorithm to efficiently (in polynomial time) under-approximate the never-worse relation from [2].

Previous and Related Work. Reductions for MDP model checking were considered in [5,7]. From the reductions studied in both papers, extremal-probability states, essential states, and end components are computable using only graph-based algorithms. In [3], learning-based techniques are proposed to obtain approximations of the maximal reachability probability in MDPs. Their algorithms, however, do rely on the actual probability values of the MDP.

This work is also related to the widely studied model of interval MDPs, where the transition probabilities are given as intervals meant to model the uncertainty of the numerical values. Numberless MDPs [11] are a particular case of the latter in which values are only known to be zero or non-zero. In the context of numberless MDPs, a special case of the question we study can be simply rephrased as the comparison of the maximal reachability values of two given states.

In [2] a preliminary version of the iterative algorithm we give in Sect. 6 was described, implemented, and shown to be efficient in practice. Proposition 1 was first stated therein. In contrast with [2], we focus chiefly on characterizing the never-worse relation and determining its computational complexity.

2 Preliminaries

We use set-theoretic notation to indicate whether a letter $b \in \Sigma$ *occurs* in a word $\alpha = a_0 \dots a_k \in \Sigma^*$, i.e. $b \in \alpha$ if and only if $b = a_i$ for some $0 \leq i \leq k$.

Consider a directed graph $\mathcal{G} = (V, E)$ and a vertex $u \in V$. We write uE for the set of *successors* of u. That is to say, $uE := \{v \in V \mid (u, v) \in E\}$. We say that a path $\pi = u_0 \dots u_k \in V^*$ in $\mathcal{G}$ *visits* a vertex v if $v \in \pi$. We also say that π is a v–T path, for $T \subseteq V$, if $u_0 = v$ and $u_k \in T$.

2.1 Stochastic Models

Let S be a finite set. We denote by $\mathbb{D}(S)$ the set of all *(rational) probabilistic distributions* on S, i.e. the set of all functions $f : S \to \mathbb{Q}_{\geq 0}$ such that $\sum_{s \in S} f(s) = 1$. A probabilistic distribution $f \in \mathbb{D}(S)$ has *full support* if $f(s) > 0$ for all $s \in S$.

Definition 1 (Markov chains). *A* Markov chain $\mathcal{C}$ *is a tuple* (Q, δ) *where* Q *is a finite set of states and* δ *is a probabilistic transition function* $\delta : Q \to \mathbb{D}(Q)$.

A *run* of a Markov chain is a finite non-empty word $\varrho = p_0 \dots p_n$ over Q. We say ϱ *reaches* q if $q = p_i$ for some $0 \leq i \leq n$. The *probability of the run* is $\prod_{0 \leq i < n} \delta(p_i, p_{i+1})$.

Let $T \subseteq Q$ be a set of states. The *probability of (eventually) reaching* T in $\mathcal{C}$ from q_0, which will be denoted by $\mathbb{P}^{q_0}_{\mathcal{C}}[\Diamond T]$, is the measure of the runs of $\mathcal{C}$ that start at q_0 and reach T. For convenience, let us first define the *probability of staying in states from* $S \subseteq Q$ *until* T *is reached*[1], written $\mathbb{P}^{q_0}_{\mathcal{C}}[S \mathbin{\mathsf{U}} T]$, as 1 if $q_0 \in T$ and otherwise

$$\sum \left\{ \prod_{0 \leq i < n} \delta(q_i, q_{i+1}) \;\middle|\; q_0 \dots q_n \in (S \setminus T)^* T \text{ for } n \geq 1 \right\}.$$

We then define $\mathbb{P}^{q_0}_{\mathcal{C}}[\Diamond T] := \mathbb{P}^{q_0}_{\mathcal{C}}[Q \mathbin{\mathsf{U}} T]$.

When all runs from q_0 to T reach some set $U \subseteq Q$ before, the probability of reaching T can be decomposed into a finite sum as in the lemma below.

Lemma 1. *Consider a Markov chain* $\mathcal{C} = (Q, \delta)$*, sets of states* $U, T \subseteq Q$*, and a state* $q_0 \in Q \setminus U$*. If* $\mathbb{P}^{q_0}_{\mathcal{C}}[(Q \setminus U) \mathbin{\mathsf{U}} T] = 0$*, then*

$$\mathbb{P}^{q_0}_{\mathcal{C}}[\Diamond T] = \sum_{u \in U} \mathbb{P}^{q_0}_{\mathcal{C}}[(Q \setminus U) \mathbin{\mathsf{U}} u] \, \mathbb{P}^{u}_{\mathcal{C}}[\Diamond T].$$

Definition 2 (Markov decision processes). *A* (finite, discrete-time) Markov decision process $\mathcal{M}$*, MDP for short, is a tuple* (Q, A, δ, T) *where* Q *is a finite set of states,* A *a finite set of actions,* $\delta : Q \times A \to \mathbb{D}(Q)$ *a* probabilistic transition function*, and* $T \subseteq Q$ *a set of* target *states.*

For convenience, we write $\delta(q|p, a)$ instead of $\delta(p, a)(q)$.

[1] $S \mathbin{\mathsf{U}} T$ should be read as "S *until* T" and not understood as a set union.

Definition 3 (Strategies). *A* (memoryless deterministic) strategy σ *in an MDP* $\mathcal{M} = (Q, A, \delta, T)$ *is a function* $\sigma : Q \to A$.

Note that we have deliberately defined only memoryless deterministic strategies. This is at no loss of generality since, in this work, we focus on maximizing the probability of reaching a set of states. It is known that for this type of objective, memoryless deterministic strategies suffice [18].

From MDPs to Chains. An MDP $\mathcal{M} = (Q, A, \delta, T)$ and a strategy σ induce the Markov chain $\mathcal{M}^\sigma = (Q, \mu)$ where $\mu(q) = \delta(q, \sigma(q))$ for all $q \in Q$.

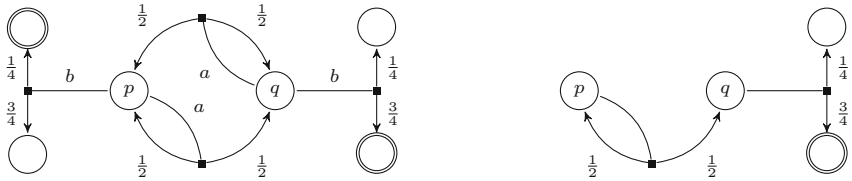

Fig. 2. On the left we have an MDP with actions $\{a, b\}$. On the right we have the Markov chain induced by the left MDP and the strategy $\{p \mapsto a, q \mapsto b\}$.

Example 1. Figure 2 depicts an MDP on the left. Circles represent states; double-circles, target states; and squares, distributions. The labels on arrows from states to distributions are actions; those on arrows from distributions to states, probabilities.

Consider the strategy σ that plays from p the action a and from q the action b, i.e. $\sigma(p) = a$ and $\sigma(q) = b$. The Markov chain on the right is the chain induced by σ and the MDP on the left. Note that we no longer have action labels.

The probability of reaching a target state from q under σ is easily seen to be 3/4. In other words, if we write $\mathcal{M}$ for the MDP and T for the set of target states then $\mathbb{P}^q_{\mathcal{M}^\sigma}[\Diamond T] = \frac{3}{4}$.

2.2 Reachability Games Against Nature

We will speak about families of MDPs whose probabilistic transition functions have the same support. To do so, we abstract away the probabilities and focus on a game played on a graph. That is, given an MDP $\mathcal{M} = (Q, A, \delta, T)$ we consider its *underlying directed graph* $\mathcal{G}_\mathcal{M} = (V, E)$ where $V := Q \cup (Q \times A)$ and $E := \{(q, \langle q, a\rangle) \in Q \times (Q \times A)\} \cup \{(\langle p, a\rangle, q) \mid \delta(q|p, a) > 0\}$. In $\mathcal{G}_\mathcal{M}$, *Nature* controls the vertices $Q \times A$. We formalize the game and the *arena* it is played on below.

Definition 4 (Target arena). *A* target arena $\mathcal{A}$ *is a tuple* (V, V_P, E, T) *such that* $(V_P, V_N := V \setminus V_P, E)$ *is a bipartite directed graph,* $T \subseteq V_P$ *is a set of* target *vertices, and* $uE \neq \varnothing$ *for all* $u \in V_N$.

Informally, there are two agents in a target arena: *Nature*, who controls the vertices in V_N, and *Protagonist*, who controls the vertices in V_P.

From Arenas to MDPs. A target arena $\mathcal{A} = (V, V_P, E, T)$ together with a family of probability distributions $\mu = (\mu_u \in \mathbb{D}(uE))_{u \in V_N}$ induce an MDP. Formally, let $\mathcal{A}_\mu$ be the MDP (Q, A, δ, T) where $Q = V_P \uplus \{\bot\}$, $A = V_N$, $\delta(q|p, a)$ is $\mu_a(q)$ if $(p, a), (a, q) \in E$ and 0 otherwise, for all $p \in V_P \cup \{\bot\}$ and $a \in A$ we have $\delta(\bot|p, a) = 1$ if $(p, a) \notin E$.

The Value of a Vertex. Consider a target arena $\mathcal{A} = (V, V_P, E, T)$ and a vertex $v \in V_P$. We define its *(maximal reachability probability) value* with respect to a family of full-support probability distributions μ as $\mathrm{Val}^\mu(v) := \max_\sigma \mathbb{P}^v_{\mathcal{A}^\sigma_\mu}[\Diamond T]$. For $u \in V_N$ we set $\mathrm{Val}^\mu(u) := \sum\{\mu_u(v)\mathrm{Val}^\mu(v) \mid v \in uE\}$.

3 The Never-Worse Relation

We are now in a position to define the relation that we study in this work. Let us fix a target arena $\mathcal{A} = (V, V_P, E, T)$.

Definition 5 (The never-worse relation (NWR)). *A subset $W \subseteq V$ of vertices is* never worse *than a vertex $v \in V$, written $v \trianglelefteq W$, if and only if*

$$\forall \mu = (\mu_u \in \mathbb{D}(uE))_{u \in V_N}, \exists w \in W : \mathrm{Val}^\mu(v) \leq \mathrm{Val}^\mu(w)$$

*where all the μ_u have **full support**. We write $v \sim w$ if $v \trianglelefteq \{w\}$ and $w \trianglelefteq \{v\}$.*

It should be clear from the definition that $\sim$ is an equivalence relation. For $u \in V$ let us denote by $\tilde{u}$ the set of vertices that are $\sim$-equivalent and belong to the same owner, i.e. $\tilde{u}$ is $\{v \in V_P \mid v \sim u\}$ if $u \in V_P$ and $\{v \in V_N \mid v \sim u\}$ otherwise.

Fig. 3. Two target arenas with $T = \{fin\}$ are shown. Round vertices are elements from V_P; square vertices, from V_N. In the left target arena we have that $p \trianglelefteq \{q\}$ and $q \trianglelefteq \{p\}$ since any path from either vertex visits t before T—see Lemma 1. In the right target arena we have that $t \trianglelefteq \{p\}$—see Proposition 1.

Example 2. Consider the left target arena depicted in Fig. 3. Using Lemma 1, it is easy to show that neither p nor q is ever worse than the other since t is visited before *fin* by all paths starting from p or q.

The literature contains various heuristics which consist in computing sets of states and "collapsing" them to reduce the size of the MDP without affecting the maximal reachability probability of the remaining states. We now show that we can collapse equivalence classes and, further, remove sub-optimal distributions using the NWR.

3.1 The Usefulness of the NWR

We will now formalize the idea of "collapsing" equivalent vertices with respect to the NWR. For convenience, we will also remove self-loops while doing so.

Consider a target arena $\mathcal{A} = (V, V_P, E, T)$. We denote by $\mathcal{A}_{/\sim}$ its $\sim$-*quotient*. That is, $\mathcal{A}_{/\sim}$ is the target arena (S, S_P, R, U) where $S_P = \{\tilde{v} \mid \exists v \in V_P\}$, $S = \{\tilde{v} \mid \exists v \in V_N\} \cup S_P$, $U = \{\tilde{t} \mid \exists t \in T\}$, and

$$\begin{aligned} R =& \{(\tilde{u}, \tilde{v}) \mid \exists (u, v) \in (V_P \times V_N) \cap E : vE \setminus \tilde{u} \neq \varnothing\} \\ & \cup \{(\tilde{u}, \tilde{v}) \mid \exists (u, v) \in (V_N \times V_P) \cap E\}. \end{aligned}$$

For a family $\mu = (\mu_u \in \mathbb{D}(uE))_{u \in V_N}$ of full-support distributions we denote by $\mu_{/\sim}$ the family $\nu = (\nu_{\tilde{u}} \in \mathbb{D}(\tilde{u}R))_{\tilde{u} \in S_N}$ defined as follows. For all $\tilde{u} \in S_N$ and all $\tilde{v} \in \tilde{u}R$ we have $\nu_{\tilde{u}}(\tilde{v}) = \sum_{w \in \tilde{v}} \mu_u(w)$, where u is any element of $\tilde{u}$.

The following property of the $\sim$-quotient follows from the fact that all the vertices in $\tilde{v}$ have the same maximal probability of reaching the target vertices.

Theorem 1. *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$. For all families $\mu = (\mu_u \in \mathbb{D}(uE))_{u \in V_N}$ of full-support probability distributions and all $v \in V_P$ we have*

$$\max_{\sigma} \mathbb{P}^{v}_{\mathcal{A}^{\sigma}_{\mu}}[\Diamond T] = \max_{\sigma'} \mathbb{P}^{\tilde{v}}_{\mathcal{B}^{\sigma'}_{\nu}}[\Diamond U],$$

where $\mathcal{B} = \mathcal{A}_{/\sim}$, $\nu = \mu_{/\sim}$, and $U = \{\tilde{t} \mid \exists t \in T\}$.

We can further remove edges that lead to sub-optimal Nature vertices. When this is done after $\sim$-quotienting the maximal reachability probabilities are preserved.

Theorem 2. *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$ such that $\mathcal{A}_{/\sim} = \mathcal{A}$. For all families $\mu = (\mu_u \in \mathbb{D}(uE))_{u \in V_N}$ of full-support probability distributions, for all $(w, x) \in E \cap (V_P \times V_N)$ such that $x \trianglelefteq (wE \setminus \{x\})$, and all $v \in V_P$ we have*

$$\max_{\sigma} \mathbb{P}^{v}_{\mathcal{A}^{\sigma}_{\mu}}[\Diamond T] = \max_{\sigma'} \mathbb{P}^{v}_{\mathcal{B}^{\sigma'}_{\mu}}[\Diamond T],$$

where $\mathcal{B} = (V, V_P, E \setminus \{(w, x)\}, T)$.

3.2 Known Efficiently-Computable Special Cases

We now recall the definitions of the set of extremal-probability states, end components, and essential states. Then, we observe that for all these sets of states their maximal probability reachability coincide and their definitions are independent of the probabilities labelling the transitions of the MDP. Hence, they are subsets of the set of the equivalence classes induced by $\sim$.

Extremal-Probability States. The set of *extremal-probability states* of an MDP $\mathcal{M} = (Q, A, \delta, T)$ consists of the set of states with maximal probability reachability 0 and 1. Both sets can be computed in polynomial time [1,4]. We give below a game-based definition of both sets inspired by the classical polynomial-time algorithm to compute them (see, e.g., [1]). Let us fix a target arena $\mathcal{A} = (V, V_P, E, T)$ for the sequel.

For a set $T \subseteq V$, let us write $\mathbf{Z}_T := \{v \in V \mid T \text{ is not reachable from } v\}$.

(Almost-Surely Winning) Strategies. A strategy for Protagonist in a target arena is a function $\sigma : V_P \to V_N$. We then say that a path $v_0 \dots v_n \in V^*$ is *consistent with* σ if $v_i \in V_P \implies \sigma(v_i) = v_{i+1}$ for all $0 \leq i < n$. Let $\mathbf{Reach}(v_0, \sigma)$ denote the set of vertices reachable from v_0 under σ, i.e. $\mathbf{Reach}(v_0, \sigma) := \{v_k \mid v_0 \dots v_k \text{ is a path consistent with } \sigma\}$.

We say that a strategy σ for Protagonist is *almost-surely winning from* $u_0 \in V$ *to* $T \subseteq V_P$ if, after modifying the arena to make all $t \in T$ into sinks, for all $v_0 \in \mathbf{Reach}(u_0, \sigma)$ we have $\mathbf{Reach}(v_0, \sigma) \cap T \neq \varnothing$. We denote the set of all such strategies by $\mathbf{Win}_T^{v_0}$.

The following properties regarding almost-surely winning strategies in a target arena follow from the correctness of the graph-based algorithm used to compute extremal-probability states in an MDP [1, Lemma 10.108].

Lemma 2 (From [1]). *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$. For all families $\mu = (\mu_u \in \mathbb{D}(uE))_{u \in V_N}$ of full-support probability distributions, for all $v \in V_P$ the following hold.*

(i) $\max_\sigma \mathbb{P}^v_{\mathcal{A}^\sigma_\mu}[\Diamond T] = 0 \iff v \in \mathbf{Z}_T$
(ii) $\forall \sigma : \sigma \in \mathbf{Win}_T^v \iff \mathbb{P}^v_{\mathcal{A}^\sigma_\mu}[\Diamond T] = 1$

End Components. Let us consider an MDP $\mathcal{M} = (Q, A, \delta, T)$. A set $S \subseteq Q$ of states is an *end component* in $\mathcal{M}$ if for all pairs of states $p, q \in S$ there exists a strategy σ such that $\mathbb{P}^p_{\mathcal{M}^\sigma}[S \cup q] = 1$.

Example 3. Let us consider the MDP shown on the left in Fig. 2. The set $\{p, q\}$ is an end component since, by playing a from both states, one can ensure to reach either state from the other with probability 1.

It follows immediately from the definition of end component that the maximal probability of reaching T from states in the same end component is the same.

Lemma 3. *Let $S \subseteq Q$ be an end component in $\mathcal{M}$. For all $p, q \in S$ we have that* $\max_\sigma \mathbb{P}^p_{\mathcal{M}^\sigma}[\Diamond T] = \max_\sigma \mathbb{P}^q_{\mathcal{M}^\sigma}[\Diamond T]$.

We say an end component is *maximal* if it is maximal with respect to set inclusion. Furthermore, from the definition of end components in MDPs and Lemma 2 it follows that we can lift the notion of end component to target arenas. More precisely, a set $S \subseteq V_P$ is an end component in $\mathcal{A}$ if and only if for some family of

full-support probability distributions μ we have that S is an end component in $\mathcal{A}_\mu$ (if and only if for all μ' the set S is an end component in $\mathcal{A}_{\mu'}$).

The set of all maximal end components of a target arena can be computed in polynomial time using an algorithm based on the strongly connected components of the graph [1,8].

Essential States. Consider a target arena $\mathcal{A} = (V, V_P, E, T)$ and let $\sqsubseteq$ be the smallest relation satisfying the following. For all $u \in V_P$ we have $u \sqsubseteq u$. For all $u_0, v \in V_P \setminus \mathbf{Z}_T$ such that $u_0 \neq v$ we have $u_0 \sqsubseteq v$ if for all paths $u_0u_1u_2$ we have that $u_2 \sqsubseteq v$ and there is at least one such path. Intuitively, $u \sqsubseteq v$ holds whenever all paths starting from u reach v. In [7], the maximal vertices according to $\sqsubseteq$ are called *essential states*[2].

Lemma 4 (From [7]). *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$. For all families $\mu = (\mu_u \in \mathbb{D}(uE))_{u \in V_N}$ of full-support probability distributions, for all $v \in V_P$ and all essential states w, if $v \sqsubseteq w$ then* $\max_\sigma \mathbb{P}^v_{\mathcal{A}^\sigma_\mu}[\Diamond T] = \max_{\sigma'} \mathbb{P}^w_{\mathcal{A}^{\sigma'}_\mu}[\Diamond T]$.

Note that, in the left arena in Fig. 3, $p \sqsubseteq t$ does not hold since there is a cycle between p and q which does not visit t.

It was also shown in [7] that the $\sqsubseteq$ relation is computable in polynomial time.

4 Graph-Based Characterization of the NWR

In this section we give a characterization of the NWR that is reminiscent of the topological-based value iteration proposed in [5]. The main intuition behind our characterization is as follows. If $v \trianglelefteq W$ does not hold, then for all $0 < \varepsilon < 1$ there is some family μ of full-support distributions such that $\mathrm{Val}^\mu(v)$ is at least $1 - \varepsilon$, while $\mathrm{Val}^\mu(w)$ is at most ε for all $w \in W$. In turn, this must mean that there is a path from v to T which can be assigned a high probability by μ while, from W, all paths go with high probability to $\mathbf{Z}_T$.

We capture the idea of separating a "good" v–T path from all paths starting from W by using partitioning of V into layers $S_i \subseteq V$. Intuitively, we would like it to be easy to construct a family μ of probability distributions such that from all vertices in S_{i+1} all paths going to vertices outside of S_{i+1} end up, with high probability, in lower layers, i.e. some S_k with $k < i$. A formal definition follows.

Definition 6 (Drift partition and vertices). *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$ and a partition $(S_i)_{0 \leq i \leq k}$ of V. For all $0 \leq i \leq k$, let $S_i^+ := \cup_{i<j} S_j$ and $S_i^- := \cup_{j<i} S_j$, and let $D_i := \{v \in S_i \cap V_N \mid vE \cap S_i^- \neq \varnothing\}$. We define the set $D := \cup_{0<i<k} D_i$ of* drift vertices. *The partition is called a* drift partition *if the following hold.*

- *For all $i \leq k$ and all $v \in S_i \cap V_P$ we have $vE \cap S_i^+ = \varnothing$.*
- *For all $i \leq k$ and all $v \in S_i \cap V_N$ we have $vE \cap S_i^+ \neq \varnothing \implies v \in D$.*

[2] This is not the usual notion of essential states from classical Markov chain theory.

Using drift partitions, we can now formalize our characterization of the negation of the NWR.

Theorem 3. *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$, a non-empty set of vertices $W \subseteq V$, and a vertex $v \in V$. The following are equivalent*

(i) $\neg\,(v \preceq W)$
(ii) There exists a drift partition $(S_i)_{0\leq i\leq k}$ and a simple path π starting in v and ending in T such that $\pi \subseteq S_k$ and $W \subseteq S_k^-$.

Before proving Theorem 3 we need an additional definition and two intermediate results.

Definition 7 (Value-monotone paths). *Let $\mathcal{A} = (V, V_P, E, T)$ be a target arena and consider a family of full-support probability distributions $\mu = (\mu_u \in \mathbb{D}(uE))_{u\in V_N}$. A path $v_0 \dots v_k$ is* μ-non-increasing *if and only if $\mathrm{Val}^\mu(v_{i+1}) \leq \mathrm{Val}^\mu(v_i)$ for all $0 \leq i < k$; it is* μ-non-decreasing *if and only if $\mathrm{Val}^\mu(v_i) \leq \mathrm{Val}^\mu(v_{i+1})$ for all $0 \leq i < k$.*

It can be shown that from any path in a target arena ending in T one can obtain a simple non-decreasing one.

Lemma 5. *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$ and a family of full-support probability distributions $\mu = (\mu_u \in \mathbb{D}(uE))_{u\in V_N}$. If there is a path from some $v \in V$ to T, there is also a simple μ-non-decreasing one.*

Additionally, we will make use of the following properties regarding vertex-values. They formalize the relation between the value of a vertex, its owner, and the values of its successors.

Lemma 6. *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$ and a family of full-support probability distributions $\mu = (\mu_u \in \mathbb{D}(uE))_{u\in V_N}$.*

(i) For all $u \in V_P$, for all successors $v \in uE$ it holds that $\mathrm{Val}^\mu(v) \leq \mathrm{Val}^\mu(u)$.
(ii) For all $u \in V_N$ it holds that

$$(\exists v \in uE : \mathrm{Val}^\mu(u) < \mathrm{Val}^\mu(v)) \implies (\exists w \in uE : \mathrm{Val}^\mu(w) < \mathrm{Val}^\mu(u)).$$

Proof (of Theorem 3). Recall that, by definition, (i) holds if and only if there exists a family $\mu = (\mu_u \in \mathbb{D}(uE))_{u\in V_N}$ of full-support probability distributions such that $\forall w \in W : \mathrm{Val}^\mu(w) < \mathrm{Val}^\mu(v)$.

Let us prove (i) $\implies$ (ii). Let $x_0 < x_1 < \dots$ be the finitely many (i.e. at most $|V|$) values that occur in the MDP $\mathcal{A}_\mu$, and let k be such that $\mathrm{Val}^\mu(v) = x_k$. For all $0 \leq i < k$ let $S_i := \{u \in V \mid \mathrm{Val}^\mu(u) = x_i\}$, and let $S_k := V \setminus \cup_{i<k} S_i$. Let us show below that the S_i form a drift partition.

- $\forall i \leq k, \forall u \in S_i \cap S_P : uE \cap S_i^+ = \varnothing$ by Lemma 6(i) (for $i < k$) and since $S_k^+ = \varnothing$.
- $\forall i \leq k, \forall u \in S_i \cap S_N : uE \cap S_i^+ \neq \varnothing \implies x \in D$ by Lemma 6(ii) (for $i < k$) and since $S_k^+ = \varnothing$.

We have that $\mathrm{Val}^\mu(w) < \mathrm{Val}^\mu(v) = x_k$ for all $w \in W$, by assumption, so $W \subseteq S_k^-$ by construction. By Lemma 5 there exists a simple μ-non-decreasing path π from v to T, so all the vertices occurring in π have values at least $\mathrm{Val}^\mu(v)$, so $\pi \subseteq S_k$.

We will prove (ii) $\implies$ (i) by defining some full-support distribution family μ. The definition will be partial only, first on $\pi \cap V_N$, and then on the drift vertices in $V \setminus S_k$. Let $0 < \varepsilon < 1$, which is meant to be small enough. Let us write $\pi = v_0 \dots v_n$ so that $v_0 = v$ and $v_n \in T$. Let us define μ on $\pi \cap V_N$ as follows: for all $i < n$, if $v_i \in V_N$ let $\mu_{v_i}(v_{i+1}) := 1 - \varepsilon$. Let σ be an arbitrary Protagonist strategy such that for all $i < n$, if $v_i \in V_P$ then $\sigma(v_i) := v_{i+1}$. Therefore

$$\begin{aligned}
(1-\varepsilon)^{|V|} &\leq (1-\varepsilon)^n && \text{since } \pi \text{ is simple} \\
&\leq \prod_{i<n, v_i \in S_N} \mu_{v_i}(v_{i+1}) && \text{by definition of } \mu \\
&\leq \mathbb{P}^v_{\mathcal{A}^\sigma_\mu}[\Diamond T] \\
&\leq \max_{\sigma'} \mathbb{P}^v_{\mathcal{A}^{\sigma'}_\mu}[\Diamond T] = \mathrm{Val}^\mu(v). && (1)
\end{aligned}$$

So, for $0 < \varepsilon < 1 - \frac{1}{\sqrt[|V|]{2}}$, we have $\frac{1}{2} < (1-\varepsilon)^{|V|} \leq \mathrm{Val}^\mu(v)$. Below we will further define μ such that $\mathrm{Val}^\mu(w) \leq 1 - (1-\varepsilon)^{|V|} < \frac{1}{2}$ for all $w \in W$ and all $0 < \varepsilon < 1 - \frac{1}{\sqrt[|V|]{2}}$, which will prove (ii) $\implies$ (i). However, the last part of the proof is more difficult.

For all $1 \leq i \leq k$, for all drift vertices $u \in S_i$, let $\varrho(u)$ be a successor of u in S_i^-. Such a $\varrho(u)$ exists by definition of the drift vertices. Then let $\mu_u(\varrho(u)) := 1 - \varepsilon$. We then claim that

$$\forall u \in D : (1-\varepsilon)(1 - \mathbb{P}^{\varrho(u)}_{\mathcal{A}^\sigma_\mu}[\Diamond T]) \leq 1 - \mathbb{P}^u_{\mathcal{A}^\sigma_\mu}[\Diamond T]. \quad (2)$$

Indeed, $1 - \mathbb{P}^u_{\mathcal{A}^\sigma_\mu}[\Diamond T]$ is the probability that, starting at u and following σ, T is never reached; and $(1-\varepsilon)(1 - \mathbb{P}^{\varrho(u)}_{\mathcal{A}^\sigma_\mu}[\Diamond T])$ is the probability that, starting at u and following σ, the second vertex is $\varrho(u)$ and T is never reached.

Now let σ be an arbitrary strategy, and let us prove the following by induction on j.

$$\forall 0 \leq j < k, \forall w \in S_j \cup S_j^- : \mathbb{P}^w_{\mathcal{A}^\sigma_\mu}[\Diamond T] \leq 1 - (1-\varepsilon)^j$$

Base case, $j = 0$: by assumption W is non-empty and included in S_k^-, so $0 < k$. Also by assumption $T \subseteq S_k$, so $T \cap S_0 = \varnothing$. By definition of a drift partition, there are no edges going out of S_0, regardless of whether the starting vertex is in V_P or V_N. So there is no path from w to T, which implies $\mathrm{Val}^\mu(w) = 0$ for all $w \in S_0$, and the claim holds for the base case. Inductive case, let $w \in S_j$, let $D' := D \cap (S_j \cup S_j^-)$ and let us argue that every path π from w to T must at some point leave $S_j \cup S_j^-$ to reach a vertex with higher index, i.e. there is some edge (π_i, π_{i+1}) from $\pi_i \in S_j \cup S_j^-$ to some $\pi_{i+1} \in S_\ell$ with $j < \ell$. By definition

of a drift partition, π_i must also be a drift vertex, i.e. $\pi_i \in D'$. Thus, if we let $F := V_P \setminus D'$, Lemma 1 implies that $\mathbb{P}^w_{\mathcal{A}^\sigma_\mu}[\Diamond T] = \sum_{u \in D'} \mathbb{P}^w_{\mathcal{A}^\sigma_\mu}[F \cup u]\, \mathbb{P}^u_{\mathcal{A}^\sigma_\mu}[\Diamond T]$. Now, since

$$
\begin{aligned}
&\sum_{u \in D'} \mathbb{P}^u_{\mathcal{A}^\sigma_\mu}[\Diamond T] \\
&= \sum_{u \in D \cap S_j^-} \mathbb{P}^u_{\mathcal{A}^\sigma_\mu}[\Diamond T] + \sum_{u \in D_j} \mathbb{P}^u_{\mathcal{A}^\sigma_\mu}[\Diamond T] && \text{by splitting the sum} \\
&\leq \sum_{u \in D \cap S_j^-} \mathbb{P}^u_{\mathcal{A}^\sigma_\mu}[\Diamond T] + \sum_{u \in D_j} (1 - (1-\varepsilon)(1 - \mathbb{P}^{\varrho(u)}_{\mathcal{A}^\sigma_\mu}[\Diamond T])) && \text{by (2)} \\
&\leq \sum_{u \in D \cap S_j^-} (1 - (1-\varepsilon)^{j-1}) + && \text{by IH and since} \\
&\quad \sum_{u \in D_j} (1 - (1-\varepsilon)(1-\varepsilon)^{j-1}) && \forall x \in D_j : \varrho(x) \in S_j^- \\
&\leq \sum_{u \in D'} (1 - (1-\varepsilon)^j) && (1-\varepsilon)^j \leq (1-\varepsilon)^{j-1}
\end{aligned}
$$

and $\sum_{u \in D'} \mathbb{P}^w_{\mathcal{A}^\sigma_\mu}[F \cup u] \leq 1$, we have that $\mathbb{P}^w_{\mathcal{A}^\sigma_\mu}[\Diamond T] \leq 1 - (1-\varepsilon)^j$. The induction is thus complete. Since σ is arbitrary in the calculations above, and since $j < k \leq |V|$, we find that $\mathrm{Val}^\mu(w) \leq 1 - (1-\varepsilon)^{|V|}$ for all $w \in W \subseteq S_k^-$.

For $0 < \varepsilon < 1 - \frac{1}{\sqrt[|V|]{2}}$ we have $\frac{1}{2} < (1-\varepsilon)^{|V|}$, as mentioned after (1), so $\mathrm{Val}^\mu(w) \leq 1 - (1-\varepsilon)^{|V|} < \frac{1}{2}$. □

5 Intractability of the NWR

It follows from Theorem 3 that we can decide whether a vertex is sometimes worse than a set of vertices by guessing a partition of the vertices and verifying that it is a drift partition. The verification can clearly be done in polynomial time.

Corollary 1. *Given a target arena $\mathcal{A} = (V, V_P, E, T)$, a non-empty set $W \subseteq V$, and a vertex $v \in V$, determining whether $v \trianglelefteq W$ is decidable and in* coNP.

We will now show that the problem is in fact coNP-complete already for Markov chains.

Theorem 4. *Given a target arena $\mathcal{A} = (V, V_P, E, T)$, a non-empty vertex set $W \subseteq V$, and a vertex $v \in V$, determining whether $v \trianglelefteq W$ is* coNP*-complete even if $|uE| = 1$ for all $u \in V_P$.*

The idea is to reduce the 2-Disjoint Paths problem (2DP) to the existence of a drift partition witnessing that $v \trianglelefteq \{w\}$ does not hold, for some $v \in V$. Recall that 2DP asks, given a directed graph $\mathcal{G} = (V, E)$ and vertex pairs

$(s_1, t_1), (s_2, t_2) \in V \times V$, whether there exists an s_1–t_1 path π_1 and an s_2–t_2 path π_2 such that π_1 and π_2 are vertex disjoint, i.e. $\pi_1 \cap \pi_2 = \varnothing$. The problem is known to be NP-complete [10,12]. In the sequel, we assume without loss of generality that (a) t_1 and t_2 are reachable from all $s \in V \setminus \{t_1, t_2\}$; and (b) t_1 and t_2 are the only sinks $\mathcal{G}$.

Proof (of Theorem 4). From the 2DP input instance, we construct the target arena $\mathcal{A} = (S, S_P, R, T)$ with $S := V \cup E$, $R := \{(u, \langle u, v\rangle), (\langle u, v\rangle, v) \in S \times S \mid (u, v) \in E \text{ or } u = v \in \{t_1, t_2\}\}$, $S_P := V \times V$, and $T := \{\langle t_1, t_1\rangle\}$. We will show there are vertex-disjoint s_1–t_1 and s_2–t_2 paths in $\mathcal{G}$ if and only if there is a drift partition $(S_i)_{0 \le i \le k}$ and a simple s_1–t_1 path π such that $\pi \subseteq S_k$ and $s_2 \in S_k^-$. The result will then follow from Theorem 3.

Suppose we have a drift partition $(S_i)_{0 \le i \le k}$ with $s_2 \in S_k^-$ and a simple path $\pi = v_0 \langle v_0, v_1\rangle \dots \langle v_{n-1}, v_n\rangle v_n$ with $v_0 = s_1, v_n = t_1$. Since the set $\{t_2, \langle t_2, t_2\rangle\}$ is *trapping* in $\mathcal{A}$, i.e. all paths from vertices in the set visit only vertices from it, we can assume that $S_0 = \{t_2, \langle t_2, t_2\rangle\}$. (Indeed, for any drift partition, one can obtain a new drift partition by moving any trapping set to a new lowest layer.) Now, using the assumption that t_2 is reachable from all $s \in V \setminus \{t_1, t_2\}$ one can show by induction that for all $0 \le j < k$ and for all $\varrho = u_0 \in S_j$ there is a path $u_0 \dots u_m$ in $\mathcal{G}$ with $u_m = t_2$ and $\varrho \subseteq S_{j+1}^-$. This implies that there is a s_2–t_2 path π_2 in $\mathcal{G}$ such that $\pi_2 \subseteq S_k^-$. It follows that π_2 is vertex disjoint with the s_1–t_1 path $v_0 \dots v_n$ in $\mathcal{G}$.

Now, let us suppose that we have s_1–t_1 and s_2–t_2 vertex disjoint paths $\pi_1 = u_0 \dots u_n$ and $\pi_2 = v_0 \dots v_m$. Clearly, we can assume both π_1, π_2 are simple. We will construct a partition $(S_i)_{0 \le i \le m+1}$ and show that it is indeed a drift partition, that $u_0 \langle u_0, u_1\rangle \dots \langle u_{n-1}, u_n\rangle u_n \subseteq S_{m+1}$, and $s_2 = v_0 \in S_{m+1}^-$. Let us set $S_0 := \{\langle v_{m-1}, v_m\rangle, v_m, \langle t_2, t_2\rangle\}$, $S_i := \{\langle v_{m-i-1}, v_{m-i}\rangle, v_{m-i}\}$ for all $0 < i \le m$, and $S_{m+1} := S \setminus \cup_{0 \le i \le m} S_i$. Since π_2 is simple, $(S_i)_{0 \le i \le m+1}$ is a partition of V. Furthermore, we have that $s_2 = v_0 \in S_{m+1}^-$, and $u_0 \langle u_0, u_1\rangle \dots \langle u_{n-1}, u_n\rangle u_n \subseteq S_{m+1}$ since π_1 and π_2 are vertex disjoint. Thus, it only remains for us to argue that for all $0 \le i \le m+1$: for all $w \in S_i \cap S_N$ we have $wR \cap S_i^+ = \varnothing$, and for all $w \in S_i \cap V_N$ we have $wR \cap S_i^+ \ne \varnothing \implies wR \cap S_i^- \ne \varnothing$. By construction of the S_i, we have that $eR \subseteq S_i$ for all $0 \le i \le m$ and all $e \in S_i \cap S_P$. Furthermore, for all $0 < i \le m$, for all $x \in S_i \cap S_N = \{v_{m-i}\}$, there exists $y \in S_{i-1} \cap S_P = \{\langle v_{m-i}, v_{m-i+1}\rangle\}$ such that $(x, y) \in R$—induced by $(v_{m-i}, v_{m-1+1}) \in E$ from π_2. To conclude, we observe that since $S_0 = \{\langle v_{m-1}, v_m\rangle, v_m = t_2, \langle t_2, t_2\rangle\}$ and $\{t_2, \langle t_2, t_2\rangle\}$ is trapping in $\mathcal{A}$, the set $t_2 R$ is contained in S_0. □

6 Efficiently Under-Approximating the NWR

Although the full NWR cannot be efficiently computed for a given MDP, we can hope for "under-approximations" that are accurate and efficiently computable.

Definition 8 (Under-approximation of the NWR). *Let $\mathcal{A} = (V, V_P, E, T)$ be a target arena and consider a relation $\preceq : V \times \mathcal{P}(V)$. The relation $\preceq$ is an* under-approximation *of the NWR if and only if $\preceq \subseteq \trianglelefteq$.*

We denote by $\preceq^*$ the *pseudo transitive closure* of $\preceq$. That is, $\preceq^*$ is the smallest relation such that $\preceq \subseteq \preceq^*$ and for all $u \in V, X \subseteq V$ if there exists $W \subseteq V$ such that $u \preceq^* W$ and $w \preceq^* X$ for all $w \in W$, then $u \preceq^* X$.

Remark 1. The empty set is an under-approximation of the NWR. For all under-approximations $\preceq$ of the NWR, the pseudo transitive closure $\preceq^*$ of $\preceq$ is also an under-approximation of the NWR.

In [2], efficiently-decidable sufficient conditions for the NWR were given. In particular, those conditions suffice to infer relations such as those in the right MDP from Fig. 3. We recall (Proposition 1) and extend (Proposition 2) these conditions below.

Proposition 1 (From [2]). *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$ and an under-approximation $\preceq$ of the NWR. For all vertices $v_0 \in V$, and sets $W \subseteq V$ the following hold.*

(i) If there exists $S \subseteq \{s \in V \mid s \preceq W\}$ such that there exists no path $v_0 \dots v_n \in (V \setminus S)^* T$, then $v_0 \trianglelefteq W$.
(ii) If $W = \{w\}$ and there exists $S \subseteq \{s \in V_P \mid w \preceq \{s\}\}$ such that $\mathbf{Win}^{v_0}_{S \cup T} \neq \varnothing$, then $w \trianglelefteq \{v_0\}$.

Proof (Sketch). The main idea of the proof of item (i) is to note that S is visited before T. The desired result then follows from Lemma 1. For item (ii), we intuitively have that there is a strategy to visit T with some probability or visit W, where the chances of visiting T are worse than before. We then show that it is never worse to start from v_0 to have better odds of visiting T. □

The above "rules" give an iterative algorithm to obtain increasingly better under-approximations of the NWR: from $\preceq_i$ apply the rules and obtain a new under-approximation $\preceq_{i+1}$ by adding the new pairs and taking the pseudo transitive closure; then repeat until convergence. Using the special cases from Sect. 3.2 we can obtain a nontrivial initial under-approximation $\preceq_0$ of the NWR in polynomial time.

The main problem is how to avoid testing all subsets $W \subseteq V$ in every iteration. One natural way to ensure we do not consider all subsets of vertices in every iteration is to apply the rules from Proposition 1 only on the successors of Protagonist vertices.

In the same spirit of the iterative algorithm described above, we now give two new rules to infer NWR pairs.

Proposition 2. *Consider a target arena $\mathcal{A} = (V, V_P, E, T)$ and $\preceq$ an under-approximation of the NWR.*

(i) For all $u \in V_N$, if for all $v, w \in uE$ we have $v \preceq \{w\}$ and $w \preceq \{v\}$, then $u \sim x$ for all $x \in uE$.
(i) For all $u, v \in V_P \setminus T$, if for all $w \in uE$ such that $w \preceq (uE \setminus \{w\})$ does not hold we have that $w \preceq vE$, then $u \trianglelefteq \{v\}$.

Proof (Sketch). Item (i) follows immediately from the definition of Val. For item (ii) one can use the Bellman optimality equations for infinite-horizon reachability in MDPs to show that since the successors of v are never worse than the non-dominated successors of u, we must have $u \trianglelefteq \{v\}$. □

Fig. 4. Two target arenas with $T = \{fin\}$ are shown. Using Propositions 1 and 2 one can conclude that $p \sim q$ in both target arenas.

The rules stated in Proposition 2 can be used to infer relations like those depicted in Fig. 4 and are clearly seen to be computable in polynomial time as they speak only of successors of vertices.

7 Conclusions

We have shown that the never-worse relation is, unfortunately, not computable in polynomial time. On the bright side, we have extended the iterative polynomial-time algorithm from [2] to under-approximate the relation. In that paper, a prototype implementation of the algorithm was used to empirically show that interesting MDPs (from the set of benchmarks included in PRISM [17]) can be drastically reduced.

As future work, we believe it would be interesting to implement an exact algorithm to compute the NWR using SMT solvers. Symbolic implementations of the iterative algorithms should also be tested in practice. In a more theoretical direction, we observe that the planning community has also studied maximizing the probability of reaching a target set of states under the name of MAXPROB (see, e.g., [16,21]). There, online approximations of the NWR would make more sense than the under-approximation we have proposed here. Finally, one could define a notion of never-worse for finite-horizon or quantitative objectives.

Acknowledgements. The research leading to these results was supported by the ERC Starting grant 279499: inVEST. Guillermo A. Pérez is an F.R.S.-FNRS Aspirant and FWA postdoc fellow.

We thank Nathanaël Fijalkow for pointing out the relation between this work and the study of interval MDPs and numberless MDPs. We also thank Shaull Almagor, Michaël Cadilhac, Filip Mazowiecki, and Jean-François Raskin for useful comments on earlier drafts of this paper.

References

1. Baier, C., Katoen, J.-P.: Principles of Model Checking. MIT Press, New York (2008)
2. Bharadwaj, S., Le Roux, S., Pérez, G.A., Topcu, U.: Reduction techniques for model checking and learning in MDPs. In: IJCAI, pp. 4273–4279 (2017)
3. Brázdil, T., Chatterjee, K., Chmelík, M., Forejt, V., Křetínský, J., Kwiatkowska, M., Parker, D., Ujma, M.: Verification of markov decision processes using learning algorithms. In: Cassez, F., Raskin, J.-F. (eds.) ATVA 2014. LNCS, vol. 8837, pp. 98–114. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11936-6_8
4. Chatterjee, K., Henzinger, M.: Faster and dynamic algorithms for maximal end-component decomposition and related graph problems in probabilistic verification. In: SODA, pp. 1318–1336. SIAM (2011)
5. Ciesinski, F., Baier, C., Größer, M., Klein, J.: Reduction techniques for model checking Markov decision processes. In: QEST, pp. 45–54 (2008)
6. Courcoubetis, C., Yannakakis, M.: The complexity of probabilistic verification. J. ACM **42**(4), 857–907 (1995)
7. D'Argenio, P.R., Jeannet, B., Jensen, H.E., Larsen, K.G.: Reachability analysis of probabilistic systems by successive refinements. In: de Alfaro, L., Gilmore, S. (eds.) PAPM-PROBMIV 2001. LNCS, vol. 2165, pp. 39–56. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44804-7_3
8. De Alfaro, L.: Formal verification of probabilistic systems. Ph.D. thesis, Stanford University (1997)
9. Dehnert, C., Junges, S., Katoen, J.-P., Volk, M.: A Storm is coming: a modern probabilistic model checker. In: Majumdar, R., Kunčak, V. (eds.) CAV 2017, Part II. LNCS, vol. 10427, pp. 592–600. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63390-9_31
10. Eilam-Tzoreff, T.: The disjoint shortest paths problem. Discret. Appl. Math. **85**(2), 113–138 (1998)
11. Fijalkow, N., Gimbert, H., Horn, F., Oualhadj, Y.: Two recursively inseparable problems for probabilistic automata. In: Csuhaj-Varjú, E., Dietzfelbinger, M., Ésik, Z. (eds.) MFCS 2014, Part I. LNCS, vol. 8634, pp. 267–278. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44522-8_23
12. Fortune, S., Hopcroft, J.E., Wyllie, J.: The directed subgraph homeomorphism problem. Theor. Comput. Sci. **10**, 111–121 (1980)
13. Fu, J., Topcu, U.: Probably approximately correct MDP learning and control with temporal logic constraints. In: RSS (2014)
14. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: a survey. JAIR **4**, 237–285 (1996)
15. Kawaguchi, K.: Bounded optimal exploration in MDP. In AAAI, pp. 1758–1764 (2016)
16. Kolobov, A., Mausam, M., Weld, D.S., Geffner, H.: Heuristic search for generalized stochastic shortest path MDPs. In: Bacchus, F., Domshlak, C., Edelkamp, S., Helmert, M. (eds.) ICAPS. AAAI (2011)
17. Kwiatkowska, M., Norman, G., Parker, D.: PRISM 4.0: verification of probabilistic real-time systems. In: Gopalakrishnan, G., Qadeer, S. (eds.) CAV 2011. LNCS, vol. 6806, pp. 585–591. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22110-1_47
18. Puterman, M.L.: Markov Decision Processes. Wiley-Interscience, Hoboken (2005)

19. Russell, S.J., Dewey, D., Tegmark, M.: Research priorities for robust and beneficial artificial intelligence. AI Mag. **36**(4), 105–114 (2015)
20. Russell, S.J., Norvig, P.: Artificial Intelligence - A Modern Approach, 3rd Int. edn., Pearson Education, London (2010)
21. Steinmetz, M., Hoffmann, J., Buffet, O.: Goal probability analysis in probabilistic planning: exploring and enhancing the state of the art. JAIR **57**, 229–271 (2016)
22. Strehl, A.L., Li, L., Littman, M.L.: Reinforcement learning in finite MDPs: PAC analysis. J. Mach. Learn. Res. **10**, 2413–2444 (2009)
23. Valiant, L.: Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. Basic Books, New York (2013)