Information and Computation 301 (2024) 105214

Contents lists available at ScienceDirect

Information and Computation

journal homepage: www.elsevier.com/locate/yinco

Entropic risk for turn-based stochastic games

Christel Baier[a], Krishnendu Chatterjee[b], Tobias Meggendorfer[c], Jakob Piribauer[a,*]

[a] Technische Universität Dresden, Germany
[b] Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria
[c] Lancaster University Leipzig (LUL), Germany

ARTICLE INFO

Article history:
Received 31 December 2023
Accepted 11 August 2024
Available online 14 August 2024

ABSTRACT

Entropic risk (ERisk) is an established risk measure in finance, quantifying risk by an exponential re-weighting of rewards. We study ERisk for the first time in the context of turn-based stochastic games with the total reward objective. This gives rise to an objective function that demands the control of systems in a risk-averse manner. We show that the resulting games are determined and, in particular, admit optimal memoryless deterministic strategies. This contrasts risk measures that previously have been considered in the special case of Markov decision processes and that require randomization and/or memory. We provide several results on the decidability and the computational complexity of the threshold problem, i.e. whether the optimal value of ERisk exceeds a given threshold. Furthermore, an approximation algorithm for the optimal value of ERisk is provided.

© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Funding This work was partly funded by the ERC CoG 863818 (ForM-SMArt), the DFG Grant 389792660 as part of TRR 248 (Foundations of Perspicuous Software Systems), the Cluster of Excellence EXC 2050/1 (CeTI, project ID 390696704, as part of Germany's Excellence Strategy), and the DFG projects BA-1679/11-1 and BA-1679/12-1.

Related version This paper extends the conference paper [1].

1. Introduction

Modern hardware and software systems have reached a level of complexity that makes it nigh impossible for humans to develop and analyze such systems without tools tailored for this task. Due to interactions with an unknown environment, concurrency, the possible failure of components, etc., such systems might exhibit *non-deterministic* as well as *probabilistic* behavior. In the area of formal verification and synthesis, such systems are modeled mathematically in order to solve analysis and synthesis tasks algorithmically on the mathematical model. A fundamental stochastic model are Markov decision processes (MDPs) [2], which extend purely stochastic Markov chains (MCs) with non-determinism to represent an agent interacting with a stochastic environment. Stochastic games (SGs) [3–5] in turn generalize MDPs by introducing an adversary, modeling the case where two agents (or, equivalently, two antagonistic coalitions of agents) engage in adversarial interaction in the presence of a stochastic environment. SGs are a useful tool to formalize synthesis problems. Notably, SGs can

* Corresponding author.
E-mail addresses: christel.baier@tu-dresden.de (C. Baier), Krishnendu.Chatterjee@ist.ac.at (K. Chatterjee), tobias@meggendorfer.de (T. Meggendorfer), jakob.piribauer@tu-dresden.de (J. Piribauer).

https://doi.org/10.1016/j.ic.2024.105214
0890-5401/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

also be used to conservatively model MDPs where transition probabilities are not known precisely [6,7]. See also [8,9,2] and [10,11] for further applications of MDPs and SGs in computer science. In addition, the formal analysis of stochastic models is ubiquitous across further disciplines of science, such as biology [12], epidemiology [13], and chemistry [14], to name a few.

Strategies and objectives In MDPs and SGs, the recipes to resolve choices are called strategies (or policy, or scheduler). *Qualitative* objectives in an SG require the agent to maximize – by choosing a strategy – the probability that the resulting play satisfies a given property against all possible strategies of the adversary. In contrast, *quantitative* objectives require the optimization of a payoff function against all possible strategies of the adversary. One of the most fundamental quantitative objectives studied in the context of MDPs and SGs is the optimization of total rewards (and the related *stochastic shortest path* problem [15]). Here, every state (or, equivalently, transition) of the stochastic model is assigned a cost or reward. These costs or rewards can be used to model quantitative aspects of a system such as resource consumption, time or utility. For the total reward objective, the payoff of a trajectory is the total sum of rewards appearing along the path. The goal of the agent is then to maximize the *expected value* of the total reward. MDPs and SGs with total reward objectives provide an appropriate model to study a wide range of applications, such as traffic optimization [16], verification of stochastic systems [17,18], or navigation and probabilistic planning [19].

Risk-ignorance of expectation Once both players fixed a strategy, the trajectory in a stochastic game and hence the payoff are determined purely probabilistically. When the goal of the players is to maximize and to minimize the expected value of the payoff, respectively, most aspects of the resulting probability distribution of payoffs are ignored. In particular, it is not taken into account whether bad outcomes for the agent might occur with relatively high probability of how bad the bad outcomes can be: An expectation maximizing agent accepts a one-in-a-million chance of extremely high rewards over a slightly worse, but guaranteed outcome. Such a behavior might be undesirable in a lot of situations. Consider a one-shot lottery where with a chance of 10^{-6} we win $2 \cdot 10^6$ times our stake and otherwise lose everything – a two-times increase in expectation. The optimal strategy w.r.t. expectation would bet all available assets, ending up broke in nearly all outcomes. Expectation of a payoff function as objective leads to ignorance of the agent towards the involved *risk*.

Risk-aware alternatives To address this issue, *risk-aware* objectives create incentives to prefer slightly smaller performance in terms of expectation in exchange for a more "stable" behavior. To this end, several variants have been studied in the verification literature of which we name a few in the sequel. First, the variance-penalized expected payoff [20,21] combines the expected value with a penalty for the variance of the resulting probability distribution. So, an agent is incentivized to achieve a high overall performance in terms of the expected value while simultaneously keeping the probability of outliers far from the expected value low. By varying the penalty factor for the variance, the degree of risk-aversion can be adjusted. Also, explicit trade-offs of the expectation and variance (and related notions) have been studied. Here, the agent is required to maximize the expectation while keeping the variance below a given threshold or to minimize the variance while keeping the expectation above a threshold [22,23]. Furthermore, quantiles have been studied in the context of risk-aversion [18]. Quantiles quantify how bad the worst outcomes are: For a given probability value p, a quantile for the total reward provides the smallest bound b such that p of the outcomes lie below the bound b. In finance, quantiles are also referred to as *value-at-risk*. Finally, the conditional value-at-risk (CVaR) [24,25] refines the value-at-risk by measuring the average of the worst p outcomes, which lie below the value-of-risk.

All of these risk-aware alternatives, however, suffer from the following three drawbacks:

1. The above studies focus on the second moment (variance) along with the first moment (mean), but do not incorporate other moments of the payoff distribution.
2. All approaches are studied only for MDPs; none of them have been extended to SGs.
3. Even in MDPs, the above problems require complicated strategies. For example, trade-offs between expectation and variance require memory and randomization [23,22], while optimizing variance-penalized expected payoffs, quantiles, or the CVaR of the total reward require exponential memory [20,26,27,24].

Entropic risk In this paper, we investigate the notion of *entropic risk* [28] with the goal to overcome these drawbacks. Entropic risk has been widely studied in finance and operation research, see e.g. [29,30]. Informally, instead of weighing each outcome uniformly and then aggregating it (as in the case for regular expectation), entropic risk re-weighs outcomes by an exponential function, then computes the expectation, and finally re-normalizes the value. Formally, for a random variable X, the entropic risk measure with parameter $\lambda > 0$ is defined as

$$\mathrm{ERisk}_\gamma(X) := -\frac{1}{\gamma}\log_b(\mathbb{E}[b^{-\gamma X}])$$

where b is a fixed basis such as 2 or e. We illustrate this in the following example.

Example 1.1. Consider a random variable X that takes values $x_1 = 1$, $x_2 = 2$, $x_3 = 4$, and $x_4 = 5$ with probability 1/4 each. Fig. 1 illustrates how the entropic risk measure of X with base e is obtained for some risk aversion factor γ: The values x_i

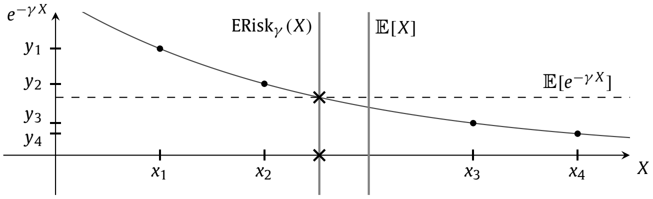

Fig. 1. Illustration of the entropic risk measure. The random variable X takes values x_1 to x_4 uniformly with probability $\frac{1}{4}$ each. Expectation considers the average of x_i, while entropic risk yields the (normalized logarithm of the) average of $y_i = e^{-\gamma x_i}$.

Table 1
Overview of the decidability and complexity results for SGs, MDPs and MCs.

<table>
<tr><td></td><td colspan="3">threshold problem</td><td colspan="2">optimal value</td></tr>
<tr><td></td><td>general instances (Theorem 4.12)</td><td>algebraic instances (Theorem 5.2)</td><td>small algebraic instances (Theorem 5.8)</td><td>computation for small algebraic instances (Theorem 5.12)</td><td>approximation with small rewards and risk aversion factor (Theorem 6.1)</td></tr>
<tr><td>SGs</td><td rowspan="3">decidable subject to Shanuel's conjecture</td><td rowspan="3">in PSPACE (in ∃ℝ)</td><td>in NP ∩ coNP</td><td rowspan="2">in polynomial space</td><td>in polynomial space</td></tr>
<tr><td>MDPs</td><td rowspan="2">in PTIME</td><td rowspan="2">in polynomial time</td></tr>
<tr><td>MCs</td><td>in polynomial time</td></tr>
</table>

are depicted on the x-axis. We now map the values x_i to values $y_i = e^{-\gamma x_i}$ on the y-axis. Then, the expected value of $e^{-\gamma X}$ can be obtained as the arithmetic mean of the values y_i. The result is mapped back to the x-axis via $y \mapsto -\frac{1}{\gamma}\log(y)$, the inverse of $x \mapsto e^{-\gamma x}$, and we obtain $\mathrm{ERisk}_\gamma(X)$.

The entropic risk always takes at most the expected value of a random variable. Due to the exponential reweighing, however, low outcomes have a stronger influence on the entropic risk than high outcomes. So, when trying to optimize the entropic risk of some quantity in a stochastic environment, one has to achieve a relatively high expected value while keeping the probability of low outcomes small. Hence, we can think of the entropic risk as a risk-adjusted performance measure.

Advantages Aside from satisfying many desirable properties of risk measures established in finance, entropic risk brings several crucial advantages in our specific setting, of which we list a few: In contrast to variance minimization, it is beneficial to increase the probability of extremely good outcomes (which would increase variance). Moreover, the entropic risk incorporates *all* moments of the distribution. In particular, even if the expectation is infinite, entropic risk still provides meaningful values (opposed to both expectation and variance). Note that the expected total reward objective is often addressed under additional assumptions excluding the case of infinite expected values [31,17]. Additionally, entropic risk is a *time-consistent* risk measure. In our situation, this means that the risk evaluation at a state is the same for *any history*. This is in stark contrast to, e.g., quantile and CVaR optimal strategies, which after a series of unfortunate events start behaving recklessly (e.g., expectation optimal). Due to these advantages, ERisk has already been studied in the context of MDPs [32,33]. However, to the best of our knowledge, neither the arising computational problems nor the more general setting of SGs have been addressed.

1.1. Our results

In this work we consider the notion of entropic risk in the context of SGs as well as the special cases of MCs and MDPs. For an overview of our complexity results, see Table 1.

1. *Determinacy and Strategy Complexity.* We establish several basic results, in particular that SGs with the entropic risk objective are determined and that pure memoryless optimal strategies exist for both players. This stands in contrast to other notions of risk, where even in MDPs strategies require memory and/or randomization.
2. *Exact Computation.* When allowing Euler's number e as the basis of exponentiation, the threshold problem, i.e. asking whether the optimal entropic risk lies above a given bound, is decidable subject to Shanuel's conjecture. If the basis of exponentiation and all other numbers in the input are rational, then all numbers resulting from the involved exponentiation are shown to be algebraic. We obtain a reduction to the existential theory of the reals and thus a PSPACE upper bound in this case.
   Furthermore, we identify a notion of *small algebraic instance* in which all occurring numbers are not only algebraic, but have a small representation and are contained in algebraic extensions of $\mathbb{Q}$ of low degree. The threshold problem for small algebraic instances of MCs and MDPs can efficiently be solved by explicit computations in an algebraic extension of $\mathbb{Q}$. We obtain polynomial-time algorithms for MCs and MDPs, and conclude that the threshold problem lies in NP

∩ co-NP for SGs in this case. For small algebraic instances, we furthermore show that an explicit closed form of the optimal value can be computed (a) in polynomial time for MCs; and consequently (b) in polynomial space for SGs.
3. *Approximate Computation.* We provide an effective way to compute an approximation, i.e. determine the optimal entropic risk up to a given precision of $\varepsilon > 0$. To this end, we show that in the general case, by considering enough bits of arising irrational numbers, we can bound the incurred error. In MDPs and MCs, the optimal value can be approximated in time polynomial in the size of the model, in $-\log(\varepsilon)$, and in the magnitude of the rewards. For SGs, this implies the existence of a polynomial-space approximation algorithm.

1.2. Related work

The entropic risk objective has been studied before in MDPs: An early formulation can be found in [32] under the name *risk-sensitive MDPs* focusing on the finite-horizon setting. The paper [34] considers an exponential utility function applied to discounted rewards and optimal strategies are shown to exist, but not to be memoryless in general. In [35], the entropic risk objective is considered for MDPs with a general Borel state space and in [33] a generalization of this objective is studied on such MDPs. To the best of our knowledge, however, all previous work in the context of MDPs focuses on optimality equations and general convergence results of value iteration, while the resulting algorithmic problems for finite-state MDPs have not been investigated. Furthermore, we are not aware of work on the entropic risk objective in SGs.

For other objectives capturing risk-aversion, algorithmic problems have been analyzed on finite-state MDPs: Variance-penalized expectation has been studied for finite-horizon MDPs with terminal rewards in [36] and for infinite-horizon MDPs with discounted rewards and mean payoffs [21], and total rewards [20]. For total rewards, optimal strategies require exponential memory and the threshold problem is in NEXPTIME and EXPTIME-hard [20].

In [22], the optimization of expected accumulated rewards under constraints on the variance is studied for finite-horizon MDPs. Possible tradeoffs between expected value and variance of mean payoffs and other notions of variability have been studied in [23].

To control the chance of bad outcomes, the problem to maximize or minimize the probability that the accumulated weight lies below a given bound w has been addressed in MDPs [26,37]. Similarly, quantile queries ask for the minimal weight w such that the weight of a path stays below w with probability at least p for the given value p under some or all schedulers [38,39]. Both of these problems have been addressed for MDPs with non-negative weights and are solvable in exponential time in this setting [38,26]. Optimal strategies require exponential memory and the decision version of these problems is PSPACE-hard [26].

The conditional value-at-risk (CVaR), a prominent risk-measure, has been investigated for mean payoff and weighted reachability in MDPs in [25] as well as for total rewards in MDPs [27,24]. The optimal CVaR of the total reward in MDPs with non-negative weights can be computed in exponential time and optimal strategies require exponential memory [27,24]. The threshold problem for optimal CVaR of total reward in MDPs with integer weights is at least as hard as the Positivity-problem for linear recurrence sequences, a well-known problem in analytic number theory whose decidability status is, since many decades, open [27].

For all these objectives capturing risk-aversion in some sense, we are not aware of any work addressing the resulting algorithmic problems on SGs.

Finally, we want to mention that there are connections of stochastic games with the entropic risk objective applied to total rewards and entropy games [40–42]. An entropy game is played on a graph where each node is controlled by one of two players. The outgoing edges from each node are labeled with a finite set of action labels. Among the transitions with the chosen label, a non-deterministic half-player “People” chooses the transition that is taken. The goal of one player is to maximize the growth rate of the number of plays of length n that People can choose as n tends to ∞, while the other player tries to minimize this growth rate. Entropy games can be seen as a special case of matrix multiplication games where two players take turns in choosing matrices from a finite set of matrices while one of the players tries to maximize – and the other player to minimize – the growth rate of the matrix product. As we will see, in stochastic games with entropic risk applied to the total reward, the auxiliary notion of *negative exponential utility* (see Section 3.2) also behaves multiplicatively along a play. However, the objective we are interested is applied to total rewards. The growth rate as objective in entropic games, in contrast, can be seen as a mean-payoff-like objective. Furthermore, the stochastic games with entropic risk as objective and entropy games are given in quite a different format. While a presentation in terms of matrix multiplications is possible in both cases, the resulting matrices are of different types (integer matrices in the case of entropy games and matrices with algebraic or transcendental entries in our case).

1.3. Outline

In Section 2, we introduce our notations for stochastic games, some commonly used objectives, and a brief overview of notions for algebraic field extensions we need later on. Section 3 formally introduces the notion of entropic risk in the setting of stochastic games. In Section 4, we establish general results on entropic risk in stochastic games. In particular, we show that the games are determined and that both players have optimal memoryless deterministic strategies. Section 5 addresses the case of rational inputs, which allows to treat the entropic risk in stochastic games via computations over algebraic numbers. Finally, Section 6 provides a method to approximate the value of a stochastic game with the entropic risk objective. We end with concluding remarks in Section 7.

2. Preliminaries

In this section, we recall the basics of (turn-based) SGs and relevant objectives. For further details, see, e.g., [2,43,17,11]. We assume familiarity with basic notions of probability theory (see, e.g., [44]). We write $\mathcal{D}(X)$ to denote the set of all *probability distributions* over a countable set X, i.e. mappings $d : X \to [0, 1]$ such that $\sum_{x \in X} d(x) = 1$. The support of a distribution d is $\mathrm{supp}(d) := \{x \in X \mid d(x) > 0\}$. For a set S, $S^\star$ and S^ω refer to the set of finite and infinite sequences of elements of S, respectively.

2.1. Markov chains, MDPs, and stochastic games

A *Markov chain (MC)* (e.g. [43]), is a tuple $\mathsf{M} = (S, \delta)$, where S is a set of *states*, and $\delta : S \to \mathcal{D}(S)$ is a *transition function* that for each state s yields a probability distribution over successor states. We write $\delta(s, s')$ instead of $\delta(s)(s')$ for the probability to move from s to s' for $s, s' \in S$. A *(infinite) path* in an MC is an infinite sequence $s_0, s_1, \ldots$ of states such that for all i, we have $\delta(s_i, s_{i+1}) > 0$. We denote the set of infinite paths by $\mathsf{Paths}_\mathsf{M}$. Together with a state s, an MC M induces a unique probability distribution $\mathsf{Pr}_{\mathsf{M},s}$ over the set of all infinite paths $\mathsf{Paths}_\mathsf{M}$ starting in s. For a random variable $f : \mathsf{Paths}_\mathsf{M} \to \mathbb{R}$, we write $\mathbb{E}_{\mathsf{M},s}(f)$ for the expected value of f under the probability measure $\mathsf{Pr}_{\mathsf{M},s}$.

A *turn-based stochastic game (SG)* (e.g. [4]) is a tuple $(S_{\max}, S_{\min}, A, \Delta)$, where $S_{\max}$ and $S_{\min}$ are finite, disjoint sets of *Maximizer* and *Minimizer* states, inducing the set of states $S = S_{\max} \cup S_{\min}$, A denotes a finite set of *actions*, furthermore overloading A to also act as a function assigning to each state s a set of non-empty *available actions* $A(s) \subseteq A$, and $\Delta : S \times A \to \mathcal{D}(S)$ is the *transition function* that for each state s and (available) action $a \in A(s)$ yields a distribution over successor states. For convenience, we write $\Delta(s, a, s')$ instead of $\Delta(s, a)(s')$. Moreover, $\mathrm{opt}^s_{a \in A(s)}$ refers to $\max_{a \in A(s)}$ if $s \in S_{\max}$ and $\min_{a \in A(s)}$ if $s \in S_{\min}$, i.e. the preference of either player in a state s. We omit the superscript s where clear from context. Given a function $f : S \to \mathbb{R}$ assigning values to states, we write $\Delta(s, a)\langle f \rangle := \sum_{s' \in S} \Delta(s, a, s') \cdot f(s')$ for the weighted sum over the successors of s under $a \in A(s)$. A *Markov decision process (MDP)* (e.g. [2]) can be seen as an SG with only one player, i.e. $S_{\max} = \emptyset$ or $S_{\min} = \emptyset$.

The semantics of SGs is given in terms of resolving choices by strategies inducing an MC with the respective probability space over infinite paths. Intuitively, a stochastic game is played in turns: In every state s, the player to whom it belongs chooses an action a from the set of available actions $A(s)$ and the play advances to a successor state s' according to the probability distribution given by $\Delta(s, a)$. Starting in a state s_0 and repeating this process indefinitely yields an infinite sequence $\rho = s_0 a_0 s_1 a_1 \cdots \in (S \times A)^\omega$ such that for every $i \in \mathbb{N}_0$ we have $a_i \in A(s_i)$ and $\Delta(s_i, a_i, s_{i+1}) > 0$. We refer to such sequences as *(infinite) paths* or *plays* and denote the set of all infinite paths in a given game G by $\mathsf{Paths}_\mathsf{G}$. Furthermore, we write ρ_i to denote the i-th state in the path ρ. *Finite paths* or *histories* $\mathsf{FPaths}_\mathsf{G}$ are finite prefixes of a play, i.e. elements of $(S \times A)^\star \times S$ consistent with A and Δ.

The decision-making of the players is captured by the notion of *strategies*. Strategies are functions mapping a given history to a distribution over the actions available in the current state. For this paper, *memoryless deterministic* strategies (abbreviated *MD strategies*, also called positional strategies) are of particular interest. These strategies choose a single action in each state, irrespective of the history, and can be identified with functions $\sigma : S \to A$. Since we show that these strategies are sufficient for the discussed notions, we define the semantics of games only for these strategies and refer the interested reader to the mentioned literature for further details. We write Π_G for the set of all strategies and $\Pi^{\mathsf{MD}}_\mathsf{G}$ for memoryless deterministic ones. We call a pair of strategies a *strategy profile*, written $\pi = (\sigma, \tau)$. We identify a profile with the induced joint strategy $\pi(s) := \sigma(s)$ if $s \in S_{\max}$ and $\tau(s)$ otherwise.

Given a profile $\pi = (\sigma, \tau)$ of MD strategies for a game G, we write G^π for the MC obtained by fixing both strategies. So, $\mathsf{G}^\pi = (S, \hat{\delta})$, where $\hat{\delta}(s) := \Delta(s, \pi(s))$. Together with a state s, the MC G^π induces a unique probability distribution $\mathsf{Pr}^\pi_{\mathsf{G},s}$ over the set of all infinite paths $\mathsf{Paths}_\mathsf{G}$. For a random variable over paths $f : \mathsf{Paths}_\mathsf{G} \to \mathbb{R}$, we write $\mathbb{E}^\pi_{\mathsf{G},s}[f]$ for the expected value of f under the probability measure $\mathsf{Pr}^\pi_{\mathsf{G},s}$.

2.2. Objectives

Usually, we are interested in finding strategies that optimize the value obtained for a particular *objective*. We introduce some objectives of interest.

Reachability A reachability objective is specified by a set of *target states* $T \subseteq S$. We define $\Diamond T = \{\rho \mid \exists i. \rho_i \in T\}$ the set of all paths eventually reaching a target state. Given a strategy profile π and a state s, the probability for this event is given by $\mathsf{Pr}^\pi_{\mathsf{G},s}[\Diamond T]$. On games, we are interested in determining the *value* $\mathsf{Val}_{\mathsf{G},\Diamond T}(s) := \max_{\sigma \in \Pi^{\mathsf{MD}}_\mathsf{G}} \min_{\tau \in \Pi^{\mathsf{MD}}_\mathsf{G}} \mathsf{Pr}^{\sigma,\tau}_{\mathsf{G},s}[\Diamond T]$ of a state s, which intuitively is the best probability we can ensure against an optimal opponent. Generally, one would consider supremum and infimum over strategies instead maximum and minimum over MD strategies. However, for reachability we know that these values coincide and the game is *determined*, i.e. the order of max and min does not matter [5]. Finally, we know that the value $\mathsf{Val}_{\mathsf{G},\Diamond T}$ is a solution of the following set of equations

$$v(s) = 0 \text{ for } s \in S_0, \quad v(s) = 1 \text{ for } s \in T,$$
$$\text{and } v(s) = \text{opt}_{a \in A(s)} \Delta(s,a)\langle v \rangle \text{ otherwise,} \tag{1}$$

where S_0 is the set of states that cannot reach T against an optimal Minimizer strategy [45].

Total reward The total reward objective is specified by a reward function $r : S \to \mathbb{R}_{\geq 0}$, assigning non-negative rewards to every state. The total reward obtained by a particular path is defined as the sum of all rewards seen along this path, $\mathsf{TR}(\rho) := \sum_{i=1}^{\infty} r(\rho_i)$. Note that since we assume $r(s) \geq 0$, this sum is always well-defined. Classically, we want to optimize the expected total reward, i.e. determine $\mathsf{Val}_{\mathsf{G},\mathbb{E}\,\mathsf{TR}}(s) := \max_{\sigma \in \Pi_{\mathsf{G}}^{\mathsf{MD}}} \min_{\tau \in \Pi_{\mathsf{G}}^{\mathsf{MD}}} \mathbb{E}_{\mathsf{G},s}^{\sigma,\tau}[\mathsf{TR}]$. This game is determined and MD strategies suffice [46]. (To be precise, that work considers a more general formulation of total reward, our case is equivalent to the case $\star = c$ and $T = \emptyset$ (Def. 3) and the quantitative rPATL formula $\langle\langle\{1\}\rangle\rangle \mathbf{R}^r_{\max=?}[\mathbf{F}^c\,\mathtt{ff}]$.)

2.3. Vector spaces and field extensions

We assume some familiarity with the algebraic field extensions. We introduce the main concepts briefly and provide our notation. For more details, see, e.g., [47].

Vectors For a vector $v \in \mathbb{K}^n$ of a field $\mathbb{K}$, we denote its components by $v_0, \ldots, v_{n-1}$. Whenever comparing two vectors x and y over an ordered field by $x \leq y$ or taking their max or min, this is understood point-wise. We often encounter functions assigning values to (finitely many) states; for convenience, we assume an implicit (arbitrary but fixed) numbering of each set of states and identify such functions with the corresponding vectors.

Field extensions Given a complex number α, the smallest field extension $\mathbb{F}$ of $\mathbb{Q}$ containing α, i.e. with $\alpha \in \mathbb{F}$, is denoted by $\mathbb{Q}(\alpha)$. The number α is called algebraic if there is a non-constant rational polynomial $P \in \mathbb{Q}[X]$ with $P(\alpha) = 0$. A complex number that is not algebraic is called transcendental. The degree of an algebraic number α is the smallest degree of a non-zero polynomial P with $P(\alpha) = 0$. There is a unique non-zero polynomial P with $P(\alpha) = 0$ of minimal degree with leading coefficient 1, which is called the minimal polynomial of α. For an algebraic number α of degree q, $\mathbb{Q}(\alpha)$ is a $\mathbb{Q}$-vector space of dimension q. One basis is given by $(\alpha^0, \ldots, \alpha^{q-1})$. For any basis $B = (b_0, \ldots, b_{q-1})$ of $\mathbb{Q}(\alpha)$ as $\mathbb{Q}$-vector space and any number $\beta \in \mathbb{Q}(\alpha)$, there is a unique vector $v \in \mathbb{Q}^q$ such that $\beta = \sum_{i=0}^{q-1} v_i \cdot b_i$. We call v the representation of β in basis B. Vice versa, given a vector $v \in \mathbb{Q}^q$, we denote the number $\sum_{i=0}^{q-1} v_i \cdot b_i$ it represents in basis B by $[v]_B$.

3. Entropic risk

As hinted in the introduction, for classical total reward we optimize the expectation and disregard other properties of the actual distribution of obtained rewards. This means that an optimal strategy may accept arbitrary risks if they yield minimal improvements in terms of expectation. To overcome this downside, we consider the entropic risk:

Definition 3.1. Let $b > 1$ a basis, X a random variable, and $\gamma > 0$ a risk aversion factor. The *entropic risk (of X with base b and factor γ)* (see, e.g., [48]) is defined as

$$\mathsf{ERisk}_\gamma(X) := -\tfrac{1}{\gamma} \log_b(\mathbb{E}[b^{-\gamma X}]).$$

One often chooses $b = e$. Nevertheless, we also consider rational values for b, which allows us to apply techniques from algebraic number theory to arising computational problems.

The definition ensures that deviations to lower values are penalized, i.e. taken into consideration more strongly, by this risk measure (see also Example 1.1). For a different perspective, we can also consider the Taylor expansion of ERisk w.r.t. γ, which is $\mathsf{ERisk}_\gamma(X) = \mathbb{E}[X] - \frac{\gamma}{2} \cdot \mathsf{Var}[X] + \mathcal{O}(\gamma^2)$ (see, e.g., [49]). The terms hidden in $\mathcal{O}(\gamma^2)$ comprise all moments of X and exhibit an asymmetry such that ERisk is roughly the expected value minus a penalty for deviations to lower values.

3.1. Entropic risk in SGs

We are interested in the case $X = \mathsf{TR}$, i.e. optimizing the risk for total rewards. We write

$$\mathsf{ERisk}^\gamma_{\mathsf{G},\hat{s}}(\pi) := -\tfrac{1}{\gamma} \log_b(\mathbb{E}^\pi_{\mathsf{G},\hat{s}}[b^{-\gamma X}])$$

to denote the entropic risk of the total reward achieved by the strategy profile π when starting in state $\hat{s}$, omitting sub- and superscripts where clear from context. Clearly, this is well defined for any profile: We have that $b^{-\gamma\,\mathsf{TR}(\rho)} = b^{-\gamma \sum_{i=1}^{\infty} r(\rho_i)} = \prod_{i=1}^{\infty} b^{-\gamma r(\rho_i)}$ and each factor lies between 0 and 1, thus the product converges (possibly with limit 0).

We also give an insightful characterization for integer rewards. If $r(s) \in \mathbb{N}$, we have

$$\mathsf{ERisk}^\gamma_{\mathsf{G},\hat{s}}(\pi) = -\tfrac{1}{\gamma} \log_b \left(\sum_{n=0}^{\infty} \mathsf{Pr}^\pi_{\mathsf{G},\hat{s}}[\mathsf{TR} = n] \cdot b^{-\gamma n} \right). \tag{2}$$

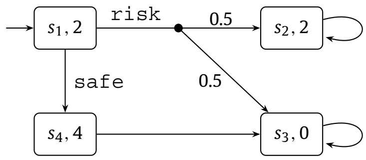

Fig. 2. Our running example to demonstrate several properties of entropic risk. For ease of presentation, the system actually is an MDP, where all states belong to Maximizer. States are denoted by boxes and their reward is written next to the state name. Transition probabilities are written next to the corresponding edges, omitting probability 1.

Naturally, our goal is to optimize the entropic risk. In this work, we mainly consider the corresponding decision variant, which we call the *entropic risk threshold problem*:

Entropic risk threshold problem: Given an SG G, state $\hat{s}$, reward function r, risk parameter γ, risk basis b, and threshold t, decide whether there exists a Maximizer strategy σ such that for all Minimizer strategies τ we have $\mathrm{ERisk}^{\gamma}_{\mathsf{G},\hat{s}}((\sigma,\tau)) \geq t$.

Note that (for now) we do not assume any particular encoding of the input. For example, the reward function r could be given symbolically, describing irrational numbers. A second variant of the threshold problem asks whether the optimal value

$$\mathrm{ERisk}^{\gamma *}_{\mathsf{G},s} := \sup_{\sigma \in \Pi_{\mathsf{G}}} \inf_{\tau \in \Pi_{\mathsf{G}}} \mathrm{ERisk}^{\gamma}_{\mathsf{G},s}((\sigma,\tau)) \tag{3}$$

is at least t for a given threshold t. We will see that SGs with the entropic risk as objective function are determined and hence the two variants are equivalent. Before proceeding with our solution approaches, we provide an illustrative example.

Example 3.2. Consider the MDP of Fig. 2. The optimal total reward is obtained by choosing action `risk` in state s_1: Then, we actually obtain an infinite total reward through state s_2. In comparison, choosing action `safe` would yield a reward of 6 in total. Now, consider the entropic risk. When choosing action `risk`, we obtain a total reward of 2 and ∞ with probability $\frac{1}{2}$ each, while action `safe` yields 6 with probability 1. Let $b = 2$. First, we consider the case $\gamma = 1$ for simplicity. Then, we obtain an entropic risk of $-\log_2(\frac{1}{2}2^{-2} + \frac{1}{2}2^{-\infty}) = 3$ under action `risk` and $-\log_2(2^{-6}) = 6$ for `safe`. Thus, action `safe` is preferable.

Let us now take a look how the risk aversion factor influences the decision. For risk aversion factor $\gamma > 0$, we obtain an entropic risk of $-\frac{1}{\gamma}\log_2(\frac{1}{2}2^{-2\gamma} + \frac{1}{2}2^{-\infty}) = \frac{1}{\gamma}\log_2(2^{2\gamma+1}) = 2 + \frac{1}{\gamma}$ under action `risk` and still $-\frac{1}{\gamma}\log_2(2^{-6\gamma}) = 6$ for `safe`. So, for small risk aversion factors $\gamma < \frac{1}{4}$, the action `risk` is preferred. When increasing the risk aversion factor to $\gamma > \frac{1}{4}$, action `safe` is preferred.

Remark 3.3. As hinted above, entropic risk is finite whenever a finite reward is obtained with non-zero probability, i.e. for any strategy profile π, $\mathrm{ERisk}^{\gamma}_{\mathsf{G},\hat{s}}(\pi) = \infty$ iff $\Pr^{\pi}_{\mathsf{G},\hat{s}}[\mathrm{TR} = \infty] = 1$. In contrast, expectation is infinite whenever there is a non-zero chance of infinite reward, i.e. $\mathbb{E}^{\pi}_{\mathsf{G},\hat{s}}[\mathrm{TR}] = \infty$ iff $\Pr^{\pi}_{\mathsf{G},\hat{s}}[\mathrm{TR} = \infty] > 0$. So, entropic risk allows us to meaningfully compare strategies which yield infinite total reward with some positive probability.

3.2. Exponential utility

Observe that the essential part of the entropic risk is the inner expectation. Thus, we consider the *negative exponential utility*

$$\mathrm{NegUtil}^{\gamma}_{\mathsf{G},\hat{s}}(\pi) := \mathbb{E}^{\pi}_{\mathsf{G},\hat{s}}[b^{-\gamma\,\mathrm{TR}}].$$

We have $\mathrm{ERisk}^{\gamma}_{\mathsf{G},\hat{s}}(\pi) = -\frac{1}{\gamma}\log_b(\mathrm{NegUtil}^{\gamma}_{\mathsf{G},\hat{s}}(\pi))$. Observe that in our case $0 \leq \mathrm{NegUtil}^{\gamma}_{\mathsf{G},\hat{s}}(\pi) \leq 1$ for any π, as $0 \leq \mathrm{TR} \leq \infty$. Moreover, $\mathrm{ERisk}^{\gamma}_{\mathsf{G},\hat{s}}(\pi) \geq t$ iff $\mathrm{NegUtil}^{\gamma}_{\mathsf{G},\hat{s}}(\pi) \leq b^{-\gamma \cdot t}$, thus, a risk-averse agent (in our case Maximizer) wants to minimize NegUtil. The optimal value is

$$\mathrm{NegUtil}^{\gamma *}_{\mathsf{G},s} := \inf_{\sigma \in \Pi_{\mathsf{G}}} \sup_{\tau \in \Pi_{\mathsf{G}}} \mathbb{E}^{\sigma,\tau}_{\mathsf{G},s}[b^{-\gamma\,\mathrm{TR}}]. \tag{4}$$

We again omit sub- and superscripts where clear from context. We show later that games with NegUtil or ERisk as payoff functions are determined. Thus, the order of sup and inf in the above definition does not matter. We call a Maximizer-strategy σ optimal if $\mathrm{ERisk}^{\gamma *}_{\mathsf{G},s} = \inf_{\tau \in \Pi_{\mathsf{G}}} \mathrm{ERisk}^{\gamma}_{\mathsf{G},s}((\sigma,\tau))$ and analogously for Minimizer-strategies.

4. Basic properties and decidability

In this section, we establish several results for SGs with entropic risk as objective functions concerning determinacy, strategy complexity, and decidability in the general case. We mainly work on games with NegUtil as payoff function. As ERisk can be obtained from NegUtil via the monotone function $-\frac{1}{\gamma}\log(\cdot)$, most results, such as determinacy or strategy complexity, will transfer directly to games with ERisk as objective function. However, note that in particular computability results do not immediately transfer since, in general, $\log(a)$ is not a rational number for a rational input a and vice versa.

First, we show that the games are determined, i.e. the order of sup and inf in Equation (3) and Equation (4) can be switched. Then, we show that games with NegUtil as payoff function can be seen as reachability games via a reduction that introduces irrational transition probabilities in general. We conclude that considering only MD strategies is sufficient to obtain the optimal value, i.e. sup and inf can be replaced with a max and min over MD strategies. From this, we derive a system of inequalities that has a solution if and only if the optimal value satisfies $\mathrm{ERisk}^* \geq t$ for a given threshold t. We conclude this section by observing that the satisfiability of this system of inequalities can be expressed as a sentence in the language of the reals with exponentiation. In this way, we obtain the conditional decidability of the entropic risk threshold problem in SGs subject to Shanuel's conjecture.

Throughout this section, fix a game G, reward function r, state $\hat{s}$, risk parameter γ, and risk basis b.

4.1. Determinacy and optimality equation

Lemma 4.1. *Stochastic games with* NegUtil *as payoff function are determined, i.e.*

$$\inf_{\sigma\in\Pi_{\mathsf{G}}}\sup_{\tau\in\Pi_{\mathsf{G}}}\mathbb{E}^{\sigma,\tau}_{\mathsf{G},s}[b^{-\gamma\,\mathrm{TR}}]=\sup_{\tau\in\Pi_{\mathsf{G}}}\inf_{\sigma\in\Pi_{\mathsf{G}}}\mathbb{E}^{\sigma,\tau}_{\mathsf{G},s}[b^{-\gamma\,\mathrm{TR}}].$$

Proof. This follows from the classical result on determinacy of Borel games [50], see [51] for a concrete formulation for stochastic games. In particular, the game is zero-sum and NegUtil is a bounded, Borel-measurable function. □

As ERisk is obtained from NegUtil via a monotone function, also games with ERisk as payoff function are determined. While ERisk^* is difficult to tackle directly due to its non-linearity, we can derive the following optimality equation for $\mathrm{NegUtil}^*$:

Lemma 4.2. *The optimal utility* $\mathrm{NegUtil}^*$ *is a solution of the following system of constraints:*

$$v(s)=b^{-\gamma r(s)}\cdot\overline{\mathrm{opt}}^{s}_{a\in A(s)}\cdot\sum_{s'\in S}\Delta(s,a,s')\cdot v(s'),\tag{5}$$

where $\overline{\mathrm{opt}}^{s}$ *is* min *for a Maximizer state* s *and* max *for a Minimizer state.*

Proof. Fix an arbitrary Maximizer state s (the statement follows analogous for Minimizer states). Let $\pi=(\sigma^*,\tau^*)$ an optimal strategy profile, ensuring a value of $\mathrm{NegUtil}^*$ in s. Observe that

$$\mathrm{NegUtil}^*(s)=\mathbb{E}^{\pi}_{\mathsf{G},s}[b^{-\gamma\,\mathrm{TR}}]=\int\limits_{\rho\in\mathsf{Paths}_{\mathsf{G}}} b^{-\gamma\sum_{i=1}^{\infty}r(\rho_i)}\,d\mathsf{Pr}^{\pi}_{\mathsf{G},s}.$$

We can rewrite $b^{-\gamma\sum_{i=1}^{\infty}r(\rho_i)}=b^{-\gamma r(\rho_1)}\cdot b^{-\gamma\sum_{i=2}^{\infty}\rho_i}$. Inserting this equality in the above equation, splitting the set of paths by the action taken and the successor state, and shifting indices by 1 in the integral, we get

$$\mathbb{E}^{\pi}_{\mathsf{G},s}[b^{-\gamma\,\mathrm{TR}}]=b^{-\gamma r(s)}\sum_{a\in A(s),s'\in S}\Delta(s,a,s')\cdot\mathbb{E}^{\pi_{(s,a)}}_{\mathsf{G},s'}[b^{-\gamma\,\mathrm{TR}}],$$

where $\pi_{(s,a)}$ is the strategy profile both players follow after seeing (s,a). Now, observe that $\mathbb{E}^{\pi_{(s,a)}}_{\mathsf{G},s'}[b^{-\gamma\,\mathrm{TR}}]=\mathrm{NegUtil}^*(s')$ for any s' where $\Delta(s,a,s')>0$: If it were different, the respective player could employ a better strategy in this particular case, ensuring a different value in s. Thus,

$$\mathrm{NegUtil}^*(s)=b^{-\gamma r(s)}\sum_{a\in A(s),s'\in S}\Delta(s,a,s')\cdot\mathrm{NegUtil}^*(s').$$

Now, it immediately follows that taking any optimal action ensures the best outcome, proving the result. □

Unfortunately, NegUtil is not the unique and not even the pointwise smallest or largest fixed point of this equation system. Consider the case where $r\equiv 0$, i.e. $b^{-\gamma r(s)}=1$. Here, every constant vector is a fixed point, however $\mathrm{NegUtil}^*\equiv 1$. More generally, as the equations are purely multiplicative, for any fixed point v, every multiple $\lambda\cdot v$ is a fixed point, too.

Example 4.3. Again consider the example of Fig. 2 with $b = 2$ and $\gamma = 1$. The (simplified) equations we get are:

$$v_1 = 2^{-2} \cdot \min\{\tfrac{1}{2} v_2 + \tfrac{1}{2} v_3, v_4\} \qquad v_2 = 2^{-2} \cdot v_2 \qquad v_3 = v_3 \qquad v_4 = 2^{-4} \cdot v_3,$$

where v_i corresponds to the value of s_i. First, for v_2, we observe that $v_2 = 0$ is the only valid assignment. Then, we have that $v_1 = 2^{-2} \cdot \min\{\frac{1}{2}0 + \frac{1}{2}v_3, 2^{-4}v_3\} = 2^{-3} \cdot \min\{v_3, 2^{-3}v_3\}$. Clearly, this system is underdetermined and we obtain a distinct solution for any value of v_3.

To solve these issues, we need to define "anchors" of the equation. We observe the resemblance of classical fixed point equations for stochastic systems. In particular, for $r \equiv 0$, Equation (5) is the same as for reachability, Equation (1).

4.2. Reduction to reachability

We define

$$S_0 = \{s \mid \max_\sigma \min_\tau \mathsf{Pr}^{\sigma,\tau}_{\mathsf{G},s}[\mathrm{TR} > 0] = 0\} \text{ and}$$
$$S_\infty = \{s \mid \max_\sigma \min_\tau \mathsf{Pr}^{\sigma,\tau}_{\mathsf{G},s}[\mathrm{TR} = \infty] = 1\}$$

the set of states in which Maximizer cannot obtain a total reward of more than 0 with positive probability against an optimal opponent strategy or ensure infinite reward with probability 1, respectively. We show later on that these sets are simple to compute and MD strategies are sufficient. Since $r(s) \geq 0$, all states in $s \in S_0$ necessarily have $r(s) = 0$. Observe that S_0 may be empty, but then $S = S_\infty$ and so $\mathrm{NegUtil}^* = 0$, $\mathrm{ERisk}^* = \infty$. Through these sets, we can connect optimizing the utility to a reachability objective.

Lemma 4.4. *For any state s in the game G, the optimal utility $\mathrm{NegUtil}^*$ is equal to the minimal probability of reaching the set S_0 from s in game G_R, defined as follows: We add a designated sink state $\underline{s}$ (which may belong to either player and only has a self-loop back to itself) and define $\Delta_R(s, a, s') = b^{-\gamma r(s')} \cdot \Delta(s, a, s')$ for $s, s' \in S$, $a \in A(s)$ and $\Delta_R(s, a, \underline{s}) = (1 - b^{-\gamma r(s)})$. There is a direct correspondence between optimal strategies.*

Sketch. We first provide a brief proof sketch. Essentially, we first show that S_0 and S_∞ are "sinks" of the play, i.e. that any optimal strategy can and will keep the play inside either set once it is reached. Moreover, they are the only sinks in that sense, i.e. under any optimal strategy either of them is reached with probability one.

The remainder of the proof then shows that the "discounted" reachability (achieved by the transitions to $\underline{s}$) corresponds to the negative utility and that optimal strategies correspond. From G to G_R, we split all paths that reach S_0 by their prefix, observing that the total reward obtained by the path equals that of the prefix (since once in S_0 no more reward is obtained), and all other paths obtain a reward of ∞ with probability 1, meaning they reach $\underline{s}$ with probability 1, too. From G_R to G, we again split the paths reaching S_0 by their prefix and notice that the probability of those prefixes exactly corresponds to the discounted total reward they would achieve in G. □

Proof. Before we tackle the actual statement, we discuss some important properties of S_0.

Properties of S_0 and S_∞: We first show that S_0 represents a "sink" of the play, in the sense that any optimal Minimizer strategy will keep the play inside S_0. Moreover, this also proves that such a strategy always exists. Observe that by definition there exists a (memoryless deterministic) strategy of Minimizer ensuring that Maximizer can only obtain a total reward of 0 for any state in S_0 and consequently a utility of 1. This already shows that by keeping the play inside S_0, Minimizer acts optimally (there is no better outcome). One can also show that remaining inside S_0 is the only optimal play, however we do not require this statement. We analogously show a similar property for S_∞, i.e. an optimal strategy of the Maximizer can always keep the play in S_∞. We note that these strategies are rather simple to obtain, indeed for any state $s \in S_0$ we can just take any action $a \in A(s)$ with $\mathrm{supp}(\Delta(s, a)) \subseteq S_0$, dually for S_∞. In particular, they are memoryless and deterministic.

Finally, we show that S_0 and S_∞ are the only two sinks, i.e. $\mathsf{Pr}^{\pi^*}_{\mathsf{G},\hat{s}}[\Diamond(S_0 \cup S_\infty)] = 1$ for any utility optimal strategy profile π^*. (Recall that both sets are absorbing for optimal strategies.) Suppose there exists some set of states S' disjoint from both S_0 and S_∞ in which paths remain forever with positive probability under π^*. There necessarily exists a subset of S' which are visited infinitely often with probability 1 under π, let S' equal this subset. If all states in S' have a reward of 0, then necessarily $S' \subseteq S_0$: Since π^* is optimal but nevertheless a reward of 0 is obtained in S', the Minimizer can ensure a total reward of 0 in S'. Dually, if there is some state with non-zero reward in S', this state occurs infinitely often with probability 1, ensuring an infinite expected reward, i.e. $S' \subseteq S_\infty$.

To summarize, we have that S_0 and S_∞ are the unique "sinks" of the game under *optimal strategies*. With this in place, we can proceed with the main proof.

We need to show equivalence of optimal values and a correspondence between strategies. In particular, we shall prove that any optimal strategy profile in G is directly equivalent to an optimal strategy in G_R, and for the other direction we only need to modify the decisions of Minimizer on S_0 to keep the play inside. We note a subtlety: We first prove that any optimal strategy in G achieves a reachability probability of NegUtil* in G_R. This alone does not yet show that NegUtil* is the optimal reachability probability. We then also prove that optimal strategies for reaching S_0 on G_R correspond to a strategy achieving NegUtil* on G. In the following, we slightly abuse notation and identify strategies on G with strategies on G_R, extended by the meaningless choice in $\underline{s}$.

From G to G_R: We start with the forward direction. Let $\pi^* = (\sigma^*, \tau^*)$ be utility-optimal strategies in G. We shall argue that $\mathrm{NegUtil}^* = \mathrm{NegUtil}(\pi^*) = \mathrm{Pr}^{\pi^*}_{\mathsf{G}_R,\hat{s}}[\Diamond S_0]$. To this end, we investigate the set $\Diamond S_0$ more closely. Every path which reaches a set does so after finitely many steps. Thus, we split this set by the finite prefix after which paths reach their respective destination (a countable set). Formally, for a prefix ϱ, we write $\Diamond_\varrho S_0$ to denote the set of all paths which begin with ϱ and reach S_0 exactly at the end of ϱ. For simplicity, we define this set to be empty if ϱ does not end in S_0 or already reached S_0 earlier. This way, we have in general that $\Diamond S_0 = \bigcup_\varrho \Diamond_\varrho S_0$ and the sets are pairwise disjoint. The probability of reaching S_0 in G_R thus can be written as $\sum_\varrho \mathrm{Pr}^{\pi^*}_{\mathsf{G}_R,\hat{s}}[\Diamond_\varrho S_0]$, the probability of all finite paths reaching S_0 (observe the similarity to the general proof of measurability for $\Diamond T$). By inserting the definition of transition probabilities in G_R and reordering, we obtain $\prod_{i=1}^{|\varrho|} b^{-\gamma r(\varrho_i)} \cdot \mathrm{Pr}^{\pi^*}_{\mathsf{G},\hat{s}}[\Diamond_\varrho S_0]$. Note that $\underline{s}$ is absorbing and thus cannot occur on any path in $\Diamond_\varrho S_0$.

Recall that no rewards are obtained once a path reaches S_0 (since τ^* is optimal), thus the total reward of any path $\rho \in \Diamond_\varrho S_0$ exactly equals $\mathrm{TR}(\rho) = \sum_{i=1}^{|\varrho|} r(\varrho_i)$. Since NegUtil is defined as expectation, we can similarly split up the set of all runs in G in a linear manner first into $\Diamond S_0$ and $\overline{\Diamond S_0}$ and then further split up $\Diamond S_0$ as above. Together, we obtain

$$\mathrm{NegUtil}^* = \mathbb{E}^{\pi^*}_{\mathsf{G},\hat{s}}[b^{-\gamma\,\mathrm{TR}}] = \sum\nolimits_\varrho b^{-\gamma\,\mathrm{TR}(\varrho)} \cdot \mathrm{Pr}^{\pi^*}_{\mathsf{G},\hat{s}}[\Diamond_\varrho S_0] + \int\limits_{\rho \in \overline{\Diamond S_0}} b^{-\gamma\,\mathrm{TR}(\rho)}\, d\mathrm{Pr}^{\pi^*}_{\mathsf{G},\hat{s}}.$$

Observe that the left hand side exactly equals the probability of $\Diamond S_0$ in G_R as we argued above. It thus remains to show that the remaining integral has a value of zero. Here, we need to exploit the optimality of σ^*. Recall that in this case almost all paths which do not reach S_0, i.e. $\overline{\Diamond S_0}$, instead reach S_∞. Thus, for all these paths the Maximizer ensures an infinite total reward, corresponding to a utility of 0.

In summary, every optimal strategy profile π^* for G reaches S_0 in G_R with probability NegUtil*.

From G_R to G: Let $\pi^* = (\sigma^*, \tau^*)$ be (memoryless deterministic) reachability-optimal strategies in G_R. Note that transition probabilities on S_0 are the same for G and G_R, since $r(s) = 0$ on all these states. Thus, there exists a Minimizer strategy which keeps the play inside S_0 in G_R and we assume w.l.o.g. that τ^* behaves in this way. This clearly does not influence the probability of reaching S_0 in the first place and thus τ^* remains an optimal strategy. Moreover, for S_∞ observe that there exists a Maximizer strategy in G ensuring that an infinite reward is obtained with probability 1. Then, following this strategy on S_∞ in G_R ensures that $\underline{s}$ is reached almost surely. This means that once S_∞ is reached, Maximizer can ensure that S_0 is never reached, since $\underline{s}$ is absorbing.

We divide the set of all possible infinite paths in G_R into three groups, namely (i) those which reach S_0, (ii) those which reach $\underline{s}$, and (iii) all others (which, as we will show, turn out to be a null set). Note that for (ii) considering those which reach S_∞ would be wrong, since $\underline{s}$ can also be reached from other states (to be precise, from any state with $r(s) > 0$). For the first kind, we apply the same reasoning as above, obtaining that $\mathrm{Pr}^{\pi^*}_{\mathsf{G}_R,\hat{s}}[\Diamond_\varrho S_0] = \prod_{i=1}^{|\varrho|} b^{-\gamma r(\varrho_i)} \cdot \mathrm{Pr}^{\pi^*}_{\mathsf{G},\hat{s}}[\Diamond_\varrho S_0]$. For the second case, observe there is no direct equivalent of $\Diamond\{\underline{s}\}$ in G. Instead, we show that the set of states for which the optimal probability to reach S_0 is 0 exactly is $S_\infty \cup \{\underline{s}\}$. As argued above, this certainly is true for all states in $S_\infty \cup \{\underline{s}\}$. Thus, choose some state $s \in S \setminus (S_0 \cup S_\infty)$. For this state, there exists a Minimizer strategy in G ensuring that a finite total reward is obtained with non-zero probability (otherwise, s would belong to S_∞). This can only be the case if S_0 is reached with non-zero probability under this strategy in G. By replicating this strategy on G_R, Minimizer can as well ensure a non-zero probability of reaching S_0, since every path in G also is possible in G_R (only with a potentially decreased probability). We now want to argue that the third kind has measure zero. Consider the Markov chain M induced by π^* (which is finite since π^* is memoryless). If M had any BSCC containing states in neither S_0 nor $S_\infty \cup \{\underline{s}\}$, we could adapt the Minimizer strategy to follow the previous strategy in this BSCC, increasing the probability to reach S_0. Recall that Minimizer wants to maximize the probability of reaching S_0, so this contradicts the optimality of π^*. Consequently, the induced Markov chain only has BSCCs which are subsets of either S_0 or $S_\infty \cup \underline{s}$, meaning almost all paths reach either of these. Together we obtain the result, i.e. that π^* obtains a utility of $\mathrm{NegUtil}^* = \mathrm{Pr}^{\pi^*}_{\mathsf{G}_R,\hat{s}}[\Diamond_\varrho S_0]$ in G.

Combining the results: From the first part, we get that any utility optimal strategy obtains a reachability probability of NegUtil* in G_R. From the second part, we dually get that the optimal reachability probability in G_R equals NegUtil* and the optimal strategies for the reachability correspond to strategies obtaining NegUtil* in G_R. Consequently, memoryless deterministic strategies are sufficient to optimize utility and we can obtain both the optimal utility as well as optimal strategies by solving the reachability game G_R. □

We note that reachability games can also be reduced to our case:

Lemma 4.5. *For any game* G *and (absorbing) reachability goal T, we have* $\mathrm{Val}_{\mathsf{G},\Diamond T}(s) = 1 - \mathrm{NegUtil}^*_{\mathsf{G}}(s)$ *with reward* $r(s) = \mathbb{1}_T(s)$ *and* $\gamma = 1$.

Proof. We assume w.l.o.g. that all states of T are absorbing in G (which can be achieved in linear time without changing the reachability probability). Now, observe that the transition structure of G_R completely agrees with G except on T, where every state has a self-loop and a transition to $\underline{s}$. As we argued in the proof before, under optimal strategies almost all runs either reach S_0 or $S_\infty \cup \underline{s}$ in G_R. Thus, since Maximizer is minimizing the probability to reach S_0, the probability to reach $S_\infty \cup \underline{s}$ is maximized. To conclude, observe that (i) $T \subseteq S_\infty$, in particular S_∞ additionally exactly contains all states from which the Maximizer can force the play into T with probability one, and (ii) $\underline{s}$ cannot be reached without reaching S_∞ first. Hence, the optimal probability of reaching $S_\infty \cup \underline{s}$ in G_R equals the probability to reach S_∞ which in turn is the optimal probability to reach T in G. □

We highlight that this reduction from entropic risk games to reachability games is *not* an effective reduction in the computational sense, since G_R comprises *irrational* transition probabilities even for entirely rational inputs. We discuss how to tackle this in the next section and first proceed to derive some useful properties from this correspondence.

Lemma 4.6. *The optimal utility* NegUtil* *is the pointwise smallest solution of*

$$\begin{aligned} &v(s) = 0 \quad \text{for } s \in S_\infty, \qquad v(s) = 1 \quad \text{for } s \in S_0, \text{ and} \\ &\quad v(s) = \overline{\mathrm{opt}}^{s}_{a \in A(s)} b^{-\gamma r(s)} \cdot \Delta(s,a)\langle v \rangle \quad \text{otherwise} \end{aligned} \tag{6}$$

Proof. Follows directly from Lemma 4.4 combined with standard result on reachability for stochastic games [45]. Note that maximizing and minimizing reachability in stochastic games is equivalent, since we can simply swap the players. □

4.3. Uniqueness of fixed points

There might be multiple fixed points to the system of equations 4.6. This is to be expected, since already reachability on MDPs exhibits this problem [52]. We identify a condition on G which ensures the reachability game G_R being "stopping" [3] (or "halting"), which implies that the fixed point is unique.

Lemma 4.7. *Equation* (6) *has a unique solution if for every strategy profile* π *and state* s*, we have* $\mathsf{Pr}^{\pi}_{\mathsf{G},s}[\Diamond(S_0 \cup \{s \mid r(s) > 0\})] > 0$*, i.e.* S_0 *or a state that yields some non-zero reward is always reached with some positive probability.*

Proof. Under the assumption, the reachability game G_R is *stopping*, i.e. no matter what either player does, the game eventually stops with probability 1. Uniqueness of the equations for reachability on G_R follows from standard results on reachability [3].

Intuitively, given the assumptions, the reachability game G_R will reach $S_0 \cup \{\underline{s}\}$ with positive probability from any state (since there is a positive probability to directly transition to $\underline{s}$ in G_R for any state with $r(s) > 0$). Now, observe that any set can only be reached with positive probability if there exists a path of length at most $|S|$ to it. This implies that every $|S|$ steps, there is some probability of reaching those states. If we repeat this ad infinitum, these states will be reached with probability 1. One can prove that the fixed point iteration corresponding to Equation (6) applied $|S|$ times is a contraction and thus has a unique fixed point. □

However, we hardly can expect to find a necessary condition: Since our problem is essentially equivalent to reachability, the remarks of [53] transfer to our case. In particular, they give an if and only if condition for uniqueness of fixed points together with reasoning why we cannot easily decide this condition through graph analysis. On the positive side, we can efficiently compute both S_0 and S_∞ as well as check the stopping criterion directly on G.

Lemma 4.8. *The sets* S_0 *and* S_∞ *as well as the property of Lemma 4.7 can be obtained / checked in time quadratic w.r.t. the number of states and transitions (assuming that the set* $\{s \mid r(s) > 0\}$ *can be determined in at most quadratic time).*

Proof. Decidability of such qualitative properties is well-known (e.g. through attractor computations on the game graph). For completeness, we rephrase the problems in terms of [54]. Observe that S_0 is equivalent to the condition "at every step we have $r(s) = 0$ with probability 1" and S_∞ means "we infinitely often see $r(s) > 0$ with probability 1", i.e. a *sure safety* and *sure Buchi* condition, respectively. Similarly, for the condition of Lemma 4.7, consider its negation: We ask if there exists a strategy under which the probability to reach $S_0 \cup \{s \mid r(s) > 0\}$ is zero – a sure safety query. All three queries can be easily answered using the approaches of [54].

We note that all computations only require knowledge of the transition structure of the underlying (hyper-)graph and the set of states with $r(s) > 0$ (which we assumed to be computable in quadratic time). □

4.4. Strategy complexity

By Lemma 4.4, the optimal negative exponential utility is obtained by reachability-optimal strategies in G_R. With the known results on reachability [4], this yields:

Theorem 4.9. *MD strategies are sufficient to optimize the negative exponential utility and thus also entropic risk. More precisely, for all SGs* G*, there is an MD strategy* σ *for the Maximizer such that* $\mathrm{ERisk}^{\gamma *}_{\mathsf{G},s} = \inf_{\tau \in \Pi_{\mathsf{G}}} \mathrm{ERisk}^{\gamma}_{\mathsf{G},s}((\sigma, \tau))$ *and analogously for the Minimizer.*

Remark 4.10. We highlight that this means that this notion of risk is history independent: Which actions are optimal does not depend on what has already "gone wrong", but purely on the potential future consequences. This is in stark contrast to, e.g., conditional value-at-risk optimal strategies for total reward, which require exponential memory and switch to a purely expectation maximizing (i.e. risk-ignorant) behavior after "enough" went wrong [24].

4.5. System of inequalities

The problem we want to solve is deciding whether the Maximizer can ensure an entropic risk of at least t. Unfortunately, the reachability game G_R is not directly computable, since even for rational rewards $b^{-\gamma r(s)}$ may be irrational. As such, we cannot use this transformation directly to prove decidability or complexity results and need to take a different route. Analogous to the classical solution to reachability, we first convert the problem to a system of inequalities. Intuitively, we replace every max with $\geq$ for all options and dually min with $\leq$ (again, recalling that Maximizer wants to minimize the value in G_R). Formally, we consider the following:

$$\begin{aligned} v(\hat{s}) \leq b^{-\gamma t}, \qquad v(s) = 0 \quad \text{for } s \in S_\infty, \qquad v(s) = 1 \quad \text{for } s \in S_0, \\ v(s) \leq b^{-\gamma r(s)} \cdot \Delta(s,a)\langle v \rangle \quad \text{for } s \in S_{\max}, a \in A(s), \\ v(s) \geq b^{-\gamma r(s)} \cdot \Delta(s,a)\langle v \rangle \quad \text{for } s \in S_{\min}, a \in A(s), \text{ and} \\ \bigvee\nolimits_{a \in A(s)} v(s) = b^{-\gamma r(s)} \cdot \Delta(s,a)\langle v \rangle \quad \text{for } s \in S \end{aligned} \tag{7}$$

Observe that this essentially is the decision variant to the standard quadratic program for reachability applied to G_R [5].

Lemma 4.11. *The system of equations (7) has a solution if and only if* $\mathrm{ERisk}^* \geq t$.

Proof. If: In this case, i.e. $\mathrm{ERisk}^* \geq t$, we have $\mathrm{NegUtil}^* \leq b^{-\gamma t}$. Thus, $\mathrm{NegUtil}^*$, which is a solution to Equation (5), immediately satisfies all equations.

Only If: Observe that the first four equations ensure that any solution actually solves Equation (6). Since $\mathrm{NegUtil}^*$ is the pointwise smallest solution by Lemma 4.6, having a vector v which satisfies the last inequality ensures that $\mathrm{NegUtil}^*$ does so, too. We conclude by noticing the equivalence of the first four equations to optimal solutions of the quadratic program for reachability, see e.g. [5] for further information. □

4.6. Decidability subject to Shanuel's conjecture

From Equation (7), we obtain a conditional decidability result for the general case:

Theorem 4.12. *Let all quantities, i.e. rewards, transition probabilities, the risk-aversion factor* γ*, and the basis* b *be given as formulas in the language of reals with exponentiation (i.e. with functions* $+$*,* $\cdot$*, and* $\exp\colon x \mapsto e^x$*). Then, the entropic risk threshold problem for SGs is decidable subject to Schanuel's conjecture.*

Proof. In this case, the existence of a solution to Equation (7) can also be expressed as a sentence in the language of the reals with exponentiation. The corresponding theory is known to be decidable subject to Shanuel's conjecture (see e.g. [55]) as shown by [56], and decidability of this theory is equivalent to the so-called "weak Schanuel's conjecture". □

In particular, this allows us to treat instances with basis $b = e$. Yet, even if all rewards, transition probabilities, and γ are given as rational values, but the basis b equals e, we do not know how to check the satisfiability of Equation (7) without relying on the theory of the reals with exponentiation. Note, however, that we do not need the "full power" of the exponential function: All values appearing in an exponent in Equation (7) are constants. So, the restricted exponential function that agrees with exp on a closed interval $[a_1, a_2]$ and is zero outside of this interval is sufficient. The theory of the reals with restricted exponentiation has some additional nice properties compared to the theory of the reals with full exponentiation: For example, it allows for quantifier elimination by [57] and related works. Nevertheless, this does not allow us to immediately obtain an unconditional decidability result.

5. The algebraic case

If all occurring values are rational, then all numbers of the system of inequalities Equation (7) are algebraic. The results of this section establish that the threshold problem for such instances is decidable. First, we will address the general algebraic case (Section 5.1). Afterwards, we will proceed to identify special cases of algebraic instances that ensure that the computations can be carried out efficiently using arithmetic on algebraic numbers (Section 5.2 and Section 5.3). We will conclude this section with a brief analysis of the case that $\gamma \cdot rew(s)$ is an integer for all states s (Section 5.4). An overview of the complexity results can also be found in Table 1.

5.1. General algebraic case

Formally, we define:

Definition 5.1. An *algebraic instance* of the entropic risk threshold problem is an instance where all occurring values, i.e. the transition probabilities of the game G, all rewards assigned by the reward function r, the risk-aversion parameter γ, the basis b, and the threshold t, are rational and encoded as the fraction of co-prime integers in binary.

In general, for algebraic instances, there is a reduction of our problem to the existential theory of the reals, leading to the following result where $\exists\mathbb{R}$ denotes the complexity class of problems that are polynomial-time reducible to the existential theory of the reals:

Theorem 5.2. *For algebraic instances, the entropic risk threshold problem is decidable in $\exists\mathbb{R}$ and thus in PSPACE.*

More precisely, the existence of a solution of Equation (7) for an algebraic instance of our problem can be expressed as an existential sentence in the language of the reals without exponentiation. Let us first illustrate this with a brief example.

Example 5.3. For the MDP depicted in Fig. 2, the set S_∞ is $\{s_2\}$ and the set S_0 is $\{s_3\}$. So, Equation (7) results in the following set of constraints for basis $b = 2$, risk aversion factor $\gamma = 1$ and for a threshold $t = 4$:

$$\begin{gathered} v(s_1) \le 2^{-4}, \qquad v(s_2) = 0, \qquad v(s_3) = 1, \\ v(s_1) \le 2^{-2} \cdot (\tfrac{1}{2} v(s_2) + \tfrac{1}{2} v(s_3)), \qquad v(s_1) \le 2^{-2} \cdot v(s_4), \qquad v(s_4) \le 2^{-4} \cdot v(s_3), \\ v(s_1) = 2^{-2} \cdot (\tfrac{1}{2} v(s_2) + \tfrac{1}{2} v(s_3)) \quad \vee \quad v(s_1) = 2^{-2} \cdot v(s_4), \\ v(s_4) = 2^{-4} \cdot v(s_3). \end{gathered} \tag{8}$$

The satisfiability of these constraints can now easily be expressed by existentially quantifying the values $v(s_i)$ for $i = 1, 2, 3, 4$. In the sequel, we show why exponentiation is not necessary to express this as a first-order sentence exploiting that all expressions using exponentiations are constants.

In general, all occurring constants are algebraic numbers of the form $c \cdot b^{\frac{p}{q}}$ with $p, q \in \mathbb{N}$ and $c \in \mathbb{Q}$ in this case. These can be represented in linear space using exponentiation by repeated squaring as shown in the following lemma.

Lemma 5.4. *Let $b = \frac{u}{v} \in \mathbb{Q}$ a rational number where $u, v \in \mathbb{N} \setminus \{0\}$. Given $p \in \mathbb{Z}$ and $q \in \mathbb{N} \setminus \{0\}$, we can define $b^{\frac{p}{q}}$ by an existential sentence in the language of the reals using $\mathcal{O}(\log p + \log q + \log u + \log v)$ symbols.*

Proof. We provide a formula defining x such that $x = b^{\frac{p}{q}}$. Let p and $q > 0$ be natural numbers with binary representations $p = \sum_{i=0}^{P} p_i \cdot 2^i$ and $q = \sum_{i=0}^{Q} q_i \cdot 2^i$ where $p_i, q_i \in \{0, 1\}$ for all $0 \le i \le P$ and Q, respectively. The formula $x = b^{\frac{p}{q}}$ with free variable x is equivalent to $x^q = b^p$. Using $d_0 = b$ as abbreviation for $v \cdot d_0 = u$, the formula $x^q = b^p$ can then be written as

$$\begin{aligned} \exists c_0, \ldots, c_Q, d_0, \ldots, d_P . c_0 = x \wedge \bigwedge\nolimits_{i=1}^{Q} (c_i = c_{i-1} \cdot c_{i-1}) \wedge \\ d_0 = b \wedge \bigwedge\nolimits_{i=1}^{P} (d_i = d_{i-1} \cdot d_{i-1}) \wedge \\ \prod\nolimits_{\{i \mid p_i = 1\}} d_i = \prod\nolimits_{\{i \mid q_i = 1\}} c_i \end{aligned}$$

This formula is uniquely satisfied by assigning $c_i = x^{2^i}$ and $d_i = b^{2^i}$. Observe that $q = \sum_{i=0}^{Q} 2^i q_i$, and consequently $x^q = \prod_{i=0}^{Q} x^{2^i \cdot q_i} = \prod_{\{i \mid q_i = 1\}} x^{2^i}$, and similar for b^p. For the size bound, observe that $Q = \lceil \log_2 q \rceil$ and $P = \lceil \log_2 p \rceil$, thus the conjunctions and multiplication require $\mathcal{O}(\log p + \log q)$ symbols. Finally, $v \cdot d_0 = u$ requires $\mathcal{O}(\log v + \log u)$ space to specify. □

Consequently, an existential sentence in the language of the reals expressing the satisfiability of Equation (7) can be computed in polynomial time from G, b, and γ. This shows Theorem 5.2.

Remark 5.5. We note that already in simple MCs, solutions might not be rational. Consider a MC yielding a distribution over total rewards of $\{0 \mapsto \frac{1}{2}, 1 \mapsto \frac{1}{2}\}$. For $b = 2$ and $\gamma = 1$, ERisk* equals $-\log_2(\frac{1}{2} \cdot 2^0 + \frac{1}{2} \cdot 2^{-1}) = \log_2(\frac{4}{3}) = 2 - \log_2(3)$, an irrational number. Note that the input comprises only small, rational values. The threshold problem for the negative exponential utility suffers from the same issue: For $b = 2$ and $\gamma = \frac{1}{2}$, NegUtil* on the same MC equals $\frac{1}{2} \cdot 2^0 + \frac{1}{2} \cdot 2^{-\frac{1}{2}} = \frac{1}{2(1+\sqrt{2})}$, again an irrational number.

On the one hand, this indicates that already for MCs, there is no obvious improvement of the complexity upper bound. Although the negative exponential utility is the solution to a linear system of equations in the MC case, the main obstacle for efficient computations is the complicated form of the occurring numbers. On the other hand, the "bottleneck" for complexity lies already in the MC case: If C is the threshold problem for MCs, the problem for SGs can be solved in $(\Sigma_2^P)^C \cap (\Pi_2^P)^C$, where $(\Sigma_2^P)^C$ and $(\Pi_2^P)^C$ denote the respective complexity classes with access to an oracle for C, due to determinacy and optimality of MD strategies: In a game G, we have $\text{ERisk}^*_G \geq t$ if, for all MD-strategies for the Minimizer, there is an MD-strategy for the Maximizer such that in the resulting MC ERisk $\geq t$, or equivalently, if there is an MD-strategy for the Maximizer such that for all MD-strategies for the Minimizer ERisk $\geq t$ in the resulting MC. This puts the problem into $(\Sigma_2^P)^C$ and $(\Pi_2^P)^C$.

5.2. Threshold problem in algebraic extensions of low degree

For Theorem 5.2, we use the standard decision procedure for the existential theory of the reals as a "black box" and do not make use of the special form of our problem. To exploit the specific structure of the system of inequalities, we note that for explicit computations on algebraic numbers the following two quantities are relevant for the resulting computational complexity: Firstly, the degree of the field extension of $\mathbb{Q}$ in which the computation can be carried out. Secondly, the bitsize of the coefficients of the minimal polynomials of the involved algebraic numbers (see, e.g., [58,59]). Alternatively, the bitsize of the representations of the algebraic numbers in a fixed basis of the field extension in which the computations can be carried out can be used. Note that the size of the basis is precisely the degree of that field extension. Motivated by these observations, we consider *small algebraic instances*, which allow us to prove that all occurring algebraic numbers have a sufficiently small representation.

Definition 5.6. A *small algebraic instance* of the entropic risk threshold problem consists of a SG G with rational transition probabilities, an integer reward function r, a rational risk-aversion parameter γ, a rational basis b, and a rational threshold t. Moreover, the rewards, γ, and t are encoded in unary, and as the fraction of co-prime integers encoded in unary, respectively. The remaining rational numbers are encoded as the fraction of co-prime integers in binary. If G is an MDP or a MC, we call the instance a small algebraic instance of an MDP or a MC.

Remark 5.7. For simplicity, we assume for small algebraic instances that all rewards are in $\mathbb{N}$. If this is not the case, we can multiply all rewards with the least common multiple D of the denominators of the rewards and use a new risk-aversion parameter $\gamma' = \gamma/D$. The resulting negative exponential utility is not affected by this transformation. The change of the optimal entropic risk by a factor of D can be addressed by also rescaling the threshold $t' = t \cdot D$. Nevertheless, note that this affects the encoding size of the risk-aversion factor γ.

Relying on algorithms for explicit computations in algebraic numbers, we will obtain the following result.

Theorem 5.8. *For small algebraic instances, the entropic risk threshold problem: (a) belongs to* NP ∩ coNP *for SGs; and (b) can be solved in polynomial time for MDPs or MCs.*

Before we begin with the proof, we show that despite the restrictions on the encoding of the input, our problem for the entropic risk in SGs is still at least as hard as for general reachability games. So, we cannot expect a polynomial-time algorithm for SGs without solving long-standing open problems.

Proposition 5.9. *The threshold problem for stochastic reachability games is polynomial-time reducible to the entropic risk threshold problem on small algebraic instances of SGs.*

Proof. This follows from Lemma 4.5: For each stochastic reachability game G, we can construct an SG G′ with rewards 0 and 1, basis $b = 2$, and $\gamma = 1$ in polynomial time such that the reachability value in G equals the optimal entropic risk in G′. Clearly, this is a small algebraic instance. □

The remainder of this section is devoted to the proof of Theorem 5.8. We consider the case of MDPs. The result for SGs will be a simple consequence afterwards. In MDPs, we will see that solving the threshold problem boils down to checking the solvability of a linear system of inequalities. Throughout this section, fix a small algebraic instance of an MDP $\mathcal{M} = (S, A, \Delta)$ with $\gamma = p/q$ and initial state $\hat{s}$. All coefficients occurring in Equation (7) now are of the form $\Delta(s, a, t) \cdot b^{-r(s)\cdot p/q}$ for a rational probability $\Delta(s, a, t)$, a natural reward $r(s)$, and the rational basis b. Consequently, all coefficients are contained in the field extension $\mathbb{Q}(b^{1/q})$.

We now determine the degree and the minimal polynomial of $b^{1/q}$. Let $b = b_1/b_2$ for co-prime natural numbers b_1 and b_2. Now, let d be the greatest divisor of q such that b_1 and b_2 have integral d-th roots d_1 and d_2. Then, we can rewrite $b^{1/q} = (d_1/d_2)^{d/q}$. We now define $q' = q/d$ and $b' = d_1/d_2$. Thus, $(b')^{1/q'} = b^{1/q}$ and q' is a natural number smaller or equal to q. Now, q' has no divisor d' such that b' has a rational d'th root. By [60, Thm 8.1.6], we conclude that $x^{q'} - b'$ is irreducible and hence the minimal polynomial of $b^{1/q}$.

In fact, we can switch to the risk aversion factor $\gamma' = p/q'$ and the basis b' instead of γ and b' without affecting the negative exponential utility. The entropic risk is changed by the integral factor $d \leq q$ by this switch. Hence, we from now on work with the following assumption:

Assumption 5.10. In a small algebraic instance with risk aversion factor $\gamma = p/q$ and basis b, we can assume w.l.o.g. that $x^q - b$ is irreducible over $\mathbb{Q}$.

In case of an MDP in which all states belong to the Minimizer, i.e. NegUtil is maximized, we obtain the following system of linear inequalities from Equation (7):

$$\begin{aligned} v(s) = 0 \quad \text{for } s \in S_\infty, \qquad v(s) = 1 \quad \text{for } s \in S_0, \\ v(s) \geq b^{-\gamma r(s)} \cdot \Delta(s, a)\langle v\rangle \quad \text{for } s \in S_{\min}, a \in A(s), \\ v(\hat{s}) \leq b^{-\gamma t}. \end{aligned} \tag{9}$$

Note that we can drop the disjunction from Equation (7) as the point-wise least solution to Equation (9) always satisfies this disjunction. However, if instead all states belong to the Maximizer we cannot drop the disjunction directly. Nevertheless, we consider the following linear program

$$\begin{aligned} \text{Maximize } v(\hat{s}) \text{ subject to} \\ v(s) = 0 \quad \text{for } s \in S_\infty, \qquad v(s) = 1 \quad \text{for } s \in S_0, \\ v(s) \leq b^{-\gamma r(s)} \cdot \Delta(s, a)\langle v\rangle \quad \text{for } s \in S_{\max}, a \in A(s). \end{aligned} \tag{10}$$

Then, in the optimal solution, we have to check whether $v(\hat{s}) \leq b^{-\gamma t}$; or in other words, all solutions to the constraints have to satisfy $v(\hat{s}) \leq b^{-\gamma t}$. By considering the dual linear program instead, we can transform this linear program into one where the objective function has to be minimized and has the same optimal value $v(\hat{s})$. Here, comparing this optimal value to the threshold by $v(\hat{s}) \leq b^{-\gamma t}$ boils down to finding one solution that satisfies all constraints and the threshold condition again. So, for both Maximizer- and Minimizer-MDPs, the entropic risk threshold problem boils down to checking the feasibility of a linear program, i.e. the satisfiability of a system of linear inequalities.

In [58,59], it is shown that the feasibility of a linear program over algebraic numbers can be checked in time polynomial in the size of the representations of the occurring algebraic numbers and in the degree of a field extension of $\mathbb{Q}$ that contains all coefficients. The representation of algebraic numbers used in [58,59] is the following: An algebraic number α is represented by its minimal polynomial P over $\mathbb{Q}$ together with a rational interval (a, b) containing only the zero α of P. The size of the representation is the bitsize of coefficients of P and the interval bounds of a and b as fractions of (co-prime) integers.

By Assumption 5.10, the degree of the field extension containing all occurring numbers is q and hence at most linear in the input size since q is given in unary. For the following lemma, we show that the coefficients in the minimal polynomials of all constants in the linear program of interest are small as well. Then, the mentioned polynomial-time algorithms for the feasibility of linear systems of inequalities over algebraic numbers from [58,59] are applicable.

Lemma 5.11. *Let $\mathcal{M}$ be a small algebraic instance of an MDP. We can decide whether* $\mathrm{ERisk}^*_{\mathcal{M}} \geq t$ *and whether* $\mathrm{ERisk}^*_{\mathcal{M}} \leq t$ *in polynomial time. Furthermore, for a rational threshold t' given in binary, we can decide whether* $\mathrm{NegUtil}^*_{\mathcal{M}} \leq t'$ *and whether* $\mathrm{NegUtil}^*_{\mathcal{M}} \geq t'$ *in polynomial time.*

Proof. In order to rely on the polynomial algorithms from [58,59], we have to show that we can obtain small representations of all coefficients in Equations (9) and (10). The coefficients have the form $\Delta(s, a, s') \cdot b^{-r(s)\cdot p/q}$ where $\Delta(s, a, s')$ and b given in binary and $r(s)$, p, and q in unary.

First of all, we note that it is sufficient to find a small representation for $b^{-r(s)\cdot p/q}$. As shown in [58,59], from this representation, a representation of $\Delta(s, a, s') \cdot b^{-r(s)\cdot p/q}$ can be found in polynomial time. So, let $n = r(s) \cdot p$. We rewrite

$n = k \cdot q + \ell$ for some natural number $\ell < q$. The numerical values of k and ℓ are polynomial in the size of the original input as $r(s)$, p and q are given in unary. Now, $b^{-n/q} = b^{-k} \cdot b^{-\ell/q}$. The number b^{-k} is a rational whose size in binary representation is linear in the size of the binary representation of b and in k. So, b^{-k} can be obtained in polynomial time and by the same argument as before, it is sufficient to find a small representation for $b^{-\ell/q}$.

The algebraic number $b^{-\ell/q}$ is a zero of the polynomial $x^q - b^{-\ell}$. Furthermore, it is the only positive real zero of this polynomial and it lies between 0 and 1. Consequently, the minimal polynomial of $b^{-\ell/q}$ has only one zero in the interval $(0, 1)$, too, which yields the interval for the representation. In order to find the minimal polynomial, note that $b^{-0/q}$, $b^{-1/q}$, ..., $b^{-(q-1)/q}$ are linearly independent over $\mathbb{Q}$ by Assumption 5.10. Let $m = q/\gcd(\ell, q)$. Now, the remainders of $0, \ell, 2\ell, 3\ell, \dots$ are periodic modulo q with period m. This means that $b^{-0\ell/q}$, $b^{-\ell/q}$, ..., $b^{-(m-1)\ell/q}$ are linearly independent over $\mathbb{Q}$ and consequently the degree of $b^{-\ell/q}$ is at least m. But $(b^{-\ell/q})^m$ is a rational as $\ell \cdot m$ is an integer multiple of q. Together, $x^m - (b^{-m\cdot\ell/q})$ is the minimal polynomial of $b^{-\ell/q}$. As $\ell \cdot m/q$ is a natural number smaller than q, the rational coefficient $b^{-\ell\cdot m/q}$ can be computed in polynomial time. Hence, the representation of $b^{-\ell/q}$ in terms of this minimal polynomial and the interval $(0, 1)$ can be obtained from $\mathcal{M}$ in polynomial time.

As the threshold t on the entropic risk in a small algebraic instance is given in unary, we can similarly obtain a small representation for the corresponding threshold $b^{-\gamma\cdot t}$ on the negative utility. If we directly want to compare the negative utility to a rational threshold t' instead, this threshold can be given in binary. Now, the polynomial time algorithms of [58,59] for checking the feasibility of linear programs with algebraic coefficients are applicable after computing the representations of all constants occurring in Equations (9) and (10) in polynomial time. □

For SGs, this result allows us to non-deterministically guess a strategy for one of the players and check the threshold condition in the resulting MDP in polynomial time. By guessing a strategy for Maximizer, we can check whether $\mathrm{ERisk}^*_{\mathsf{G}} \geq t$ in NP, by guessing a strategy for the Minimizer, we obtain a coNP-upper bound analogously, finishing the proof of Theorem 5.8. Note that for a rational threshold t' given in binary, the problem to decide whether $\mathrm{NegUtil}^*_{\mathsf{G}} \geq t'$ for a small algebraic instance G of an SG is in NP ∩ coNP as well by analogous reasoning and Lemma 5.11.

5.3. Optimal solution in algebraic extensions of low degree

While the mentioned results concern the threshold problem, we can even go a step further in small algebraic instances of MCs. Here, the system of inequalities simplifies to a linear system of equations, which we can solve *explicitly* in the algebraic numbers. For small algebraic instances, this is possible in polynomial time yielding the following main result of this section.

Theorem 5.12. *For small algebraic instances, an explicit representation of* NegUtil* *can be computed in: (a) polynomial time for MCs; and (b) in polynomial space for SGs and MDPs.*

To prove this result, we take a closer look at how the equation system we obtain for MCs can be solved with computations in the algebraic numbers. In the end, we state the complexity-theoretic consequences for SGs that we obtain from this solution for MCs.

Let $\mathsf{M} = (S, \delta)$ be a small algebraic instance of a MC with reward function $r : S \to \mathbb{N}$, initial state $\hat{s}$, $\gamma = p/q \in \mathbb{Q}$ a risk-aversion parameter for co-prime integers p and q, and $b \in \mathbb{Q}$ a basis. Again, we work with Assumption 5.10 as justified in the previous section. The following lemma makes the statement of Theorem 5.12 about MCs more precise.

Lemma 5.13. *A representation of the negative exponential utility* NegUtil* *in* M *in the basis* $B = (1, b^{1/q}, b^{2/q}, \dots, b^{q-1/q})$ *of* $\mathbb{Q}(b^{1/q})$ *can be computed in polynomial time.*

In the sequel, we prove this lemma using a sequence of further lemmata below. The first step to compute an explicit representation of NegUtil* in M is the computation of sets S_0 and S_∞ of states from which reward 0 or ∞, respectively, is received almost surely. This can be done by the analysis of bottom strongly connected components and the computation of reachability probabilities in MCs in polynomial time, but of course it is also a special case of the analogous computation of such sets in games above. The system of inequalities Equation (7) simplifies to

$$\begin{aligned} v(s) &= 0 \quad \text{for } s \in S_\infty, \qquad v(s) = 1 \quad \text{for } s \in S_0, \\ v(s) &= \sum\nolimits_{t \in S} b^{-\gamma\cdot r(s)} \cdot \delta(s,t) \cdot v(t) \quad \text{for } s \in S \setminus (S_0 \cup S_\infty). \end{aligned} \tag{11}$$

This is a linear equation system $Ax = v$ where the entries of A and v are of the form $c \cdot b^{-\gamma w}$ for rational c and natural numbers w. Due to pre-processing of S_0 and S_∞, we know that $Ax = v$ has a unique solution and that this solution contains the negative exponential utility from each state.

As we assume that all rewards are natural numbers, all entries in A and v are elements of the number field $\mathbb{Q}(b^{1/q})$. By Assumption 5.10, $\mathbb{Q}(b^{1/q})$ is a q-dimensional $\mathbb{Q}$-vector space. The tuple $B = (1, b^{1/q}, b^{2/q}, \dots, b^{(q-1)/q})$ forms a basis of $\mathbb{Q}(b^{1/q})$. So, each element $a \in \mathbb{Q}(b^{1/q})$ can be represented by a vector $z \in \mathbb{Q}^q$ with $[z]_B = \sum_{i=0}^{q-1} z_i \cdot b^{i/q} = a$.

Now, our goal is to compute an explicit representation of the value $v(\hat{s})$ in the unique solution of Equation (11), i.e. a vector $z \in \mathbb{Q}^q$ such that $[z]_B = v(\hat{s})$. For this purpose, we solve Equation (11) via Gaussian elimination by explicit computations in $\mathbb{Q}(b^{1/q})$ using representations in the basis B. The following three technical results show how this can be done and prove Lemma 5.13.

Lemma 5.14. *For a number of the form $c \cdot b^{-n/q}$ with $c \in \mathbb{Q}$ and $n \in \mathbb{N}$ we can compute its representation in basis B in time polynomial in the length of the binary representation of c, and polynomial in the numerical values of n and q.*

Proof. We can write $-n/q$ as $-k + \ell/q$ for natural numbers k and ℓ with $\ell < q$. Then, $c \cdot b^{-n/q} = c \cdot b^{-k} \cdot b^{\ell/q}$ which has the representation $c \cdot b^{-k} \cdot e_\ell$ in $\mathbb{Q}^q$ where e_ℓ is the ℓ-th standard basis vector. The coefficient $c \cdot b^{-k}$ can be computed in time polynomial in the representation of c and in the numerical value of k which is bounded by the numerical value of n. □

Lemma 5.15. *Given two representations $z, z' \in \mathbb{Q}^q$ in basis B of numbers $y = [z]_B$ and $y' = [z']_B$ in $\mathbb{Q}(b^{1/q})$, we can compute the representation of $y \cdot y'$ in time polynomial in the size of the binary representations of z and z'. Furthermore, we can compute the representation of y^{-1} in time polynomial in the size of the binary representations of z.*

Proof. Multiplication is a bi-linear map. Thus, in order to determine a representation of $[z]_B \cdot [z']_B$ for given $z, z' \in \mathbb{Q}^q$, it is sufficient to know the q^2-many representations $m_{\ell,h}$ of $[e_\ell]_B \cdot [e_h]_B$ for standard basis vectors e_ℓ and e_h with $0 \le \ell, h \le q-1$. The representation of $[z]_B \cdot [z']_B$ can then be computed as the sum

$$\sum\nolimits_{0 \le \ell, h \le q-1} z_\ell \cdot z'_h \cdot m_{\ell,h}, \tag{12}$$

where the vectors $m_{\ell,h}$ are the representations of $b^{(\ell-1)/q} \cdot b^{h/q} = b^{(\ell+h)/q}$. If $\ell + h < q$, the representation is $m_{\ell,h} = e_{\ell+h}$. If $q \le \ell + h < 2q$, the value is $b \cdot b^{\ell+h-q}$ and its representation is $b \cdot e_{\ell+h-q}$. This proves the first claim.

To compute the representation of the inverse y^{-1}, we take a vector of variables x. The equation

$$\sum\nolimits_{0 \le \ell, h \le q-1} z_\ell \cdot x_h \cdot m_{\ell,h} = e_0$$

yields a rational equation system with q variables which has a unique solution as the inverse is unique. Given z, the system can be constructed in polynomial time using the q^2-many vectors $m_{\ell,h}$. Consequently, the representation in basis B of y^{-1} can be computed in time polynomial in the binary representation of z, proving the second claim. □

Lemma 5.16. *Using Gaussian elimination on $Ax = v$, we obtain a representation of x in the basis B in time polynomial in the encoding size of the given small algebraic instance of a MC.*

Proof. We perform Gaussian elimination on the matrix A and vector v in a way that ensures that all intermediate numbers have a small representation. In A and v, we write representations from $\mathbb{Q}^q$ at every entry. Writing down the system with representations can be done in polynomial time by Lemma 5.14. In [61], it is shown that it is possible to perform Gaussian elimination in a way such that all numbers occurring during the computation are determinants of a submatrix of the original input. If the problem dimension is n, this means that all numbers occurring during the computation are sums of $n!$ products of at most n numbers from the original input. In our case, n is the number of states of the Markov chain M.

Let d be the least common multiple of all denominators of numbers in A and v. The bit size of d is linear in the bit sizes of the denominators. Generalizing Equation (12) in the proof of Lemma 5.15, we obtain that the product of n numbers given as representations $x^i \in \mathbb{Q}^q$ for $1 \le i \le n$ is

$$y = \sum_{\substack{0 \le \ell_i \le q-1 \\ 1 \le i \le n}} \prod_{i=1}^{n} (x^i)_{\ell_i} \cdot b^{\lfloor \sum_{i=1}^{n} (x^i)_{\ell_i} / q \rfloor} \cdot e_{(\sum_{i=1}^{n} (x^i)_{\ell_i} \bmod q)}.$$

If m is the maximal absolute value of any entry in one of the $x^i \in \mathbb{Q}^q$ for $1 \le i \le n$, each component of this vector y has absolute value less than $q^n \cdot m^n \cdot b^n$. Furthermore, if $b = b_1/b_2$, each component of y is an integer multiple of $1/(b_2^n \cdot d)$. So, each component y_i is the fraction of an integer less than $q^n \cdot m^n \cdot b^n \cdot b_2^n \cdot d$ and an integer less than $b_2^n \cdot d$. The bit size of these numbers is bounded by $n \cdot (\log_2(q) + \log_2(m) + \log_2(b) + \log_2(b_2)) + \log_2(d)$. Thus, the bit size of all components of y is at most polynomial in the bit sizes of the entries of the vectors $x^i \in \mathbb{Q}^q$ for $1 \le i \le n$.

Together, any product of n entries of A and v has a representation in basis B of polynomial size. Furthermore, all numbers that can occur are integer multiples of $1/(b_2^n \cdot d)$. Consequently, if we want to add $n!$-many such numbers, we can rewrite all rational numbers to denominator $1/(b_2^n \cdot d)$ and afterwards add the integer enumerators component-wise. This increases the bit size by a factor of at most $\log_2(n!) < \log_2(n) \cdot n$. So, all intermediate numbers that occur when performing Gaussian elimination as in [61] have a representation whose bit size is bounded by a polynomial in the input size. As $Ax = v$ has a unique solution, the Gaussian elimination produces this solution in polynomially many steps and all necessary multiplications and divisions can be carried out in polynomial time by Lemma 5.15 and the fact that all intermediate numbers occurring have a polynomially large representation. □

Put together, this finishes the proof of Lemma 5.13. So, on small algebraic instances of MCs, we can compute and explicit representation of the negative exponential utility in polynomial time. This allows us to conclude that on a small algebraic instance of an SG, an explicit representation of ERisk^* can be computed in polynomial space: In polynomial space, we can go through all MD-strategies σ for the Maximizer. For each strategy σ, we compute a representation of NegUtil for each Minimizer MD-strategy τ in the resulting MC and compare it to the least value we have seen so far that the Minimizer can enforce against σ. To compare the explicit representations computed in the process, we can rely on the algorithms in [58,59] to compare algebraic numbers. Once we have found the value the Maximizer can enforce with σ, we consider the next strategy of the Maximizer and keep track of the best value found so far. This concludes the proof of Theorem 5.12.

5.4. Integer exponents

To conclude, we consider the special case that b is a rational number and $\gamma \cdot r(s)$ is an integer for all states s. Then, the transition probabilities $b^{-\gamma r(s)}$ are rational. In this case, we can directly compute G_R, requiring space linear in the numerical values $\gamma \cdot r(s)$. Then, we can apply standard methods to decide whether the Maximizer can ensure a reachability probability of at most $b^{-\gamma t}$. In particular, for a fixed upper bound on $\gamma \cdot r(s)$, this yields a polynomial procedure for MCs and MDPs. Note that even if γt is not an integer, we could compute the optimal reachability precisely and then check whether this (rational) value is larger or smaller than the threshold by computing sufficiently many digits of $b^{-\gamma t}$.

6. Approximation algorithms

The results of the previous section suggest, depending on the form of the input, a polynomial-space algorithm or even worse in the general case. Clearly, this is somewhat unsatisfactory for practical applications. Recall that the difficulties are due to the occurring irrational transition probabilities. In the hope that we can work with approximations of these numbers, we now aim to identify an approach which allows us to approximate the correct answer, i.e. compute a value close to the optimal entropic risk that the Maximizer can ensure. Again, fix an SG G, reward function r, risk parameter γ, and risk basis b throughout this section. Then, given precision $\varepsilon > 0$, we aim to compute a value v such that $|\mathrm{ERisk}^* - v| < \varepsilon$, i.e. an approximation with small absolute error.

Since entropic risk is the logarithm of utility, we need to obtain an approximation of $\mathrm{NegUtil}^*$ to a sufficiently small *relative* error. Concretely, we need to compute a value v_U such that $b^{-\gamma\varepsilon} \leq v_U / \mathrm{NegUtil}^* \leq b^{\gamma\varepsilon}$. Then, $v = -\frac{1}{\gamma}\log_b(v_U)$ yields an approximation, since

$$\mathrm{ERisk}^* - v = -\frac{1}{\gamma}\log_b(\mathrm{NegUtil}^*) + \frac{1}{\gamma}\log_b(v_U) = \frac{1}{\gamma}\log_b(v_U / \mathrm{NegUtil}^*)$$

and so

$$\mathrm{ERisk}^* - v \geq \frac{1}{\gamma}\log_b(b^{-\gamma\varepsilon}) = -\varepsilon \quad \text{and} \quad \mathrm{ERisk}^* - v \leq \frac{1}{\gamma}\log_b(b^{\gamma\varepsilon}) = \varepsilon.$$

(When we are interested in a concrete value for v, we need to determine v_U with a slightly higher precision and then approximate $\log_b(v_U)$ sufficiently.) Now, in order to approximate $\mathrm{NegUtil}^*$, we still need to deal with a system comprising potentially irrational transition probabilities. We argue that the occurring values $b^{-\gamma r(s)}$ can be "rounded" to a sufficient precision while keeping the overall relative error small.

Using techniques from [62], we will provide an effective way to compute a game $\mathsf{G}_\approx$, which behaves "similarly" to the reachability game G_R from Lemma 4.4. Once $\mathsf{G}_\approx$ is computed, we can employ classical solution methods, such as linear equation solving for MCs, linear programming for MDPs, or, e.g., quadratic programming for SGs leading to the following result:

Theorem 6.1. *In MCs and MDPs, the optimal value* ERisk^* *can be approximated up to an absolute error of* ε *in time polynomial in the size of the system,* $-\log(\varepsilon)$, $\log b$, $\gamma \cdot r_{\max}$, *and* $1/(\gamma \cdot r_{\min})$, *where* $r_{\max}$ *and* $r_{\min}$ *are the largest and smallest occurring non-zero rewards, respectively. For SGs, this is possible in polynomial space.*

In particular, for fixed b and γ, and bounded rewards (both from above and below), we obtain a PTIME solution for MC and MDP. In general, the procedure is exponential for SG. Alternatively, we can also apply different approaches such as value iteration [63].

Remark 6.2. We note the connection to the small algebraic case: The "limiting factor" in both cases is the (size of the) product of γ and the state rewards. If these are fixed or given in unary, respectively, the complexity of our proposed algorithms is significantly reduced.

Furthermore, recall that we do not assume γ or the transition probabilities to be rational. We only require that we can expand their binary representation to arbitrary precision. Then, we can conservatively approximate their logarithm to evaluate the required rounding precision and approximate the transition probabilities of $\mathsf{G}_{\approx}$ in the same way.

Robustness for structurally equivalent games As mentioned before, in the following we require some tools of [62]. Intuitively, we want to show that only slightly changing each transition of a game does not change its overall value too much. This allows us to prove that "rounding" irrational transition probabilities only incur a bounded error.

The exact definitions are rather technical, we present a summary of relevant concepts here and refer to [62] for a more complete picture. We note that the results of [62] apply to *concurrent* stochastic games with *parity* objective, which are a significant generalization of *turn-based* stochastic games with *reachability* objective. We rephrase the definitions relative to our model.

To begin, we introduce the notion of *structurally equivalent games*. Intuitively, this means that the states, actions, and supports of transitions are equivalent; in other words, the induced graphs are the same. For such equivalent games G_1 and G_2 one can define the *relative difference* $\mathrm{dist}_R(\mathsf{G}_1, \mathsf{G}_2)$, which refers to the largest quotient of the probabilities of two corresponding transitions minus 1. In turn, this distance bounds the difference in reachability values.

For a formal definition, fix two SGs $\mathsf{G}_1 = (S^1_{\max}, S^1_{\min}, A_1, \Delta_1)$ and $\mathsf{G}_2 = (S^2_{\max}, S^2_{\min}, A_2, \Delta_2)$. We say that these two games are *structurally equivalent* if they induce the same graph, formally $S^1_{\max} = S^2_{\max}$, $S^1_{\min} = S^2_{\min}$, $A_1 = A_2$, and $\mathrm{supp}(\Delta_1(s,a)) = \mathrm{supp}(\Delta_2(s,a))$ for all $s \in S^1_{\max} \cup S^1_{\min}$ and $a \in A_1(s)$. Since the set of states and actions is equal, we omit the subscripts in the following. Moreover, let $S = S^1_{\max} \cup S^2_{\min} = S^2_{\max} \cup S^2_{\min}$.

We define the *distance* between two structurally equivalent games as

$$\mathrm{dist}_R(\mathsf{G}_1, \mathsf{G}_2) \;:=\; \max\left\{\frac{\Delta_1(s,a,t)}{\Delta_2(s,a,t)}, \frac{\Delta_2(s,a,t)}{\Delta_1(s,a,t)} \;\middle|\; s \in S, a \in A(s), t \in \mathrm{supp}(\Delta_1(s,a,t))\right\} - 1.$$

Since the games are structurally equivalent, the fractions are always well defined. Moreover, the value is always non-negative and it is equal to zero if and only if the two games are equal.

With this, we can state our desired robustness result.

Lemma 6.3. *Let* G_1, G_2 *two structurally equivalent games together with a reachability objective* T*. Set* $0 \le d = \mathrm{dist}_R(\mathsf{G}_1, \mathsf{G}_2)$*. Then*

$$(1+d)^{-2|S|} \le \frac{\mathsf{Val}(\mathsf{G}_1, T)}{\mathsf{Val}(\mathsf{G}_2, T)} \le (1+d)^{2|S|}.$$

Proof. We modify proofs of [62] as follows.

A useful tool employed there is the *mean-discounted time*, which we also recall. For a state s, discount vector $\lambda : S \to \mathbb{R}$, and infinite path ρ, this refers to the discounted time the path is in that state, formally

$$\mathsf{MDT}(\lambda, s)(\rho) := \frac{\sum_{j=0}^{\infty}(\prod_{i=0}^{j} \lambda(\rho_i)) \cdot \mathbb{1}_s(\rho_j)}{\sum_{j=0}^{\infty}(\prod_{i=0}^{j} \lambda(\rho_i))}.$$

The starting point to proving the result in [62] then is to show that for Markov chains the expected mean-discounted time can be expressed by a rational function comprising polynomials of bounded degree. In consequence, it is shown that for two Markov chains which are "close" w.r.t. dist_R the difference between the expected mean-discounted time can be bounded, too. The result follows by the known result that reachability (and parity) can be obtained as limit of mean-discounted time by taking the values of λ to 1.

To obtain our result, we only need to adapt the proofs of [62] slightly. First, consider [62, Lem. 3]: During the proof, we get that

$$(1+d)^{-2|S|} \le \frac{\mathsf{Val}(\mathsf{M}_1, \mathsf{MDT}(\lambda, r))(s)}{\mathsf{Val}(\mathsf{M}_2, \mathsf{MDT}(\lambda, r))(s)} \le (1+d)^{2|S|},$$

where M_1 and M_2 are two structurally equivalent Markov chains and $d = \mathrm{dist}_R(\mathsf{M}_1, \mathsf{M}_2)$ their relative difference.

Continuing with [62, Thm. 4], we obtain through [62, Thm. 2] that the same inequality also holds for the value of parity objectives and, as a special case, for reachability. By then applying the reasoning of [62, Thm. 5], i.e. considering the Markov chain obtained by fixing two optimal memoryless deterministic strategies, the inequality transfers to games. □

Proof of Theorem 6.1 With these notions at hand, the remainder of this section presents the proof of Theorem 6.1. Let G_R the reachability game from Lemma 4.4. We define a new game $\mathsf{G}_{\approx}$, only multiplicatively changing the transition probabilities of G_R. In particular, we have $\Delta^{\approx}(s,a,t) = \Delta^R(s,a,t) \cdot (1+\delta_s) \in \mathbb{Q}$ and $|\delta_s|$ is small. We intuitively see that Lemma 6.3 is applicable in this case, and, by choosing this factor δ_s small enough, the value of the rounded game will not deviate too much from the original value. However, note that we have two competing goals here: We want to find a factor δ_s such that on the one hand $b^{-\gamma r(s)} \cdot (1+\delta_s)$ is rational and, in particular, has a sufficiently small denominator to be computationally viable, and, on the other hand, the obtained value is not changed too much.

Lemma 6.4. *Fix a precision requirement $\varepsilon > 0$, let $r_{\min}$ and $r_{\max}$ equal the minimal and maximal occurring non-zero rewards, respectively, $N = |S|$ the number of states, and $p_{\min}$ the smallest occurring non-zero transition probability.*

Then, there exists a rounded game $\mathsf{G}_\approx$ such that (i) the reachability probability in $\mathsf{G}_\approx$ relatively differs from G_R by at most $b^{\gamma\varepsilon}$ and (ii) all transition probabilities are rational quantities with a denominator of bit size

$$-\log_2 p_{\min} - \min(\log_2 b^{-\gamma r_{\max}}, \log_2(1 - b^{-\gamma r_{\min}})) - \log_2 \gamma - \log_2 \varepsilon + \log_2 N - \log_2 \log b.$$

Proof. We want to show that

$$b^{-\gamma\varepsilon} \leq \mathsf{Val}(\mathsf{G}_R, T)/\mathsf{Val}(\mathsf{G}_\approx, T) \leq b^{\gamma\varepsilon}.$$

By Lemma 6.3, this holds if the transition probabilities have a relative difference d with

$$b^{-\gamma\varepsilon} \leq (1+d)^{-2|S|} \quad \text{and} \quad (1+d)^{2|S|} \leq b^{\gamma\varepsilon}.$$

Or, rearranged, $1 + d \leq b^{\gamma\varepsilon/(2|S|)}$ (and $0 \leq d$). To ease notation, we define $z = \gamma\varepsilon/(2|S|)$.

Fix some state $s \in S$. Recall that the reachability game G_R features two kinds of transition probabilities in each state. First, transition probabilities of the original game multiplied by $b^{-\gamma r(s)}$, and second the transition to the introduced trap state, multiplying by $(1 - b^{-\gamma r(s)})$.

Let us focus on the first kind and suppose the transition probability is given by $b^{-\gamma r(s)} \cdot p$ and assume that $r(s) \neq 0$. We want to show that there exists a rational number with a sufficiently small representation in the neighborhood of this transition probability, i.e. in the interval $I = [b^{-\gamma r(s)} p \cdot b^{-z}, b^{-\gamma r(s)} p \cdot b^{z}]$. Such a number necessarily exists if this interval is sufficiently large. Thus, consider $b^z - b^{-z}$. By change of base, we obtain

$$b^z - b^{-z} = e^{z \cdot \log b} - e^{-z \cdot \log b} = 2\sinh(z \log b) \geq 2z \log b,$$

using that $\sinh(x) \geq x$. Thus, the interval has at least size $|I| \geq b^{-\gamma r(s)} p \cdot 2z \log b$. Now, if n satisfies $2^{-n} \leq |I|$, then for some m we have that $\frac{m}{2^n} \in I$. In other words, n is an upper bound on the bit size of the smallest denominator that can be found in that interval. Taking $\log_2$ and inserting z, we arrive at

$$n \geq -\log_2 b^{-\gamma r(s)} - \log_2 \gamma - \log_2 \varepsilon + \log_2 p + \log_2 |S| + 1 - \log_2 \log b.$$

Thus, I contains a rational number with a denominator of that bit size. For the other type of transition, observe that the interval of concern is centered around $1 - b^{-\gamma r}$ and we similarly get

$$n \geq -\log_2(1 - b^{-\gamma r(s)}) - \log_2 \gamma - \log_2 \varepsilon + \log_2 p + \log_2 |S| + 1 - \log_2 \log b.$$

Recall that we assumed $r(s) \neq 0$. If instead we have $r(s) = 0$ and p is rational, we are done, since the transition probabilities from this state in G_R are equal to those of the original game G, i.e. rational and of size $-\log_2 p$. In case $p \in \mathbb{R} \setminus \mathbb{Q}$, observe that we only need to round the first kind of transitions (since the second kind is zero). However, here we can apply the same reasoning and analogously obtain the first inequality, noting that $\log b^{-\gamma r(s)} = 0$ in this case.

Taking the maximum of all inequalities over all states yields the result. □

Observe that Theorem 6.1 follows directly: The representation size of $\mathsf{G}_\approx$ is polynomial in the given quantities. If G is an MC or MDP, so is $\mathsf{G}_\approx$, and this rounded system can be solved in PTIME by standard means, i.e. solving an equation system or linear program. This also yields a PSPACE algorithm by trying out all memoryless strategies and solving the induced system (recall that memoryless strategies are sufficient by Theorem 4.9).

Remark 6.5. For practical application, note that we do not (and cannot) compute the concrete factor $b^{-\gamma r(s)}$ or the rounding δ_s. Instead, we determine as many digits of $b^{-\gamma r(s)}$ as required by the inequality in the proof. This guarantees that an appropriate δ_s exists, or, in other words, the obtained value is within the required interval I. For "reasonable" encodings of b, γ, and $r(s)$, these digits can be efficiently obtained through standard means.

For example, with $|S| = 10{,}000$, $\varepsilon = 10^{-4}$, $p_{\min} = 0.01$, $b = 2$, $\gamma = 2$, $r_{\min} = 0$, and $r_{\max} = 5$, we require an approximation to 43 bit precision, which is easily achieved by regular IEEE 754 64 bit doubles and associated mathematical functions.

7. Conclusion

We applied the entropic risk to total rewards in SGs to capture risk-averse behavior in these games. The objective forces agents to achieve a good overall performance while keeping the chance of particularly bad outcomes small. We showed that SGs with the entropic risk as payoff function are determined and admit optimal MD-strategies. This reflects the time-consistency of entropic risk and makes entropic risk an appealing objective as, in contrast, the optimization of other risk-averse objective functions that have been studied on MDPs in the literature require strategies with large memory or complicated randomization.

Computationally, difficulties arise due to the involved exponentiation leading to irrational or even transcendental numbers. For the general case, we obtained decidability of the threshold problem only subject to Shanuel's conjecture while for purely rational inputs, the problem can be solved via a reduction to the existential theory of the reals. Additional restrictions on the encoding of the input allowed us to obtain better upper bounds. Further, we provided an approximation algorithm for the optimal value. For an overview of the results, see Table 1.

A question that is left open is whether the entropic risk threshold problem for algebraic instances of MCs can be solved more efficiently than by the polynomial-time reduction to the existential theory of the reals. This case constitutes a bottleneck in the complexity. Furthermore, we worked with non-negative rewards, which made a reduction from games with the entropic risk objective to reachability games possible. Dropping the restriction to non-negative rewards constitutes an interesting direction of future research, in which additional difficulties arise and a reduction to reachability is not possible any more. A further direction for future work is the experimental evaluation of the proposed algorithms to assess their practical applicability as well as to investigate the behavior of the resulting optimal strategies by implementing the overall approach into existing model checkers such as PRISM [64]. In particular, it might be interesting to investigate the "cost" of risk-awareness, namely how much the expected total reward of a risk-aware strategy differs from a purely expectation maximizing one on realistic systems. Additionally, due to the specific shape of the reachability game (usually, every action has a non-zero probability to reach an absorbing state), partial exploration approaches as described in [65] and implemented in PET [66] might further enhance practical applicability.

CRediT authorship contribution statement

Christel Baier: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Krishnendu Chatterjee:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Tobias Meggendorfer:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Jakob Piribauer:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Christel Baier reports financial support was provided by German Research Foundation. Krishnendu Chatterjee reports financial support was provided by European Research Council. Co-author serves in an editorial capacity for 'Information and Computation' - K.C. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Baier, K. Chatterjee, T. Meggendorfer, J. Piribauer, Entropic risk for turn-based stochastic games, in: 48th International Symposium on Mathematical Foundations of Computer Science (MFCS 2023), in: J. Leroux, S. Lombardy, D. Peleg (Eds.), Leibniz International Proceedings in Informatics (LIPIcs), vol. 272, Schloss Dagstuhl – Leibniz-Zentrum Für Informatik, Dagstuhl, Germany, 2023, pp. 15:1–15:16 [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.MFCS.2023.15.
[2] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley Series in Probability and Statistics, Wiley, 1994.
[3] L.S. Shapley, Stochastic games, Proc. Natl. Acad. Sci. 39 (10) (1953) 1095–1100.
[4] A. Condon, On algorithms for simple stochastic games, Adv. Comput. Complex. Theor. 13 (1990) 51–72.
[5] A. Condon, The complexity of stochastic games, Inf. Comput. 96 (2) (1992) 203–224.
[6] K. Chatterjee, K. Sen, T.A. Henzinger, Model-checking omega-regular properties of interval Markov chains, in: FoSSaCS, in: Lecture Notes in Computer Science, vol. 4962, Springer, 2008, pp. 302–317.
[7] M. Weininger, T. Meggendorfer, J. Kretínský, Satisfiability bounds for ω-regular properties in bounded-parameter Markov decision processes, in: CDC, IEEE, 2019, pp. 2284–2291.
[8] D.J. White, A survey of applications of Markov decision processes, J. Oper. Res. Soc. 44 (11) (1993) 1073–1096.
[9] D.J. White, Real applications of Markov decision processes, Interfaces 15 (6) (1985) 73–83.
[10] K. Chatterjee, T.A. Henzinger, A survey of stochastic ω-regular games, J. Comput. Syst. Sci. 78 (2) (2012) 394–413.
[11] J. Filar, K. Vrieze, Competitive Markov Decision Processes, Springer Science & Business Media, 2012.
[12] J. Paulsson, Summing up the noise in gene networks, Nature 427 (6973) (2004) 415–418.
[13] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, Y. Moreno, Discrete-time Markov chain approach to contact-based disease spreading in complex networks, Europhys. Lett. 89 (3) (Feb 2010).

[14] D.T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, J. Comput. Phys. 22 (4) (1976) 403–434.
[15] D.P. Bertsekas, J.N. Tsitsiklis, An analysis of stochastic shortest path problems, Math. Oper. Res. 16 (3) (1991) 580–595.
[16] L. Fu, L.R. Rilett, Expected shortest paths in dynamic and stochastic traffic networks, Transp. Res., Part B, Methodol. 32 (7) (1998) 499–516.
[17] V. Forejt, M.Z. Kwiatkowska, G. Norman, D. Parker, Automated verification techniques for probabilistic systems, in: SFM, in: Lecture Notes in Computer Science, vol. 6659, Springer, 2011, pp. 53–113.
[18] M. Randour, J. Raskin, O. Sankur, Variations on the stochastic shortest path problem, in: VMCAI, in: Lecture Notes in Computer Science, vol. 8931, Springer, 2015, pp. 1–18.
[19] F. Teichteil-Königsbuch, U. Kuter, G. Infantes, Incremental plan aggregation for generating policies in mdps, in: AAMAS, IFAAMAS, 2010, pp. 1231–1238.
[20] J. Piribauer, O. Sankur, C. Baier, The variance-penalized stochastic shortest path problem, in: ICALP, in: LIPIcs, vol. 229, Schloss Dagstuhl - Leibniz-Zentrum Für Informatik, 2022, pp. 129:1–129:19.
[21] J.A. Filar, L.C.M. Kallenberg, H. Lee, Variance-penalized Markov decision processes, Math. Oper. Res. 14 (1) (1989) 147–161.
[22] S. Mannor, J.N. Tsitsiklis, Mean-variance optimization in Markov decision processes, in: ICML, Omnipress, 2011, pp. 177–184.
[23] T. Brázdil, K. Chatterjee, V. Forejt, A. Kucera, Trading performance for stability in Markov decision processes, J. Comput. Syst. Sci. 84 (2017) 144–170.
[24] T. Meggendorfer, Risk-aware stochastic shortest path, in: AAAI, AAAI Press, 2022, pp. 9858–9867.
[25] J. Kretínský, T. Meggendorfer, Conditional value-at-risk for reachability and mean payoff in Markov decision processes, in: LICS, ACM, 2018, pp. 609–618.
[26] C. Haase, S. Kiefer, The odds of staying on budget, in: ICALP (2), in: Lecture Notes in Computer Science, vol. 9135, Springer, 2015, pp. 234–246.
[27] J. Piribauer, C. Baier, On Skolem-hardness and saturation points in Markov decision processes, in: ICALP, in: LIPIcs, vol. 168, Schloss Dagstuhl - Leibniz-Zentrum Für Informatik, 2020, pp. 138:1–138:17.
[28] H. Föllmer, A. Schied, Convex measures of risk and trading constraints, Finance Stoch. 6 (4) (2002) 429–447.
[29] H. Föllmer, T. Knispel, Entropic risk measures: coherence vs. convexity, model ambiguity and robust large deviations, Stoch. Dyn. 11 (02n03) (2011) 333–351.
[30] M. Brandtner, W. Kürsten, R. Rischau, Entropic risk measures and their comparative statics in portfolio selection: coherence vs. convexity, Eur. J. Oper. Res. 264 (2) (2018) 707–716.
[31] D.P. Bertsekas, J.N. Tsitsiklis, Neuro-Dynamic Programming, Optimization and Neural Computation Series, vol. 3, Athena Scientific, 1996.
[32] R.A. Howard, J.E. Matheson, Risk-sensitive Markov decision processes, Manag. Sci. 18 (7) (1972) 356–369.
[33] N. Bäuerle, U. Rieder, More risk-sensitive Markov decision processes, Math. Oper. Res. 39 (1) (2014) 105–120.
[34] S.C. Jaquette, A utility criterion for Markov decision processes, Manag. Sci. 23 (1) (1976) 43–49.
[35] G.B. Di Masi, L. Stettner, Risk-sensitive control of discrete-time Markov processes with infinite horizon, SIAM J. Control Optim. 38 (1) (1999) 61–78.
[36] E. Collins, Finite-horizon variance penalised Markov decision processes, OR Spektrum 19 (1) (1997) 35–39.
[37] C. Haase, S. Kiefer, M. Lohrey, Computing quantiles in Markov chains with multi-dimensional costs, in: LICS, IEEE Computer Society, 2017, pp. 1–12.
[38] M. Ummels, C. Baier, Computing quantiles in Markov reward models, in: FoSSaCS, in: Lecture Notes in Computer Science, vol. 7794, Springer, 2013, pp. 353–368.
[39] M. Randour, J. Raskin, O. Sankur, Percentile queries in multi-dimensional Markov decision processes, Form. Methods Syst. Des. 50 (2–3) (2017) 207–248, https://doi.org/10.1007/s10703-016-0262-7 [Online].
[40] E. Asarin, J. Cervelle, A. Degorre, C. Dima, F. Horn, V. Kozyakin, Entropy games and matrix multiplication games, in: 33rd Symposium on Theoretical Aspects of Computer Science (STACS 2016), in: N. Ollinger, H. Vollmer (Eds.), Leibniz International Proceedings in Informatics (LIPIcs), vol. 47, Schloss Dagstuhl – Leibniz-Zentrum Für Informatik, Dagstuhl, Germany, 2016, pp. 11:1–11:14 [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.STACS.2016.11.
[41] M. Akian, S. Gaubert, J. Grand-Clément, J. Guillaud, The operator approach to entropy games, Theory Comput. Syst. 63 (2019) 1089–1130.
[42] M. Akian, S. Gaubert, U. Naepels, B. Terver, Solving irreducible stochastic mean-payoff games and entropy games by relative Krasnoselskii-Mann iteration, in: 48th International Symposium on Mathematical Foundations of Computer Science (MFCS 2023), in: J. Leroux, S. Lombardy, D. Peleg (Eds.), Leibniz International Proceedings in Informatics (LIPIcs), vol. 272, Schloss Dagstuhl – Leibniz-Zentrum Für Informatik, Dagstuhl, Germany, 2023, pp. 10:1–10:15 [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.MFCS.2023.10.
[43] C. Baier, J. Katoen, Principles of Model Checking, MIT Press, 2008.
[44] P. Billingsley, Probability and Measure, John Wiley & Sons, 2008.
[45] K. Chatterjee, T.A. Henzinger, Value iteration, in: 25 Years of Model Checking, in: Lecture Notes in Computer Science, vol. 5000, Springer, 2008, pp. 107–138.
[46] T. Chen, V. Forejt, M.Z. Kwiatkowska, D. Parker, A. Simaitis, Automatic verification of competitive stochastic systems, Form. Methods Syst. Des. 43 (1) (2013) 61–92.
[47] S. Lang, Algebra, 3rd ed., Addison-Wesley, 1993.
[48] H. Föllmer, A. Schied, T. Lyons, Stochastic finance. An introduction in discrete time, Math. Intell. 26 (4) (2004) 67–68.
[49] H. Asienkiewicz, A. Jaśkiewicz, A note on a new class of recursive utilities in Markov decision processes, Appl. Math. 44 (2017) 149–161.
[50] D.A. Martin, Borel determinacy, Ann. Math. 102 (2) (1975) 363–371.
[51] A. Maitra, W. Sudderth, Stochastic games with Borel payoffs, in: Stochastic Games and Applications, Springer, 2003, pp. 367–373.
[52] S. Haddad, B. Monmege, Interval iteration algorithm for mdps and imdps, Theor. Comput. Sci. 735 (2018) 111–131.
[53] J. Eisentraut, E. Kelmendi, J. Kretínský, M. Weininger, Value iteration for simple stochastic games: stopping criterion and learning algorithm, Inf. Comput. 285 (2022) 104886.
[54] L. de Alfaro, T.A. Henzinger, Concurrent omega-regular games, in: LICS, IEEE Computer Society, 2000, pp. 141–154.
[55] S. Lang, Introduction to Transcendental Numbers, Addison-Wesley Publishing Company, 1966.
[56] A. Macintyre, A.J. Wilkie, On the decidability of the real exponential field, in: P. Odifreddi (Ed.), Kreiseliana. About and Around Georg Kreisel, A K Peters, 1996, pp. 441–467.
[57] L. van den Dries, A. Macintyre, D. Marker, The elementary theory of restricted analytic fields with exponentiation, Ann. Math. 140 (1) (1994) 183–205.
[58] I. Adler, P.A. Beling, Polynomial algorithms for linear programming over the algebraic numbers, Algorithmica 12 (6) (1994) 436–457.
[59] P.A. Beling, Exact algorithms for linear programming over algebraic extensions, Algorithmica 31 (4) (2001) 459–478.
[60] G. Karpilovsky, Topics in Field Theory, Elsevier, 1989.
[61] J. Edmonds, Systems of distinct representatives and linear algebra, J. Res. Natl. Bur. Stand. B 71 (4) (1967) 241–245.
[62] K. Chatterjee, Robustness of structurally equivalent concurrent parity games, in: International Conference on Foundations of Software Science and Computational Structures, Springer, 2012, pp. 270–285.
[63] J. Kretínský, T. Meggendorfer, M. Weininger, Stopping criteria for value iteration on stochastic games with quantitative objectives, in: LICS, 2023.
[64] M. Kwiatkowska, G. Norman, D. Parker, G. Santos, Prism-games 3.0: stochastic game verification with concurrency, equilibria and time, in: S.K. Lahiri, C. Wang (Eds.), Computer Aided Verification - 32nd International Conference, CAV 2020, Los Angeles, CA, USA, July 21-24, 2020, Proceedings, Part II, in: Lecture Notes in Computer Science, vol. 12225, Springer, 2020, pp. 475–487 [Online].

[65] T. Brázdil, K. Chatterjee, M. Chmelik, V. Forejt, J. Kretínský, M.Z. Kwiatkowska, D. Parker, M. Ujma, Verification of Markov decision processes using learning algorithms, in: F. Cassez, J. Raskin (Eds.), Automated Technology for Verification and Analysis - 12th International Symposium, ATVA 2014, Sydney, NSW, Australia, November 3-7, 2014, Proceedings, in: Lecture Notes in Computer Science, vol. 8837, Springer, 2014, pp. 98–114 [Online].

[66] T. Meggendorfer, PET - a partial exploration tool for probabilistic verification, in: A. Bouajjani, L. Holík, Z. Wu (Eds.), Automated Technology for Verification and Analysis - 20th International Symposium, ATVA 2022, Virtual Event, October 25-28, 2022, Proceedings, in: Lecture Notes in Computer Science, vol. 13505, Springer, 2022, pp. 320–326 [Online].