# Entropic Risk for Turn-Based Stochastic Games

**Christel Baier**
Technische Universität Dresden, Germany

**Krishnendu Chatterjee**
Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria

**Tobias Meggendorfer**
Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria
Technische Universität München, Germany

**Jakob Piribauer**
Technische Universität Dresden, Germany
Technische Universität München, Germany

## 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. In the most general case, the problem is decidable subject to Shanuel's conjecture. If all inputs are rational, the resulting threshold problem can be solved using algebraic numbers, leading to decidability via a polynomial-time reduction to the existential theory of the reals. Further restrictions on the encoding of the input allow the solution of the threshold problem in $\mathsf{NP} \cap \mathsf{coNP}$. Finally, an approximation algorithm for the optimal value of ERisk is provided.

**2012 ACM Subject Classification** Theory of computation → Logic and verification

**Keywords and phrases** Stochastic games, risk-aware verification

**Digital Object Identifier** 10.4230/LIPIcs.MFCS.2023.15

**Related Version** *Full Version*: https://arxiv.org/abs/2307.06611 [3]

**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.

## 1 Introduction

**Stochastic Models.** Formal analysis of stochastic models is ubiquitous across disciplines of science, such as computer science [4], biology [43], epidemiology [29], and chemistry [28], to name a few. In computer science, a fundamental stochastic model are Markov decision processes (MDPs) [46], which extend purely stochastic Markov chains (MCs) with non-determinism to represent an agent interacting with a stochastic environment. Stochastic games (SGs) [49, 18, 19] in turn generalize MDPs by introducing an adversary, modelling the case where two agents engage in adversarial interaction in the presence of a stochastic environment. Notably, SGs can also be used to conservatively model MDPs where transition probabilities are not known precisely [15, 53]. See also [46] and [14, 21] for further applications of MDPs and SGs.

48th International Symposium on Mathematical Foundations of Computer Science (MFCS 2023).
Editors: Jérôme Leroux, Sylvain Lombardy, and David Peleg; Article No. 15; pp. 15:1–15:16
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

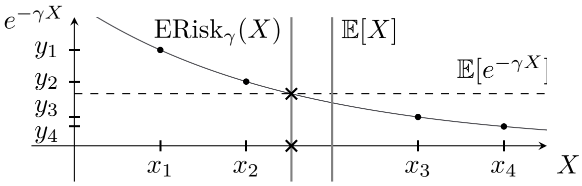

**Figure 1** Illustration of the entropic risk measure. The random variable $X$ takes values $x_1$ to $x_4$ uniformly with probability $\frac{1}{4}$ each. Expectation considers the average of $x_i$, while entropic risk yields the (normalized logarithm of the) average of $y_i = e^{-\gamma x_i}$.

**Strategies and Objectives.** In MDPs and SGs, the recipes to resolve choices are called strategies. The objective of the agent is to optimize a payoff function against all possible strategies of the adversary. One of the most fundamental problems studied in the context of MDPs and SGs is the optimization of total reward (and the related *stochastic shortest path* problem [7]). Here, every state (or, equivalently, transition) of the stochastic model is assigned a cost or reward and the payoff of a trajectory is the total sum of rewards appearing along the path. MDPs and SGs with total reward objectives provide an appropriate model to study a wide range of applications, such as traffic optimization [27], verification of stochastic systems [26, 47], or navigation / probabilistic planning [50].

**Risk-Ignorance of Expectation.** Typically, the expectation of the obtained total reward is optimized. However, the expectation measure is ignorant towards aspects of risk; an expectation maximizing agent accepts a one-in-a-million chance of extremely high rewards over a slightly worse, but guaranteed outcome. Such a behaviour 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.

**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” behaviour. To this end, several variants have been studied in the verification literature, such as (a) variance-penalized expected payoff [45, 22] that combines the expected value with a penalty for the variance of the resulting probability distribution; (b) trade-off of the expectation and variance for various notions of variance [40, 11]; (c) quantiles and conditional value-at-risk (CVaR) [47, 42, 35]; to name a few.

**Drawbacks.** The current approaches 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 [11, 40], while optimizing variance-penalized expected payoffs, quantiles, or the CVaR of the total reward require exponential memory [45, 30, 44, 42].

**Entropic Risk.** The notion of entropic risk [24] has been widely studied in finance and operation research, see e.g. [23, 10]. 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. We illustrate this in Figure 1. The exact definition of entropic risk is introduced later on.

**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: Compared to expectation, "bad" outcomes are penalized more than "good" outcomes add value. Thus, an agent optimizing entropic risk seeks to reduce the chances of particularly bad outcomes while also being interested in a good overall performance. 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 this case [8, 26]. 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 [33, 5]. 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 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 an 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 $\cap$ 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

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

| | threshold problem | | | optimal value | |
|---|---|---|---|---|---|
| | general instances (Thm. 14) | algebraic instances (Thm. 16) | small algebraic instances (Thm. 19) | computation for small algebraic instances (Thm. 20) | approximation with small rewards and risk aversion factor (Thm. 21) |
| SGs | decidable subject to Shanuel's conjecture | in PSPACE (in $\exists\mathbb{R}$) | in NP ∩ coNP | in polynomial space | in polynomial space |
| MDPs | | | in PTIME | | in polynomial time |
| MCs | | | | in polynomial time | |

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 [33] 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 [20], the entropic risk objective is considered for MDPs with a general Borel state space and in [5] 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 [17] and for infinite-horizon MDPs with discounted rewards and mean payoffs [22], and total rewards [45]. For total rewards, optimal strategies require exponential memory and the threshold problem is in NEXPTIME and EXPTIME-hard [45].

In [40], the optimization of expected accumulated rewards under constraints on the variance are studied for finite-horizon MDPs. Possible tradeoffs between expected value and variance of mean payoffs and other notions of variability have been studied in [11].

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 [30, 31]. 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 [51, 48]. Both of these problems have been addressed for MDPs with non-negative weights and are solvable in exponential time in this setting [51, 30]. Optimal strategies require exponential memory and the decision version of these problems is PSPACE-hard [30].

The conditional value-at-risk (CVaR), a prominent risk-measure, has been investigated for mean payoff and weighted reachability in MDPs in [35] as well as for total rewards in MDPs [44, 42]. 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 [44, 42]. 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 [44].

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

## 2 Preliminaries

In this section, we recall the basics of (turn-based) SGs and relevant objectives. For further details, see, e.g., [46, 4, 26, 21]. We assume familiarity with basic notions of probability theory (see, e.g., [9]). 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 $\mathrm{supp}(d) \coloneqq \{x \in X \mid d(x) > 0\}$. For a set $S$, $S^\star$ and $S^\omega$ refer to the set of finite and infinite sequences of elements of $S$, respectively.

### Markov Chains, MDPs, and Stochastic Games

A *Markov chain (MC)* (e.g. [4]), 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 $\mathsf{M}$ 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. [18]) is a tuple $(S_{\max}, S_{\min}, A, \Delta)$, where $S_{\max}$ and $S_{\min}$ are 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 $\Delta(s, a)\langle f\rangle \coloneqq \sum_{s' \in S} \Delta(s, a, s') \cdot f(s')$ for the weighted sum over the successors of $s$ under $a \in A(s)$. A *Markov decision process (MDP)* (e.g. [46]) 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 $\mathsf{G}$ by $\mathsf{Paths}_\mathsf{G}$. Furthermore, we write $\rho_i$ to denote the $i$-th state in the path $\rho$. *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 $\Delta$.

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 $\Pi_{\mathsf{G}}$ for the set of all strategies and $\Pi_{\mathsf{G}}^{\mathsf{MD}}$ 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 $\pi(s) \coloneqq \sigma(s)$ if $s \in S_{\max}$ and $\tau(s)$ otherwise.

Given a profile $\pi = (\sigma, \tau)$ of MD strategies for a game $\mathsf{G}$, we write $\mathsf{G}^\pi$ for the MC obtained by fixing both strategies. So, $\mathsf{G}^\pi = (S, \hat{\delta})$, where $\hat{\delta}(s) \coloneqq \Delta(s, \pi(s))$. Together with a state $s$, the MC $\mathsf{G}^\pi$ induces a unique probability distribution $\mathsf{Pr}_{\mathsf{G},s}^{\pi}$ 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 $\mathbb{E}_{\mathsf{G},s}^{\pi}[f]$ for the expected value of $f$ under the probability measure $\mathsf{Pr}_{\mathsf{G},s}^{\pi}$.

## 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 $\pi$ and a state $s$, the probability for this event is given by $\mathsf{Pr}_{\mathsf{G},s}^{\pi}[\Diamond T]$. On games, we are interested in determining the *value* $\mathsf{Val}_{\mathsf{G},\Diamond T}(s) \coloneqq \max_{\sigma \in \Pi_{\mathsf{G}}^{\mathsf{MD}}} \min_{\tau \in \Pi_{\mathsf{G}}^{\mathsf{MD}}} \mathsf{Pr}_{\mathsf{G},s}^{\sigma,\tau}[\Diamond T]$ 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 value coincide and the game is *determined*, i.e. the order of max and min does not matter [19]. Finally, we know that the value $\mathsf{Val}_{\mathsf{G},\Diamond T}$ is a solution of the following set of equations

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

where $S_0$ is the set of states that cannot reach $T$ against an optimal Minimizer strategy [13].

**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, $\mathrm{TR}(\rho) \coloneqq \sum_{i=1}^{\infty} r(\rho_i)$. 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 $\mathsf{Val}_{\mathsf{G},\mathbb{E}\,\mathrm{TR}}(s) \coloneqq \max_{\sigma \in \Pi_{\mathsf{G}}^{\mathsf{MD}}} \min_{\tau \in \Pi_{\mathsf{G}}^{\mathsf{MD}}} \mathbb{E}_{\mathsf{G},s}^{\sigma,\tau}[\mathrm{TR}]$. This game is determined and MD strategies suffice [16]. (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 $\langle\langle\{1\}\rangle\rangle \mathbf{R}_{\max=?}^{r}[\mathbf{F}^{c}\,\mathtt{ff}]$.)

# 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 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 $\gamma$) *(see, e.g., [25]) is defined as*

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

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.

▶ **Example 2.** 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. Figure 1 illustrates how the entropic risk measure of $X$ with base $e$ is obtained for some risk aversion factor $\gamma$: The values $x_i$ 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 example shows that deviations to lower values are penalized, i.e. taken into consideration more strongly, by this risk measure. For a different perspective, we can also consider the Taylor expansion of ERisk w.r.t. $\gamma$, which is $\mathrm{ERisk}_\gamma(X) = \mathbb{E}[X] - \frac{\gamma}{2} \cdot \mathrm{Var}[X] + \mathcal{O}(\gamma^2)$ (see, e.g., [2]). 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 $X = \mathrm{TR}$, i.e. optimizing the risk for total rewards. We write

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

to denote the entropic risk of the total reward achieved by the strategy profile $\pi$ 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 $b^{-\gamma\,\mathrm{TR}(\rho)} = b^{-\gamma\sum_{i=1}^{\infty} r(\rho_i)} = \prod_{i=1}^{\infty} b^{-\gamma r(\rho_i)}$ 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

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

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 $\mathsf{G}$, state $\hat{s}$, reward function $r$, risk parameter $\gamma$, risk basis $b$, and threshold $t$, decide whether there exists a Maximizer strategy $\sigma$ such that for all Minimizer strategies $\tau$ 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

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

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.** Consider the MDP of Figure 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 $\infty$ with probability $\frac{1}{2}$ each, while action `safe` yields 6 with probability 1. Let $b = 2$ and $\gamma = 1$ 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.

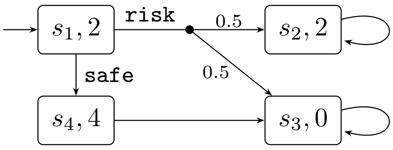

**Figure 2** Our running example to demonstrate several properties of entropic risk. For ease of presentation, the system actually is an MDP, where all states belong to Maximizer. States are denoted by boxes and their reward is written next to the state name. Transition probabilities are written next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

▶ Remark 4. As hinted above, entropic risk is finite whenever a finite reward is obtained with non-zero probability, i.e. for any strategy profile $\pi$, $\mathrm{ERisk}^{\gamma}_{\mathsf{G},\hat{s}}(\pi) = \infty$ iff $\mathrm{Pr}^{\pi}_{\mathsf{G},\hat{s}}[\mathrm{TR} = \infty] = 1$. 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 $\mathrm{Pr}^{\pi}_{\mathsf{G},\hat{s}}[\mathrm{TR} = \infty] > 0$. 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 $\pi$, 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 $\sigma$ 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.

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 $\mathsf{G}$, reward function $r$, state $\hat{s}$, risk parameter $\gamma$, and risk basis $b$. Omitted proofs can be found in the extended version [3].

## 4.1 Determinacy and Optimality Equation

▶ **Lemma 5.** *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 [41], see [39] 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 NegUtil$^*$:

▶ **Lemma 6.** *The optimal utility* NegUtil$^*$ *is a solution of the following system of constraints:*

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

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

Unfortunately, NegUtil is not the unique or, at least, 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 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 7.** Again consider the example of Figure 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 $S_0 = \{s \mid \max_\sigma \min_\tau \mathsf{Pr}^{\sigma,\tau}_{\mathsf{G},s}[\mathrm{TR} > 0] = 0\}$ 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 NegUtil$^* = 0$, ERisk$^* = \infty$. Through these sets, we can connect optimizing the utility to a reachability objective.

▶ **Lemma 8.** *For any state $s$ in the game $\mathsf{G}$, the optimal utility* $\mathrm{NegUtil}^*$ *is equal to the minimal probability of reaching the set $S_0$ from $s$ in game $\mathsf{G}_R$, 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.*

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

▶ **Lemma 9.** *For any game $\mathsf{G}$ and (absorbing) reachability goal $T$, we have* $\mathsf{Val}_{\mathsf{G},\Diamond T}(s) = 1 - \mathrm{NegUtil}^*_{\mathsf{G}}(s)$ *with reward $r(s) = \mathbb{1}_T(s)$ and $\gamma = 1$.*

We highlight that this reduction from entropic risk games to reachability games is *not* an effective reduction in the computational sense, since $\mathsf{G}_R$ 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 10.** *The optimal utility* $\mathrm{NegUtil}^*$ *is the pointwise smallest solution of*

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

Yet, there might be multiple fixed points to the system of equations. This is to be expected, since already reachability on MDPs exhibits this problem [32]. We provide a discussion of these issues together with a sufficient condition for uniqueness in the extended version [3].

## 4.3 Strategy Complexity

By Lemma 8, the optimal negative exponential utility is achieved by reachability-optimal strategies in $\mathsf{G}_R$. With the known results on reachability [18], this yields:

▶ **Theorem 11.** *MD strategies are sufficient to optimize the negative exponential utility and thus also entropic risk. More precisely, for all SGs $\mathsf{G}$, there is an MD strategy $\sigma$ 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 12. 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) behaviour after "enough" went wrong [42].

## 4.4 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 $\mathsf{G}_R$ 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 $\mathsf{G}_R$). 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 $\mathsf{G}_R$ [19].

▶ **Lemma 13.** *The system of equations 7 has a solution if and only if* $\mathrm{ERisk}^* \geq t$*.*

## 4.5 Decidability Subject to Shanuel's Conjecture

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

▶ **Theorem 14.** *Let all quantities, i.e. rewards, transition probabilities, the risk-aversion factor $\gamma$, 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. [37]) as shown by [38], 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 $\gamma$ 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 [52] 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. A detailed exposition of the results can be found in the extended version [3]; an overview of the complexity results can also be found in Table 1. Formally, we define:

▶ **Definition 15.** *An* algebraic instance *of the entropic risk threshold problem is an instance where all occurring values, i.e. the transition probabilities of the game $\mathsf{G}$, all rewards assigned by the reward function $r$, the risk-aversion parameter $\gamma$, 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 16.** *For algebraic instances, the entropic risk threshold problem is decidable in $\exists\mathbb{R}$ and thus in PSPACE.*

Already for Markov chains, it is unclear whether the upper complexity bound can be improved. For a discussion on this issue, see also the extended version [3].

For Theorem 16, 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., [1, 6]). 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 17.** *A* small algebraic instance *of the entropic risk threshold problem consists of a SG* $\mathsf{G}$ *with rational transition probabilities, an integer reward function $r$, a rational risk-aversion parameter $\gamma$, a rational basis $b$, and a rational threshold $t$. Moreover, the rewards, $\gamma$, 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* $\mathsf{G}$ *is an MDP or a MC, we call the instance a small algebraic instance of an MDP or a MC.*

▶ Remark 18. 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 $\gamma$.

Relying on algorithms for explicit computations in algebraic numbers [1, 6], we obtain:

▶ **Theorem 19.** *For small algebraic instances, the entropic risk threshold problem: (a) belongs to* $\mathsf{NP} \cap \mathsf{coNP}$ *for SGs; and (b) can be solved in polynomial time for MDPs or MCs.*

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 result.

▶ **Theorem 20.** *For small algebraic instances, an explicit representation of* $\mathrm{NegUtil}^*$ *can be computed in: (a) polynomial time for MCs; and (b) in polynomial space for SGs and MDPs.*

## 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 $\mathsf{G}$, reward function $r$, risk parameter $\gamma$, and risk basis $b$ throughout this section. Then, given a 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

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

(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 [12], we provide an effective way to compute a game $\mathsf{G}_\approx$, which behaves "similarly" to the reachability game $\mathsf{G}_R$ from Lemma 8, in the extended version [3]. 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 an algorithm in polynomial space for SGs.

▶ **Theorem 21.** *In MCs and MDPs, the optimal value* $\mathrm{ERisk}^*$ *can be approximated up to an absolute error of* $\varepsilon$ *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 $\gamma$, 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 [36].

▶ Remark 22. We note the connection to the small algebraic case: The "limiting factor" in both cases is the (size of the) product of $\gamma$ and the state rewards. If these are fixed or given in unary, respectively, the complexity of our proposed algorithms is significantly reduced.

As a final note, recall that we do not assume $\gamma$ 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.

## 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 anymore. 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. 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.

## References

1 Ilan Adler and Peter A. Beling. Polynomial algorithms for linear programming over the algebraic numbers. *Algorithmica*, 12(6):436–457, 1994.

2 Hubert Asienkiewicz and Anna Jaśkiewicz. A note on a new class of recursive utilities in Markov decision processes. *Applicationes Mathematicae*, 44:149–161, 2017.

3 Christel Baier, Krishnendu Chatterjee, Tobias Meggendorfer, and Jakob Piribauer. Entropic risk for turn-based stochastic games, 2023. arXiv preprint. arXiv:2307.06611.

4 Christel Baier and Joost-Pieter Katoen. *Principles of model checking.* MIT Press, 2008.

5 Nicole Bäuerle and Ulrich Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

6 Peter A. Beling. Exact algorithms for linear programming over algebraic extensions. *Algorithmica*, 31(4):459–478, 2001.

7 Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Math. Oper. Res.*, 16(3):580–595, 1991.

8 Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-dynamic programming*, volume 3 of *Optimization and neural computation series.* Athena Scientific, 1996. URL: https://www.worldcat.org/oclc/35983505.

9 Patrick Billingsley. *Probability and measure.* John Wiley & Sons, 2008.

10 Mario Brandtner, Wolfgang Kürsten, and Robert Rischau. Entropic risk measures and their comparative statics in portfolio selection: Coherence vs. convexity. *European Journal of Operational Research*, 264(2):707–716, 2018.

11 Tomás Brázdil, Krishnendu Chatterjee, Vojtech Forejt, and Antonín Kucera. Trading performance for stability in Markov decision processes. *J. Comput. Syst. Sci.*, 84:144–170, 2017.

12 Krishnendu Chatterjee. Robustness of structurally equivalent concurrent parity games. In *International Conference on Foundations of Software Science and Computational Structures*, pages 270–285. Springer, 2012.

13 Krishnendu Chatterjee and Thomas A. Henzinger. Value iteration. In *25 Years of Model Checking*, volume 5000 of *Lecture Notes in Computer Science*, pages 107–138. Springer, 2008.

14 Krishnendu Chatterjee and Thomas A. Henzinger. A survey of stochastic $\omega$-regular games. *J. Comput. Syst. Sci.*, 78(2):394–413, 2012.

15 Krishnendu Chatterjee, Koushik Sen, and Thomas A. Henzinger. Model-checking omega-regular properties of interval Markov chains. In *FoSSaCS*, volume 4962 of *Lecture Notes in Computer Science*, pages 302–317. Springer, 2008.

16 Taolue Chen, Vojtech Forejt, Marta Z. Kwiatkowska, David Parker, and Aistis Simaitis. Automatic verification of competitive stochastic systems. *Formal Methods Syst. Des.*, 43(1):61–92, 2013. doi:10.1007/s10703-013-0183-7.

17 E.J. Collins. Finite-horizon variance penalised Markov decision processes. *Operations-Research-Spektrum*, 19(1):35–39, 1997.

18 Anne Condon. On algorithms for simple stochastic games. *Advances in computational complexity theory*, 13:51–72, 1990.

19 Anne Condon. The complexity of stochastic games. *Information and Computation*, 96(2):203–224, 1992.

20 Giovanni B Di Masi and Lukasz Stettner. Risk-sensitive control of discrete-time Markov processes with infinite horizon. *SIAM Journal on Control and Optimization*, 38(1):61–78, 1999.

21 Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes.* Springer Science & Business Media, 2012.

22 Jerzy A. Filar, Lodewijk C. M. Kallenberg, and Huey-Miin Lee. Variance-penalized Markov decision processes. *Math. Oper. Res.*, 14(1):147–161, 1989.

23 Hans Föllmer and Thomas Knispel. Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations. *Stochastics and Dynamics*, 11(02n03):333–351, 2011.

24 Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and stochastics*, 6(4):429–447, 2002.

25 Hans Föllmer, Alexander Schied, and T Lyons. Stochastic finance. an introduction in discrete time. *The Mathematical Intelligencer*, 26(4):67–68, 2004.

26 Vojtech Forejt, Marta Z. Kwiatkowska, Gethin Norman, and David Parker. Automated verification techniques for probabilistic systems. In *SFM*, volume 6659 of *Lecture Notes in Computer Science*, pages 53–113. Springer, 2011.

27 Liping Fu and Larry R Rilett. Expected shortest paths in dynamic and stochastic traffic networks. *Transportation Research Part B: Methodological*, 32(7):499–516, 1998.

28 Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

29 S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno. Discrete-time Markov chain approach to contact-based disease spreading in complex networks. *EPL (Europhysics Letters)*, 89(3), February 2010.

30 Christoph Haase and Stefan Kiefer. The odds of staying on budget. In *ICALP (2)*, volume 9135 of *Lecture Notes in Computer Science*, pages 234–246. Springer, 2015.

31 Christoph Haase, Stefan Kiefer, and Markus Lohrey. Computing quantiles in Markov chains with multi-dimensional costs. In *LICS*, pages 1–12. IEEE Computer Society, 2017.

32 Serge Haddad and Benjamin Monmege. Interval iteration algorithm for mdps and imdps. *Theor. Comput. Sci.*, 735:111–131, 2018.

33 Ronald A Howard and James E Matheson. Risk-sensitive Markov decision processes. *Management science*, 18(7):356–369, 1972.

34 Stratton C Jaquette. A utility criterion for Markov decision processes. *Management Science*, 23(1):43–49, 1976.

35 Jan Kretínský and Tobias Meggendorfer. Conditional value-at-risk for reachability and mean payoff in Markov decision processes. In *LICS*, pages 609–618. ACM, 2018.

36 Jan Kretínský, Tobias Meggendorfer, and Maximilian Weininger. Stopping criteria for value iteration on stochastic games with quantitative objectives. In *LICS*, 2023.

37 Serge Lang. *Introduction to transcendental numbers.* Addison-Wesley Publishing Company, 1966.

**38** Angus Macintyre and Alex J. Wilkie. On the decidability of the real exponential field. In Piergiorgio Odifreddi, editor, *Kreiseliana. About and Around Georg Kreisel*, pages 441–467. A K Peters, 1996.

**39** A Maitra and W Sudderth. Stochastic games with borel payoffs. In *Stochastic Games and Applications*, pages 367–373. Springer, 2003.

**40** Shie Mannor and John N. Tsitsiklis. Mean-variance optimization in Markov decision processes. In *ICML*, pages 177–184. Omnipress, 2011.

**41** Donald A Martin. Borel determinacy. *Annals of Mathematics*, 102(2):363–371, 1975.

**42** Tobias Meggendorfer. Risk-aware stochastic shortest path. In *AAAI*, pages 9858–9867. AAAI Press, 2022.

**43** Johan Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–418, 2004.

**44** Jakob Piribauer and Christel Baier. On skolem-hardness and saturation points in Markov decision processes. In *ICALP*, volume 168 of *LIPIcs*, pages 138:1–138:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

**45** Jakob Piribauer, Ocan Sankur, and Christel Baier. The variance-penalized stochastic shortest path problem. In *ICALP*, volume 229 of *LIPIcs*, pages 129:1–129:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

**46** Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley Series in Probability and Statistics. Wiley, 1994.

**47** Mickael Randour, Jean-François Raskin, and Ocan Sankur. Variations on the stochastic shortest path problem. In *VMCAI*, volume 8931 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 2015.

**48** Mickael Randour, Jean-François Raskin, and Ocan Sankur. Percentile queries in multi-dimensional Markov decision processes. *Formal Methods Syst. Des.*, 50(2-3):207–248, 2017. doi:10.1007/s10703-016-0262-7.

**49** Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

**50** Florent Teichteil-Königsbuch, Ugur Kuter, and Guillaume Infantes. Incremental plan aggregation for generating policies in mdps. In *AAMAS*, pages 1231–1238. IFAAMAS, 2010.

**51** Michael Ummels and Christel Baier. Computing quantiles in Markov reward models. In *FoSSaCS*, volume 7794 of *Lecture Notes in Computer Science*, pages 353–368. Springer, 2013.

**52** Lou van den Dries, Angus Macintyre, and David Marker. The elementary theory of restricted analytic fields with exponentiation. *Annals of Mathematics*, 140(1):183–205, 1994.

**53** Maximilian Weininger, Tobias Meggendorfer, and Jan Kretínský. Satisfiability bounds for $\omega$-regular properties in bounded-parameter Markov decision processes. In *CDC*, pages 2284–2291. IEEE, 2019.