Solvency Markov Decision Processes with Interest

Tomáš Brázdil*[1], Taolue Chen[2], Vojtěch Forejt†[2], Petr Novotný[1], and Aistis Simaitis[2]

1 Faculty of Informatics, Masaryk University, Czech Republic
2 Department of Computer Science, University of Oxford, UK

Abstract

Solvency games, introduced by Berger et al., provide an abstract framework for modelling decisions of a risk-averse investor, whose goal is to avoid ever going broke. We study a new variant of this model, where, in addition to stochastic environment and fixed increments and decrements to the investor's wealth, we introduce interest, which is earned or paid on the current level of savings or debt, respectively.

We study problems related to the minimum initial wealth sufficient to avoid bankruptcy (i.e. steady decrease of the wealth) with probability at least p. We present an exponential time algorithm which approximates this minimum initial wealth, and show that a polynomial time approximation is not possible unless P = NP. For the qualitative case, i.e. $p = 1$, we show that the problem whether a given number is larger than or equal to the minimum initial wealth belongs to NP $\cap$ coNP, and show that a polynomial time algorithm would yield a polynomial time algorithm for mean-payoff games, existence of which is a longstanding open problem. We also identify some classes of solvency MDPs for which this problem is in P. In all above cases the algorithms also give corresponding bankruptcy avoiding strategies.

1998 ACM Subject Classification G.3 Probability and statistics.

Keywords and phrases Markov decision processes, algorithms, complexity, market models.

1 Introduction

Markov decision processes (MDP) are a standard model of complex decision-making where results of decisions may be random. An MDP has a set of *states*, where each state is assigned a set of *enabled actions*. Every action determines a distribution on the set of successor states. A run starts in a state; in every step, a *controller* chooses an enabled action and the process moves to a new state chosen randomly according to the distribution assigned to the action. The functions that describe decisions of the controller are called *strategies*. They may depend on the whole history of the computation and the choice of actions may be randomized.

MDPs form a natural model of decision-making in the financial world. To model nuances of financial markets, various MDP-based models have been developed (see e.g. [15, 2, 3]). A common property of these models is that actions correspond to investment choices and result in (typically random) payoffs for the controller. One of the common aims in this area is to find a *risk-averse* controller (investor) who strives to avoid undesirable events [12, 13].

In this paper we consider a model based on standard reward structures for MDPs, which is closely related to solvency games studied in [3]. The model is designed so that it captures essential properties of risk-averse investments. We assume finite-state MDPs and assign a (real) reward to every action which is collected whenever the action is chosen. The states

[*] Tomáš Brázdil is supported by the Czech Science Foundation, grant No P202/12/P612.
[†] Vojtěch Forejt is also affiliated with FI MU, Brno, Czech Republic.

of the MDP capture the global situation on the market, prices of assets, etc. Note that it is usually plausible to model the prices by a finite-state stochastic process (see e.g. [15]). Rewards model money received (positive rewards) and money spent (negative rewards) by the controller. Controllers are then compared w.r.t. their ability to collect the reward over finite or infinite runs.

Standard objectives such as the *total reward*, or the *long-run average reward* are not suitable for modelling the behaviour of a risk-averse investor as they allow temporary loss of an arbitrary amount of money (i.e., a long sequence of negative rewards), which is undesirable, because normally the controller's access to credit is limited. The authors of [3] consider a "bankruptcy-avoiding" objective defined as follows: Starting with an initial amount of wealth W_0, in the n-th step, the current wealth W_n is computed from W_{n-1} by adding the reward collected in the n-th step. The goal is to find a controller which maximizes the probability of having $W_n > 0$ for all n.

Although the model of [3] captures basic behaviour of a risk-averse investor, it lacks one crucial aspect usually present in the financial environment, i.e., the *interest*. Interests model the value that is received from holding a certain amount of cash, or conversely, the cost of having a negative balance. To accommodate interests, we propose the following extension of the bankruptcy-avoiding objective: Fix an interest rate $\varrho > 1$.[1] Starting with an initial wealth W_0, in the n-th step, compute the current wealth W_n from W_{n-1} by adding not only the collected reward but also the interest $(\varrho - 1)W_{n-1}$. The economical motivation for such a model is that the controller can earn additional amount of wealth by lending its assets for a fixed interest, and conversely, when the controller is in debt, it has to pay interest to its creditors (for the clarity of presentation, we suppose the interest earned from positive wealth is the same as the interest paid on debts).

Hence, the objective is to "manage" the wealth so that it stays above some threshold and does not keep decreasing to negative infinity. More precisely, we want to maximize the probability of having $\liminf_{n\to\infty} W_n > -\infty$. Intuitively, $\liminf_{n\to\infty} W_n \geq 0$ means that the controller ultimately does not need to borrow money, and $-\infty < \liminf_{n\to\infty} W_n < 0$ means that the controller is able to sustain interest payments from its income. If $\liminf_{n\to\infty} W_n = -\infty$, then the controller cannot sustain interest payments and bankrupts.

An important observation is that this objective is closely related to another well-studied objective concerning the *discounted total reward*. Concretely, given a discount factor $0 < \beta < 1$, the discounted total reward T accumulated on a run is defined to be the weighted sum of rewards of all actions on the run where the weight of the n-th action is β^n. In particular, the *threshold problem* asks to maximize the probability of $T \geq t$ for a given threshold t. This problem has been considered in, e.g., [16, 10, 17, 18]. A variant of the threshold problem is the *value-at-risk* problem [4] which asks, for a given probability p, what is the infimum threshold, such that maximal probability of discounted reward surpassing the threshold is at least p? We show that for every controller, the probability of $T \geq t$ with discount factor β is equal to the probability of $\liminf_{n\to\infty} W_n > -\infty$ with $W_0 = -t$ for the interest rate $\varrho := \frac{1}{\beta}$. This effectively shows interreducibility of these problems. Note that the interpretation of the discount factor as the inverse of the interest is natural in financial mathematics.

Contribution. We introduce a model of solvency MDPs with interests (referred to as *solvency MDPs* for brevity), which allows to capture the complex dynamics of wealth management under uncertainty. We show that for every solvency MDP there is a bound on

[1] For notational convenience, we define the interest rate to be the number $1 + r$, where $r > 0$ is the usual interest rate, i.e. the percentage of money paid/received over a unit of time.

wealth such that above this bound the bankruptcy is surely avoided (no matter what the controller is doing), and another bound on wealth below which the bankruptcy is inevitable. Nevertheless, we also show that there still might be infinitely many reachable values of wealth between these two bounds.

The main results of our paper concentrate on the complexity of computing minimal wealth with which the controller can stay away from bankruptcy. Let $\mathbf{W}(s_0, p)$ be the *infimum* of all initial wealths W_0 such that starting in the state s_0 with W_0 the controller can avoid bankruptcy (i.e., $\liminf_{n\to\infty} W_n > -\infty$) with probability at least p. Our overall goal is to compute this number $\mathbf{W}(s_0, p)$. Solution to this problem is important for a risk-averse investor, whose aim is to keep the risk of bankruptcy below some acceptable level.

First we consider the *qualitative case*, i.e. $\mathbf{W}(s_0, 1)$. For this case we show a connection with two-player (non-stochastic) games with discounted total reward objectives. Then, using the results of [19] we show that there is an *oblivious* strategy (i.e., the one that looks only at the current state but is independent of the wealth accumulated so far) which starting in some state s_0 with wealth $\mathbf{W}(s_0, 1)$ avoids bankruptcy with probability one. The problem whether $W \geq \mathbf{W}(s_0, 1)$ for a given W (encoded in binary) is in $NP \cap coNP$ (we also obtain a reduction from discounted total reward games, showing that improving this complexity bound might be difficult). In addition, the number $\mathbf{W}(s_0, 1)$ can be computed in pseudo-polynomial time. Further it follows that for a restricted class of solvency Markov chains (i.e. when there is only one enabled action in every state) the value $\mathbf{W}(s_0, 1)$ can be computed in polynomial time.

The main part of our paper concerns the *quantitative case*, i.e. $\mathbf{W}(s_0, p)$ for an arbitrary probability bound p.

- We give an exponential-time algorithm that approximates $\mathbf{W}(s_0, p)$ up to a given absolute error $\varepsilon > 0$. We actually show that the algorithm runs in time polynomial in the number of control states and exponential in $\log(1/(\varrho - 1))$, $\log(1/\varepsilon)$ and $\log(r_{\max})$, where ϱ is the interest rate and $r_{\max}$ is the maximal $|r|$ where r is a reward associated to some action.
- Employing a reduction from the Knapsack problem, we show that the above complexity cannot be lowered to polynomial in either $\log(1/\varepsilon)$ or $\log(\varrho - 1) + \log(r_{\max})$ unless P=NP.
- We give an exponential-time algorithm that for a given $\varepsilon > 0$ and initial wealth W_0 computes v such that if the initial wealth is increased by ε, then the probability of avoiding bankruptcy is at least v (i.e. $W_0 + \varepsilon \geq \mathbf{W}(s_0, v)$) and $v \geq \sup\{v' \mid W_0 \in \mathbf{W}(s_0, v')\}$.

Moreover, via the aforementioned interreducibility between discounted and solvency MDPs we establish new complexity bounds for value-at-risk approximation in discounted MDPs.

We note that the aforementioned algorithms employ a careful rounding of numbers representing the current wealth W_n. Choosing the right precision for this rounding is quite an intricate step, since a naive choice would only yield a doubly-exponential algorithm.

The paper is organized as follows: after introducing necessary definitions and clarifying the relation with the discounted MDPs in Section 2 we summarise the results for qualitative problem in Section 3. In Section 4 we give the contributions for the quantitative problem.

Related work. Processes involving interests and their formal models naturally emerge in the field of financial mathematics. An MDP-based model of a financial market is presented, e.g., in Chapter 3 of [2]. There, in every step the investor has to allocate his current wealth between riskless bonds, on which he receives an interest according to some fixed interest rate, and several risky stocks, whose price is subject to random fluctuations. Optimization of the investor's portfolio with respect to various utility measures was studied. However, this portfolio optimization problem was considered only in the finite-horizon case, where the trading stops after some fixed number of steps. In contrast, we concentrate on the long-term stability of the investor's wealth. Also, the model in [2] was analysed mainly from the mathematical perspective (e.g., characterizing the form of optimal portfolios), while we focus on an efficient algorithmic computation of the optimal investor's behaviour.

The issues of a long-term stability and algorithms were considered for other related models, all of which concern total accumulated reward properties. Our model is especially close to *solvency games* [3], which are in fact MDPs with a single control state, where the investor aims to keep the total accumulated reward non-negative. In *energy games* (see e.g. [7, 8, 9]), there are two competing players, but no stochastic behaviour. In *one-counter* MDPs [6], the counter can be seen as a storage for the current value of wealth. All these models differ from the topic studied in this paper in that they do not consider interest on wealth. This makes them fundamentally different in terms of their properties, e.g. in our setting the set of all wealths reachable from a given initial wealth can have nontrivial limit points. Also, in all the three aforementioned models, the objective is to stay in the positive wealth. Here we focus on a different objective to capture the idea that it is admissible to be in debt as long as it is possible to maintain the debt above some limit.

As mentioned before, our work is also related to the threshold discounted total reward objectives, which were considered in [16, 10, 17, 18], where the authors studied finite- and infinite-horizon cases. In the finite-horizon case, in particular [18] gave an algorithm to compute the probability, but a careful analysis shows that their algorithm has a doubly-exponential worst-case complexity when the planning horizon (i.e., the number of steps after which the process halts) is encoded in binary. In [5] they proposed to approximate the probability through the discretisation of wealth, but in the worst the error of approximation is 1, no matter how small discretisation step is taken. In [18], the optimality equation characterising optimal probabilities has been provided for the infinite-horizon case, but no algorithm was proposed. Moreover, [4] considered the "value-at-risk" problem, but again only for the finite-horizon case, giving a doubly-exponential approximation algorithm. Although we consider only infinite-horizon MDPs, the exponential-time upper bound for the $\mathbf{W}(s,p)$ approximation and the NP-hardness lower bound can be easily carried over to the finite-horizon case. Thus, we establish new complexity bounds for value-at-risk approximation in both finite and infinite-horizon discounted MDPs. We also mention [11] which introduced the percentile performance criteria where the controller aims to find a strategy achieving a specified value of the long-run limit average reward at a specified probability level (percentile).

2 Preliminaries

We denote by $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ the sets of all natural, integer, rational and real numbers, respectively. For an index set I, its member i and vector $\mathbf{V} \in \mathbb{R}^I$ we denote by $\mathbf{V}(i)$ the i-component of $\mathbf{V}$. The encoding size of an object B is denoted by $||B||$. We use $\log x$ to refer to the binary logarithm of x. We assume that all numbers are represented in binary and that rational numbers are represented as fractions of binary-encoded integers.

We assume familiarity with basic notions of probability theory. Given an at most countable set X, we use $\mathit{dist}(X)$ to denote all probability distributions on X.

▶ **Definition 1** (MDP). A *Markov decision process* (MDP) is a tuple $M = (V, A, T)$ where V is at most countable set of *vertices*, A is a finite set of *actions*, and $T : V \times A \to \mathit{dist}(V)$ is a partial *transition function*. We assume that for every $v \in V$ the set $A(v)$ of all actions available at v (i.e., the set off all actions a s.t. $T(v, a)$ is defined) is nonempty.

We denote by $\mathit{Succ}(v, a) = \{u \mid T(v, a)(u) > 0\}$ the *support* of $T(v, a)$. A *Markov chain* is an MDP with one action per vertex, i.e., $|A(v)| = 1$ for all $v \in V$.

From a given initial vertex $v_0 \in V$ the MDP evolves as follows. An *infinite path* (or *run*) is a sequence $v_0 a_1 v_1 a_2 v_2 \cdots \in (V \times A)^\omega$ such that $a_{i+1} \in A(v_i)$ and $v_{i+1} \in \mathit{Succ}(v_i, a_{i+1})$ for all i. A *finite path* (or *history*) is a prefix of a run ending with a vertex, i.e. a word of the form $(V \times A)^* V$. We refer to the set of all runs as Runs_M and to the set of all histories as Hist_M. For a finite or infinite path $\omega = v_0 a_1 v_1 a_2 v_2 \ldots$ and $i \in \mathbb{N}$ we denote by ω_i the finite path $v_0 a_1 \cdots a_i v_i$.

A *strategy* in M is a function that to every history w assigns a distribution on actions available in the last vertex of w. A strategy is *deterministic* if it always assigns distributions that choose some action with probability 1, and *memoryless* if it only depends on the last vertex of history. We use Σ_M (or just Σ) for the set of all strategies of M.

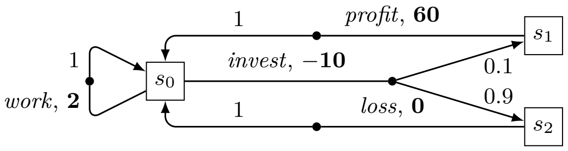
Each history $w \in \mathsf{Hist}_M$ determines the set $\mathsf{Cone}(w)$ consisting of all runs having w as a prefix. To an MDP M, its vertex v and strategy σ we associate the probability space $(\mathsf{Runs}_M, \mathcal{F}, \mathbb{P}^\sigma_{M,v})$, where $\mathcal{F}$ is the σ-field generated by all $\mathsf{Cone}(w)$, and $\mathbb{P}^\sigma_{M,v}$ is the unique probability measure such that for every history $w = v_0 a_1 \ldots a_k v_k$ we have $\mathbb{P}^\sigma_{M,v}(\mathsf{Cone}(w)) = \mu(v_0) \cdot \prod_{i=1}^k x_i$, where $\mu(v_0)$ is 1 if $v_0 = v$ and 0 otherwise, and where $x_i = \sigma(w_{i-1})(a_i) \cdot T(v_{i-1}, a_i)(v_i)$ for all $1 \leq i \leq k$ (the empty product is equal to 1). We drop M from the subscript when the MDP is clear from the context.

▶ **Definition 2** (Solvency MDP). A *solvency Markov decision process* is a tuple (S, A, T, F, ϱ) where S is a finite set of *states*, A and T are such that (S, A, T) is an MDP, $F : S \times A \to \mathbb{Q}$ is a partial *gain function* and $\varrho \in \mathbb{Q} \cap (1, \infty)$ is an *interest rate*.

We stipulate that for every $(s, a) \in S \times A$ the value $F(s, a)$ is defined iff $a \in A(s)$. A *solvency Markov chain* is a solvency MDP with one action per state, i.e. $|A(s)| = 1$ for all $s \in S$. A *configuration* of a solvency MDP $M = (S, A, T, F, \varrho)$ is represented as a state-wealth pair (s, x) where $s \in S$ and $x \in \mathbb{Q}$. The semantics of M is given by an infinite-state MDP $M_\varrho = (S \times \mathbb{Q}, A, T_\varrho)$ where for every $(s, x) \in S \times \mathbb{Q}$ and $a \in A(s)$ we define $T_\varrho((s, x), a)(s', \varrho \cdot x + F(s, a)) = p$ whenever $T(s, a)(s') = p$. We sometimes do not distinguish between M and M_ϱ and refer to strategies or runs of M where strategies or runs of M_ϱ are intended. A strategy σ for M_ϱ is *oblivious* if it is memoryless and does not make its decision based on the current wealth, i.e. for all $w \cdot (s, x)$ and (s, x') we have $\sigma(w \cdot (s, x)) = \sigma((s, x'))$.

Objectives. Given an solvency MDP M and its initial configuration (s_0, x_0), we are interested in the set of runs in which the wealth always stays above some finite bound, denoted by $\mathit{Win} = \mathsf{Runs}_M \setminus \{(s_0, x_0) a_1 (s_1, x_1) \cdots \in \mathsf{Runs}_M \mid \liminf_{n \to \infty} x_n = -\infty\}$. Intuitively, this objective models the ability of the investor not to go bankrupt, i.e. to compensate for the incurred interest by obtaining sufficient gains. We denote $\mathit{Val}_M(s_0, x_0) = \sup_\sigma \mathbb{P}^\sigma_{M,(s_0,x_0)}(\mathit{Win})$ the maximal probability of winning with a given wealth, and $\mathbf{W}_M(s, p) = \inf\{x \mid \mathit{Val}_M(s, x) \geq p\}$ the infimum of wealth sufficient for winning with probability p. In this paper we are mainly interested in the problems of computing or approximating the values of $\mathbf{W}_M(s, p)$. We also address the problem of computing a convenient risk-averse strategy for an investor with a given initial wealth x_0. A precise definition of what we mean by a convenient strategy is given in Section 4 (Theorem 11). We say that a strategy is p-winning (in an initial configuration (s_0, x_0)) if $\mathbb{P}^\sigma_{M,(s_0,x_0)}(\mathit{Win}) \geq p$. A 1-winning strategy is called *almost surely winning*, and strategy σ with $\mathbb{P}^\sigma_{M,(s_0,x_0)}(\mathit{Win}) = 0$ is called *almost surely losing*.

▶ **Example 3.** Consider the following solvency MDP $M = (S, A, T, F, \varrho)$:



Here $S = \{s_0, s_1, s_2\}$, $A = \{\mathit{work}, \mathit{invest}, \mathit{profit}, \mathit{loss}\}$, T is depicted by the arrows in the figure, for example $T(s_0, \mathit{invest}) = [s_1 \mapsto 0.1, s_2 \mapsto 0.9]$, the function F is given by the bold numbers next to the actions, e.g. $F(s, \mathit{work}) = 2$, and $\varrho = 2$ (we take this extremely large value to keep the example computations simpler). The MDP models the choices of a person who can either work, which ensures certain but relatively small income, or can invest a larger amount of money but take a significant risk. Starting in the configuration $(s_0, -10)$ (i.e. in debt), an example strategy σ is the strategy which always chooses *work* in s_0, but as can be easily seen, we get $\mathbb{P}^{\sigma}_{M,(s_0,-10)}(\mathit{Win}) = 0$ since the constant gains are not high enough to cover the interest incurred by the debt. An optimal strategy here is to pick *work* only in histories ending with a configuration (s_0, x) for $x \geq -2$, and to pick *invest* otherwise. Such strategy shows that $\mathit{Val}_M(s_0, -10) = 0.1$. Now suppose that the investor wants to find out what is the wealth needed to make sure the probability of winning is at least 0.7, i.e. wants to compute $\mathbf{W}_M(s_0, 0.7)$. This number is equal to -2. To see this, observe that for any configuration (s_0, y) where $y < -2$ the optimal strategy must pick *invest*, which with probability 0.9 results in a debt from which it is impossible to recover. Finally, observe that $\mathit{Val}_M(s_0, -2) = 1$ since a strategy that always chooses *work* is 1-winning in $(s_0, -2)$. This demonstrates that the function $\mathit{Val}(s, \cdot)$ for a given state s may not be continuous.

Relationship with discounted MDPs. The problems we study for solvency MDPs are closely related to another risk-averse decision making model, so called *discounted MDPs with threshold objectives.* A discounted MDP is a tuple $D = (S, A, T, F, \beta)$, where the first four components are as in a solvency MDP and $0 < \beta < 1$ is a *discount factor.* The semantics of a discounted MDP is given by a finite-state MDP $D^{\beta} = (S, A, T)$ and a reward function $\mathit{disc}(\cdot)$ which to every run $\omega = s_0 a_1 s_1 a_2 \ldots$ in D^{β} assigns its *total discounted reward* $\mathit{disc}(\omega) = \sum_{i=1}^{\infty} F(s_{i-1}, a_i) \cdot \beta^i$. The threshold objective asks the controller to maximize, for a given threshold $t \in \mathbb{Q}$, the probability of the event $\mathit{Thr}(t) = \{\omega \in \mathit{Run}(D^{\beta}) \mid \mathit{disc}(\omega) \geq t\}$.

Now consider a solvency MDP $M = (S, A, T, F, \varrho)$ with an initial configuration (s_0, x_0) and a discounted MDP $D = (S, A, T, F, 1/\varrho)$ with a threshold objective $\mathit{Thr}(-x_0)$. Note that once an initial configuration $(s_0, x_0) \in S \times \mathbb{Q}$ is fixed, there is a natural one-to-one correspondence between runs in M_{ϱ} initiated in (s_0, x_0) and runs in $D^{1/\varrho}$ initiated in s: we identify a run $(s_0, x_0) a_1 (s_1, x_1) a_2 \ldots$ in M_{ϱ} with a run $s_0 a_1 s_1 a_2 \ldots$ in $D^{1/\varrho}$. This correspondence naturally extends to strategies in both MDPs, so we assume that these MDPs have identical sets of runs and strategies.

▶ **Proposition 4.** *Let M, D be as above. Then $\mathbb{P}^{\sigma}_{M,(s,x)}(\mathit{Win}) = \mathbb{P}^{\sigma}_{D,s}(\mathit{Thr}(-x))$ for all $\sigma \in \Sigma$.*

Proof. It suffices to show that for every run ω we have $\omega \in \mathit{Win} \Leftrightarrow \mathit{disc}(\omega) \geq -x$. Fix a run $\omega = (s_0, x_0) a_1 (s_1, x_1) a_2 \ldots$, and define, for every $n \geq 0$, $\mathit{disc}_n(\omega) \overset{\text{def}}{=} \sum_{i=1}^{n} F(s_{i-1}, a_i) \cdot \frac{1}{\varrho^i}$ (an empty sum is assumed to be equal to 0). Obviously, for every $n \geq 0$ we have $x_n = \varrho^n \cdot (\mathit{disc}_n(\omega) + x_0)$. Thus, if $\mathit{disc}(\omega) = \lim_{n\to\infty} \mathit{disc}_n(\omega) > -x_0$, then $\lim_{n\to\infty} x_n$ exists and it is equal to $+\infty$. Similarly, if $\mathit{disc}(\omega) < -x_0$, then $\lim_{n\to\infty} x_n = -\infty$. If $\mathit{disc}(\omega) = -x_0$, the infimum wealth x_n along ω is finite (see Appendix A.1), and so $\omega \in \mathit{Win}$. ◀

It follows that many natural problems for solvency MDPs (value computation etc.) are polynomially equivalent to similar natural problems for discounted MDPs with threshold objectives. In particular, our problem of computing/approximating $\mathbf{W}_M(s_0, p)$ is interreducible with the *value-at-risk* problem in discounted MDPs, where the aim is to compute/approximate the supremum threshold t such that under suitable strategy the probability (risk) of the discounted reward being $\leq t$ is at most $1-p$.

3 Qualitative Case

In this section we establish a connection between the qualitative problem for solvency MDPs (i.e., determining whether $x \geq \mathbf{W}_M(s,1)$ for a given state s and number x) and the problem of determining the winner in non-stochastic discounted games.

▶ **Definition 5** (Discounted game)**.** A finite *discounted game* is a tuple $G = (S_1, S_2, s_0, T, R, \beta)$ where S_1 and S_2 are sets of player 1 and 2 states, respectively; $s_0 \in S_1$ is the initial state; $T \subseteq (S_1 \times S_2) \cup (S_2 \times S_1)$ is a transition relation; $R : (S_1 \cup S_2) \to \mathbb{R}$ is a reward function; and $0 < \beta < 1$ is a discount factor.

A strategy for player $i \in \{1,2\}$ in a discounted game is a function $\zeta_i : (S_1 \cup S_2)^* \cdot S_i \to (S_1 \cup S_2)$ such that $(s, \zeta_i(ws)) \in T$ for every s and w. A strategy is *memoryless* if it only depends on the last state. A pair of strategies ζ_1 and ζ_2 for players 1 and 2 yields a unique run $run(\zeta_1,\zeta_2) = s_0 s_1 \ldots$ in the game, given by $s_j = \zeta_i(s_0 \ldots s_{j-1})$ where i is 1 or 2 depending on whether $s_{j-1} \in S_1$ or $s_{j-1} \in S_2$. The discounted total reward of the run is defined to be $disc(s_0 s_1 \ldots) := \sum_{i=0}^{\infty} \beta^{i+1} R(s_i)$. The *discounted game* problem asks, given a game G and a value x, whether there is a strategy ζ_1 for player 1 such that for all strategies ζ_2 of player 2 we have $disc(run(\zeta_1,\zeta_2)) \geq x$. Such a strategy ζ_1 is then called *winning*.

By Proposition 4 the problem of determining whether $x \geq \mathbf{W}_M(s,1)$ for a state s of a solvency MDP M is interreducible (in polynomial time) with the problem of determining whether there is $\sigma \in \Sigma_D$ such that $\mathbb{P}^{\sigma}_{D,s}(Thr(-x)) = 1$ in the corresponding discounted MDP D. We show that the latter is interreducible[6] with the discounted game problem.

Let us first fix a discounted MDP $D = (S, A, T, F, \beta)$. We say that a run $\omega = s_0 a_1 s_1 \ldots$ of D is *realisable* under a strategy σ if $\sigma(s_0 a_1 \ldots s_n)(a_{n+1}) > 0$, and $T(s_n, a_{n+1})(s_{n+1}) > 0$ for all n. The idea of the reduction relies on the following lemma, which is proved in Appendix A.

▶ **Lemma 6.** *If $\sigma \in \Sigma_D$ satisfies $\mathbb{P}^{\sigma}_{D,s}(Thr(x)) = 1$, then* all *runs realisable under σ are in $Thr(x)$.*

Using the lemma above we can construct a game G from D by stipulating that the results of actions are chosen by player 2 instead of being chosen randomly, and vice versa. The technical details of the reduction are presented in Appendix A. The next theorem follows from the reduction and the fact that memoryless (deterministic) strategies suffice in discounted games.

▶ **Theorem 7.** *For every solvency MDP M there exists an oblivious deterministic strategy which is almost-surely winning in every configuration (s,x) with $x \geq \mathbf{W}_M(s,1)$.*

The discounted game problem is in $NP \cap coNP$ and there exists a pseudopolynomial algorithm computing the optimal value [19]. Also, when one of the players controls no states in a game, the problem can be solved in polynomial time [19]. Hence, we get the following theorem.

▶ **Theorem 8.** *The qualitative problem for solvency MDPs is in $NP \cap coNP$. Moreover, there is a pseudopolynomial algorithm that computes $\mathbf{W}_M(s,1)$ for every state s of M. For the*

[6] Actually, we use slightly different variants of the discounted game problem in reductions *from* and *to* the discounted MDPs problem, respectively. Nevertheless, they establish the desired complexity bounds.

restricted class of solvency Markov chains, to compute $\mathbf{W}_M(s,1)$ and to decide the qualitative problem can be done in polynomial time.

Note that the existence of a reduction from mean-payoff games to discounted games [1] suggests that improving the above complexity to polynomial-time is difficult, since a polynomial-time algorithm for solvency MDPs would give a polynomial-time algorithm for mean-payoff games, existence of which is a longstanding open problem in the area of graph games.

4 Quantitative Case

This section formulates results on quantitative questions for solvency MDPs. We start with a proposition showing that we can restrict our attention to some subset of $S \times \mathbb{Q}$, since for every state there are two values below and above which all strategies are almost-surely winning or losing, respectively. Intuitively, these values represent wealth (positive or negative) for which losses/gains from the interest dominate gains/losses from the gain function F. An important consequence of the proposition, when combined with [14], is that deterministic strategies suffice to maximize the probability of winning. Therefore, in the rest of this section we consider only deterministic strategies. The proposition is proved in Appendix B.1.

▶ **Proposition 9.** *For every state s of the solvency MDP M there are rational numbers*

$$U(M,s) \stackrel{\text{def}}{=} \arg\inf_{x\in\mathbb{R}} \forall\sigma\,.\,\mathbb{P}^{\sigma}_{M,(s,x)}(\mathit{Win}) = 1 \quad \textit{and} \quad L(M,s) \stackrel{\text{def}}{=} \arg\sup_{x\in\mathbb{R}} \forall\sigma\,.\,\mathbb{P}^{\sigma}_{M,(s,x)}(\mathit{Win}) = 0,$$

of encoding size polynomial in $||M||$, and they can be computed in polynomial time using linear programming techniques. Moreover, we have $\mathbb{P}^{\sigma}_{M,(s,U(M,s))}(\mathit{Win}) = 1$ for every strategy σ.

To illustrate the proposition, we return to Example 3 and note that $U(M,s_0) = \frac{20}{3}$ and $L(M,s_0) = -\frac{40}{3}$. Obviously, for every s we have $K \geq U(M,s) \geq L(M,s) \geq -K$ where $K = \max_{(s,a)\in S\times A} \frac{|F(s,a)|}{\varrho-1}$, but as Example 3 shows, using $U(M,s)$ and $L(M,s)$ we can restrict the set of interesting configurations more than with the trivial bounds K and $-K$.

We also define the global versions of the bounds, i.e., $L(M) \stackrel{\text{def}}{=} \min_{s\in S} L(M,s)$ and $U(M) \stackrel{\text{def}}{=} \max_{s\in S} U(M,s)$. In accordance with the economic interpretation of our model, we call any configuration of the form (s,x) with $x \geq U(M,s)$ a *rentier configuration*. From Proposition 9 it follows that every run which visits a rentier configuration belongs to Win.

Note that although Proposition 9 suggests that we can restrict our analysis to the configurations (s,x) where $L(M,s) \leq x \leq U(M,s)$, the set of reachable configurations between these bounds is still infinite in general as the following example shows.

▶ **Example 10.** Consider a solvency MDP $M = (\{s\},\{a,b\},T,F,\frac{3}{2})$ with $T(s,a) = T(s,b) = s$, and $F(s,a) = \frac{1}{2}$ and $F(s,b) = -\frac{1}{2}$. We have $L(M) = -1$ and $U(M) = 1$. We will show that for any $n \in \mathbb{N}$ there is a configuration (s,x_n) where $x_n = k/2^n$ that is reachable in exactly n steps from an initial configuration $(s,\frac{1}{2})$ and satisfies $k \in \mathbb{N}_0$, $0 \leq k < 2^n$, $2 \nmid k$. Hence the reachable state space from $(s,\frac{1}{2})$ is infinite as the numbers x_n are pairwise different.

We set $x_0 = \frac{1}{2}$, and let (s,x_n) be a reachable configuration where x_n is of the form $k/2^n$ satisfying the above conditions. In one step we can reach configurations (s,x') where $x' = \varrho x_n \pm \frac{1}{2} = \frac{3k\pm 2^n}{2^{n+1}}$. Clearly $2 \nmid 3k \pm 2^n$; otherwise we would have $2 \mid 3k$ and thus $2 \mid k$ which contradicts the definition of x_n. It remains to show that one of the values of x' again satisfies the above conditions; this is a simple exercise, and we give a proof in Appendix B.2.

Note that if the interest ϱ is restricted to be an integer, the reachable configuration space between $L(M)$ and $U(M)$ is finite, because for the initial configuration (s,x) it holds $x = \frac{p}{q}$

where $p, q \in \mathbb{Z}$, and $\varrho \cdot x + y = \frac{\varrho \cdot p + y \cdot q}{q}$. Hence, any reachable wealth is a multiple of $\frac{1}{q}$, and there are only finitely many such numbers between $L(M)$ and $U(M)$. This means that one can use off-the-shelf algorithms for finite-state MDPs, i.e., minimising the probability to reach configuration with (s, x), where $x < L(M, s)$. However, for the general case, this is not possible and we need to devise new techniques.

4.1 Approximation Algorithms

In this subsection we show how to approximate $\mathbf{W}(s, p)$. Our algorithm depends on the following theorem, which allows us, in a certain sense that will be explained soon, to approximate the function $Val_M(s_0, \cdot)$.

▶ **Theorem 11.** *There is an algorithm that computes, for a solvency MDP M with initial configuration (s_0, x_0) and a given $\varepsilon > 0$, a rational number v and a strategy σ such that:*

1. *$v \geq Val_M(s_0, x_0)$.*
2. *Strategy σ is v-winning from configuration $(s_0, x_0 + \varepsilon)$.*

The running time of the algorithm is polynomial in $|S| \cdot |A| \cdot \log\left(p_{\min}^{-1}\right)$ where $p_{\min} = \min_{(s,s',a) \in S^2 \times A} T(s, a)(s')$, and exponential in $\log(|r_{\max}|/(\varrho - 1))$ and $\log(1/\varepsilon)$ where $r_{\max} = \max_{(s,a) \in S \times A} |F(s, a)|$.

We will prove Theorem 11 later, but first we argue that the theorem is important in its own right. Consider the following scenario. Suppose that an investor starts with wealth x_0. It is plausible to assume that this initial wealth is not strictly fixed. Instead, one can assume that the investor is willing to acquire some small additional amount of wealth (represented by ε), in exchange for some substantial benefit. Here, the benefit consists of the fact that the small difference in the initial wealth allows the investor to compute and execute a strategy, under which the risk of bankruptcy is provably no greater than the lowest risk achievable with the original wealth. Note that the strategy σ may not be $Val_M(s_0, x_0 + \varepsilon)$-winning from $(s_0, x_0 + \varepsilon)$. We now proceed with the theorem providing the approximation of $\mathbf{W}(s, x)$

▶ **Theorem 12.** *For a given solvency MDP M, its state s and rational numbers $\delta > 0$, $p \in [0, 1]$, it is possible to approximate $\mathbf{W}(s, p)$ up to the absolute error δ in time polynomial in $(|S| \cdot |A|)^{\mathcal{O}(1)} \cdot \log\left(p_{\min}^{-1}\right)$, and exponential in $\log(|r_{\max}|/(\varrho - 1))$ and $\log(1/\delta)$, where $p_{\min}$ an $r_{\max}$ are as in Theorem 11.*

Proof. Suppose that we already know that $a \leq \mathbf{W}(s, p) \leq b$, for some a, b. We can use the algorithm of Theorem 11 for $s_0 = s$, $x_0 = a + (b - a)/2$ and $\varepsilon = (b - a)/4$. If the algorithm returns $v \leq p$, we know that $a + (b - a)/2 \leq \mathbf{W}(s, p) \leq b$, otherwise we can conclude that $a \leq \mathbf{W}(s, p) \leq a + 3(b - a)/4$. Initially we know that $L(M) \leq \mathbf{W}(s, p) \leq U(M)$, so in order to approximate $\mathbf{W}(s, p)$ with absolute error δ it suffices to perform $\mathcal{O}(\log((U(M) - L(M))/\delta))$ iterations of this procedure, finishing when $\varepsilon \leq \delta/4$. ◀

Later we will show that the time complexity of the algorithm cannot be improved to polynomial in either $\log(|r_{\max}|/(\varrho - 1))$ or $\log(1/\delta)$ unless P=NP.

Proof of Theorem 11. For the rest of this section we fix a solvency MDP $M = (S, A, T, F, \varrho)$ and its initial configuration (s_0, x_0). First we establish the existence of a strategy that, given a small additional amount of wealth, reaches a rentier configuration in at most exponential number of steps with probability at least $Val_M(s_0, x_0)$. Then, we will show how to compute such a strategy in exponential time.

To establish the proof of the following proposition, we use a suitable Bellman functional whose unique fixed point is equal to $\mathbf{W}$. The proof can be found in Appendix B.3.

▶ **Proposition 13.** *For every initial configuration (s,x) and every $\varepsilon > 0$ there is a strategy σ_ε such that starting in $(s,x+\varepsilon/2)$, σ_ε ensures hitting of a rentier configuration in at most $n = \lceil \frac{\log(U(M)-L(M))+\log\varepsilon^{-1}+2}{\log\varrho} \rceil$ steps with probability at least $Val_M(s,x)$. In particular, $\mathbb{P}^{\sigma_\varepsilon}_{(s,x+\varepsilon/2)}(Win) \geq Val_M(s,x)$.*

The previous proposition shows that the number v and strategy σ of Theorem 11 can be computed by examining the possible behaviours of M during the first n steps. However, since $\log\varrho \approx \varrho - 1$ for ϱ close to 1, the number n can be exponential in $||M||$. Thus, the trivial algorithm, that unfolds the MDP from the initial configuration $(s_0, x_0 + \varepsilon/2)$ into a tree of depth n, and on this tree computes a strategy maximising the probability of reaching a rentier configuration, has a doubly-exponential complexity. The key idea allowing to reduce this complexity to singly-exponential is to round the numbers representing the wealth in the configurations of M to numbers of polynomial size. If the size is chosen carefully, the error introduced by the rounding is not large enough to thwart the computation. In the following we assume that $\log\varrho < \log(U(M) - L(M)) + \log(\varepsilon^{-1}) + 2$, since otherwise $n = 1$ and we can compute the strategy σ and number v by computing an action that maximizes the one-step probability of reaching a rentier configuration from $(s_0, x_0 + \varepsilon/2)$.

We now formalise the notion of rounding the numbers appearing in configurations of M. Let λ be a rational number. We say that two configurations (s,x), (s',x') are λ-*equivalent*, denoted by $(s,x) \sim_\lambda (s',x')$, if $s = s'$ and one of the following conditions holds:

- both x and x' are greater than $U(M,s)$ or less than or equal to $L(M,s)$; or
- $L(M,s) < x, x' \leq U(M,s)$ and there is $k \in \mathbb{Z}$ such that both $x, x' \in (k\lambda, (k+1)\lambda]$.

Clearly, $\sim_\lambda$ is indeed an equivalence on the set $S \times \mathbb{Q}$, and every member of the quotient set $(S \times \mathbb{Q})/{\sim_\lambda}$ is a tuple of the form (s,D), with $s \in S$ and D being either a half-open interval of length at most λ or one of the intervals $(U(M,s),+\infty)$, $(-\infty, L(M,s)]$. For such D, we denote by w_D the maximal element of D (putting $w_{(U(M,s),+\infty)} = +\infty$). We also denote by $[s,x]_\lambda$ the equivalence class of (s,x).

Now let n be as in Proposition 13. We define an MDP $M_{\lambda,n}$ representing an unfolding of M into a DAG of depth n, in which the current wealth w is always rounded up to the least integer multiple of λ greater than w, with configurations exceeding the upper or dropping below the lower threshold of Proposition 9 being immediately recognized as winning or losing. The unfolded MDP $M_{\lambda,n}$ is formally defined as follows.

▶ **Definition 14.** [Unfolded MDP] Let $M = (S, A, T, F, \varrho)$ be an solvency MDP, and $n > 0$ and $\lambda > 0$ two numbers. We define an MDP $M_{\lambda,n} = (S', A, T')$ where S' is $((S \times \mathbb{Q})/{\sim_\lambda}) \times \{0, 1, \ldots, n\}$, and the transition function T' is the unique function satisfying the following:

- for all $(s, D, i) \in S'$ and $a \in A$ where $i < n$ and D is a bounded interval, the distribution $T'((s,D,i),a)$ is defined iff $a \in A(s)$, and assigns $T(s,a)(s')$ to $([s', \varrho\cdot w_D + F(s,a)]_\lambda, i+1)$
- for every other vertex $(s,D,i) \in S'$ there is only a self loop on this vertex under every action, i.e., $T'((s,D,i),a)$ is given by $[(s,D,i) \mapsto 1]$ for every action $a \in A$.

The size of $M_{\lambda,n}$ as well as the time needed to construct it is $(|S| \cdot |A| \cdot \log(p_{\min}^{-1}) \cdot n \cdot \lambda^{-1})^{\mathcal{O}(1)}$.

Now we denote by *Hit* the set of all runs in $M_{\lambda,n}$ that contain a vertex of the form $(t, (U(M,t), \infty), i)$, and by $Ar(z)$ (for “almost rentier”) the set of all runs in M that hit a configuration of the form (t,y) with $y \geq U(M,t) - z$ in at most n steps. In particular, $Ar(0)$ is the event of hitting a rentier configuration in at most n steps. The following lemma (proved in Appendix B.4) shows that $M_{\lambda,n}$ adequately approximates the behaviour of M.

▶ **Lemma 15.** *Let (s,y) be an arbitrary configuration of M. Then the following holds:*

1. *For every $\sigma \in \Sigma_M$ there is $\pi \in \Sigma_{M_{\lambda,n}}$ such that $\mathbb{P}^{\pi}_{M_{\lambda,n},([s,y]_\lambda,0)}(Hit) \geq \mathbb{P}^{\sigma}_{M,(s,y)}(Ar(0))$.*

```
Input: MDP $M$, state $s$, number $p \in [0,1]$, $\delta > 0$
Output: number and strategy satisfying conditions of Theorem 12
$a := L(M,s)$; $b := U(M,s)$ ;                                   // see Proposition 9
do
    $\varepsilon := (b-a)/4$;    $n := \lceil(\log(U(M)-L(M)) + \log\varepsilon^{-1} + 2)/\log\varrho\rceil$;
    $y := a + (b-a)/2$;    $\lambda := \lceil(64 \cdot n \cdot (U(M)-L(M))^2)/\varepsilon^3\rceil^{-1}$;
    $M' := M_{\lambda,n}$ and initial configuration $t := [s,y]_\lambda$ ;          // see Definition 14
    $v := \sup_\pi \mathbb{P}^\pi_{M',t}(Hit)$; ;        // $Val_M(s,y) \leq v \leq Val_M(s,y+\varepsilon)$ by Theorem 11
    if $v \leq p$ then $a := a + (b-a)/2$ else $b := a + 3(b-a)/4$;
while $(b-a)/4 > \delta$;
Compute $\sigma$ s.t. $\mathbb{P}^\sigma_{M,(s,y)}(Ar(n \cdot \lambda \cdot \varrho^n)) \geq v$ ;        // see Lemma 15
return $a$ and the wealth-independent strategy (see p. 11) given by $\sigma$ and $(s,y)$;
```

Algorithm 1: Algorithm approximating $\mathbf{W}(s,p)$

2. *There is $\sigma \in \Sigma_M$ such that $\mathbb{P}^\sigma_{M,(s,y)}(Ar(n \cdot \lambda \cdot \varrho^n)) \geq \sup_\pi \mathbb{P}^\pi_{M_{\lambda,n},([s,y]_\lambda,0)}(Hit) \stackrel{\text{def}}{=} v$, where the supremum is taken over $\Sigma_{M_{\lambda,n}}$. Moreover, the number v and a finite representation of the strategy σ can be computed in time $||M_{\lambda,n}||^{\mathcal{O}(1)}$.*

We can now finish the proof of Theorem 11. Let us put $\lambda = \lceil(64 \cdot n \cdot (U(M) - L(M))^2)/\varepsilon^3\rceil^{-1}$. An easy computation (shown in Appendix B.5) proves that $n \cdot \lambda \cdot \varrho^n \leq \frac{\varepsilon}{2}$ thanks to our assumption that $\log\varrho < \log(U(M) - L(M)) + \log(\varepsilon^{-1}) + 2$.

By Proposition 13 there is a strategy σ_ε in M with $\mathbb{P}^\sigma_{M,(s_0,x_0+\varepsilon/2)}(Ar(0)) \geq Val_M(s_0,x_0)$, and so from Lemma 15 (1.) we get $\sup_\pi \mathbb{P}^\pi_{M_{\lambda,n},([s_0,x_0+\varepsilon/2]_\lambda,0)}(Hit) \geq Val_M(s_0,x_0)$. By part (2.) of the same lemma we can compute, in time $||M_{\lambda,n}||^{\mathcal{O}(1)}$, a strategy σ in M and a number v such that $\mathbb{P}^\sigma_{M,(s_0,x_0+\varepsilon/2)}(Ar(\varepsilon/2)) \geq v \geq Val_M(s_0,x_0)$. In other words, from $(s_0, x_0+\varepsilon/2)$ the strategy σ reaches with probability at least v a configuration that is only $\varepsilon/2$ units of wealth away from being rentier. Note that once an initial configuration is fixed, any strategy can be viewed as being *wealth-independent*, i.e. being only a function of a sequence of states and actions in the history, since the current wealth can be inferred from this sequence and the initial wealth. Suppose now that we fix the initial configuration $(s_0, x_0+\varepsilon)$ instead of $(s_0, x_0+\varepsilon/2)$, keeping the same strategy σ (i.e., we use a strategy that selects the same action as σ after observing the same sequence of states and actions). It is then obvious that we reach a rentier configuration with probability at least v, i.e., $\mathbb{P}^\sigma_{(s,x+\varepsilon)}(Win) \geq v$ as required.

It remains to analyse the complexity of the construction. The analysis is merely technical and is postponed to Appendix B.6. ◀*(Thm. 11)*

The results described in this section are summarised in a form of pseudocode in Algorithm 1.

4.2 Lower Bounds

Now we complement the positive results given above with lower complexity bounds.

▶ **Theorem 16.** *The problem of deciding whether $\mathbf{W}(s,p) \leq x$ for a given x is NP-hard. Furthermore, existence of any of the following algorithms is not possible unless P=NP:*

1. *An algorithm approximating $\mathbf{W}(s,p)$ up to the absolute error δ in time polynomial in $|S| \cdot |A| \cdot \log\left(p_{\min}^{-1}\right)$ and $\log(|r_{\max}|/(\varrho-1))$ and exponential in $\log(1/\delta)$.*
2. *An algorithm approximating $\mathbf{W}(s,p)$ up to the absolute error δ in time polynomial in $|S| \cdot |A| \cdot \log\left(p_{\min}^{-1}\right)$ and $\log(1/\delta)$ and exponential in $\log(|r_{\max}|/(\varrho-1))$.*

Above, the numbers $r_{\max}$ and $p_{\min}$ are as in Theorem 11.

Proof sketch. We show how to construct, for a given instance of the Knapsack problem, a solvency MDP M in which the item values are suitably encoded into probabilities of certain transitions, while the item weights are encoded as rewards associated to some actions. We then show that the instance of Knapsack has a solution if and only if for a certain state s of M and a certain number p (which can be computed from the instance) it holds that $\mathbf{W}(s,p) \leq 0$. We also show that in order to decide this inequality it suffices (for the constructed MDP M) to approximate $\mathbf{W}(s,p)$ up to the absolute error $\frac{1}{4}$. (Intuitively, this corresponds to the well-known fact that no polynomial approximation algorithm for Knapsack can achieve a constant absolute error.) To get part (2.) we use a slight modification of the same approach.

Let us note that a crucial component of the aforementioned reductions is that $\mathit{Val}_M(t,\cdot)$ may not be a continuous function (see example 3). Intuitively, this allows us to recognise whether the current wealth, which in M always encodes weight of some set of items, surpasses some threshold. The proof can be found in Appendix B.7. ◀

Note that thanks to the interreducibility from Proposition 4, the (suitably rephrased) results of Theorems 12 and 16 hold also for the value-at-risk approximation in discounted MDPs.

5 Conclusions

We have introduced solvency MDPs, a model apt for analysis of systems where interest is paid or received for the accumulated wealth. We have analysed the complexity of fundamental problems, and proposed algorithms that approximate the minimum wealth needed to win with a given probability and compute a strategy that achieves the goal. As a by-product, we obtained new results for the *value-at-risk* problem in discounted MDPs.

There are several important directions of future study. One question deserving attention is to find an algorithm computing or approximating $\mathit{Val}(s,x)$. The usual approaches of discretising the state space do not work in this case since the function $\mathit{Val}(s,\cdot)$ is not continuous and thus it is difficult to bound the error introduced by the discretisation. Another direction is the implementation of the algorithms and their evaluation on case-studies.

References

1 D. Andersson and P. B. Miltersen. The complexity of solving stochastic games on graphs. In *Proceedings of the ISAAC '09*, pages 112–121, Berlin, Heidelberg, 2009. Springer-Verlag.

2 N. Bäuerle and U. Rieder. *Markov Decision Processes with Applications to Finance.* Springer, 2011.

3 N. Berger, N. Kapur, L. Schulman, and V. Vazirani. Solvency games. In *Proceedings of FST&TCS 2008*, volume 2 of *LIPIcs*, pages 61–72. Schloss Dagstuhl, 2008.

4 K. Boda and J. A. Filar. Time consistent dynamic risk measures. *Math. Meth. of OR*, 63(1):169–186, 2006.

5 K. Boda, J. A. Filar, Y. Lin, and L. Spanjers. Stochastic target hitting time and the problem of early retirement. *IEEE Trans. Automat. Contr.*, 49(3):409–419, 2004.

6 T. Brázdil, V. Brožek, K. Etessami, and A. Kučera. Approximating the termination value of one-counter mdps and stochastic games. *Inf. Comput.*, 222:121–138, 2013.

7 A. Chakrabarti, L. de Alfaro, T. A. Henzinger, and M. Stoelinga. Resource interfaces. In *Proc. of EMSOFT 2003*, volume 2855 of *LNCS*, pages 117–133, Heidelberg, 2003. Springer.

8 K. Chatterjee and L. Doyen. Energy parity games. In *Proceedings of ICALP 2010, Part II*, volume 6199 of *LNCS*, pages 599–610. Springer, 2010.

9 K. Chatterjee and L. Doyen. Energy and mean-payoff parity Markov decision processes. In *Proceedings of MFCS 2011*, volume 6907 of *LNCS*, pages 206–218. Springer, 2011.

10 K. Chung and M. J. Sobel. Discounted mdps: Distribution functions and exponential utility maximization. *SIAM J. Contr. Optim.*, 25:49–62, 1987.

11 J. Filar, D. Krass, and K. Ross. Percentile performance criteria for limiting average markov decision processes. *IEEE Trans. Automat. Contr.*, 40(1):2–10, 1995.

12 J. D. Hamilton. *Time series analysis*, volume 2. Cambridge Univ Press, 1994.

13 J. Hull. *Options, futures, and other derivatives.* Pearson, 2009.

14 D. Martin. The determinacy of Blackwell games. *J. of Symb. Logic*, 63(4):1565–1581, 1998.

15 M. Schäl. Markov decision processes in finance and dynamic options. *International Series in Operations Research & Management Science*, 40:461–487, 2002.

16 M. J. Sobel. The variance of discounted markov decision processes. *J. Appl. Probab.*, 19:794–802, 1982.

17 D. J. White. Minimizing a threshold probability in discounted markov decision processes. *J. Math. Anal. Appl.*, 173:634–646, 1993.

18 C. Wu and Y. Lin. Minimizing risk models in markov decision processes with policies depending on target values. *J. Math. Anal. Appl.*, 231(1):47–67, 1999.

19 U. Zwick and M. Paterson. The complexity of mean payoff games on graphs. *TCS*, 158(1–2):343–359, 1996.

Technical Appendix

A Proofs for Section 3

A.1 Proof of Proposition 4

We show the missing part of the proof of Proposition 4. We need to show that the infimum wealth x_n along ω is finite. This follows from the fact that for every $n \geq 0$ we have

$$\begin{aligned} x_n &= \varrho^n \cdot (\mathit{disc}_n(\omega) - \mathit{disc}(\omega)) = -\varrho^n \cdot \Big(\sum_{i=n+1}^{\infty} \frac{1}{\varrho^i} \cdot F(s_{i-1}, a_i) \Big) \\ &\geq -\frac{\varrho^n}{\varrho^{n+1}} \cdot \frac{\max_{(s,a)\in S\times A} |F(s,a)|}{1 - \frac{1}{\varrho}} = -\frac{\max_{(s,a)\in S\times A} |F(s,a)|}{\varrho - 1}. \end{aligned}$$

A.2 Proof of Lemma 6

Lemma 6. *Let σ be a strategy in D such that $\mathbb{P}^{\sigma}_{D,s}(\mathit{Thr}(x)) = 1$. Then* all *runs realisable under σ are in $\mathit{Thr}(x)$.*

Proof. Suppose the lemma does not hold, and let $\omega = s_0 a_0 s_1 \dots$ be a run realisable under σ such that $\mathit{disc}(\omega) = x - \varepsilon$ for some $\varepsilon > 0$. Let $M := \sum_{i=0}^{\infty} \beta^i \max_{s,a} |F(s,a)| = \frac{\max_{s,a} |F(s,a)|}{1-\beta}$ and let k be such that $\beta^{k+1} \cdot 2 \cdot M < \varepsilon$. Let ω' be any run of the form $s_0 a_1 \dots a_k s_k b_{k+1} t_{k+1} b_{k+2} \dots$ (i.e. $\omega_k = \omega'_k$). Then, denoting $t_k := s_k$ we have

$$\mathit{disc}(\omega') = \mathit{disc}(\omega) - \Big(\sum_{i=k+1}^{\infty} \beta^i \cdot F(s_{i-1}, a_i) \Big) + \Big(\sum_{i=k+1}^{\infty} \beta^i \cdot F(t_{i-1}, b_i) \Big) \leq x - \varepsilon + \beta^{k+1} \cdot 2 \cdot M < x.$$

However, the probability of such runs is nonzero, a contradiction with the assumption that $\mathbb{P}^{\sigma}_{D,s}(\mathit{Thr}(x)) = 1$. ◀

A.3 Interreducibility of discounted MDPs and discounted games

Let us first fix a discounted MDP $D = (S, A, T, F, \beta)$ We define a game $G = (S, (S \times A), s_0, T_G, R_G, \sqrt{\beta})$ with

- $(s, (s, a)) \in T_G$ whenever $T(s, a)$ is defined;
- $((s, a), s') \in T_G$ whenever $T(s, a)(s') > 0$; and
- $R_G(s) = 0$ for all $s \in S$, and $R_G((s, a)) = F(s, a)$

For the clarity of presentation we first assume that $\sqrt{\beta}$ is a rational number of polynomial encoding size. Then we will show how to get rid of this assumption.

Let σ be a strategy in D such that $\mathbb{P}^{\sigma}_{D,s}(\mathit{Thr}(x)) = 1$. We define a strategy σ^G for the player 1 in G by $\sigma^G(s_0(s_0, a_1)s_1 \dots s_n) = (s_n, a_{n+1})$ where $a_n \in A$ is an arbitrary action satisfying $\sigma(s_0 a_1 \dots s_n)(a_{n+1}) > 0$. For all player 2 strategies ζ_2 we have that to $\mathit{run}(\sigma^G, \zeta_2) = s_0(s_0, a_1)s_1 \dots$ corresponds the run $\omega = s_0 a_1 s_1 \dots$ in D which is realisable under σ, and $\mathit{disc}(\mathit{run}(\sigma^G, \zeta_2)) = \mathit{disc}(\omega)$. Because every run ω realisable under σ is in $\mathit{Thr}(x)$, we have that $\mathit{disc}(\mathit{run}(\sigma^G, \zeta_2)) \geq x$. For the other direction, let ζ_1 be a winning player 1 strategy, by [19] we can assume that it is is memoryless. We define a strategy σ for D by $\sigma(s) = \zeta_1(s)$ for all $s \in S$. Assume $\mathbb{P}^{\sigma}_{D,s}(\mathit{Thr}(x) < 1)$, and let $\omega = s_0 a_1 s_1 \dots$ be a run realisable under σ such that $\mathit{disc}(\omega) < x$. Then we can fix a strategy ζ_2 for player

2 in G defined by $\zeta_2(s_0(s_0,a_1)\dots s_n) = a_{n+1}$. We can easily show that $disc(run(\zeta_1,\zeta_2)) = disc(\omega) < x$, which contradicts that ζ_1 is winning.

Now we drop the assumption that $\sqrt{\beta}$ is a rational number of polynomial encoding size. In such a case we represent the number $\sqrt{\beta}$ symbolically, as a triplet $(P(x),0,1)$, where $P(x) = x^2 - \beta$ is the minimal polynomial of $\sqrt{\beta}$ over $\mathbb{Q}$ and the numbers $0,1$ represent the fact that we are interested in the positive (i.e., the one lying in the interval $[0,1]$) root of $P(x)$. Now consider again the aforementioned game G with $\sqrt{\beta}$ represented as above. Surely, determining the winner in such a game is at least as hard as determining the winner in "standard" discounted games, since all rational numbers can be also represented in the triplet form. It thus remains to show that the problem of determining the winner is in $NP \cap coNP$. Let us recall how the NP-algorithm for standard games (see [19]) works: first, it guesses a winning memoryless deterministic strategy of player 1 and then it verifies, using linear programming techniques, that against this strategy the player 2 cannot decrease the discounted reward below x. Now linear programs with coefficients represented in the triplet form can be solved on a Turing machine in time polynomial in the encoding size of the triplets and in the degree of the algebraic extension defined by adjoining all the coefficients in the program to $\mathbb{Q}$ (see [**?**, Theorem 21]). The linear program obtained by guessing the strategy in G contains only one coefficient which may be irrational, namely $\sqrt{\beta}$, which generates an extension of degree at most 2. Thus, we can again verify that the guessed strategy is winning in polynomial time. For the $coNP$ upper bound we proceed similarly.

For the other direction of interreducibility, for a discount game $G = (S_1, S_2, s_0, T_G, R, \beta)$ we define a discounted MDP $D = (S_1, S_2, T, F, \beta^2)$ where

- T is an arbitrary function satisfying that $T(s,t)(s') > 0$ iff $(s,t) \in T$ and $(t,s') \in T$;
- $F(s,t) = R(s)/\beta + R(t)$

The rest of the proof proceeds in the same way as above.

▶ Remark. If $\sqrt{\beta}$ is a rational number, then the pseudopolynomial algorithm for the qualitative problem in solvency MDPs can be immediately obtained from the pseudopolynomial algorithm for discounted games in [19]. The algorithm in that paper iterates, for a pseudopolynomial number of steps, a suitable Bellman functional, where each iteration performs a polynomial number of additions, multiplications, divisions and comparisons, which involve the discount (i.e., in our reduction, the number $\sqrt{\beta}$). Since all of these operations can be computed in polynomial time for algebraic numbers in the triplet form [**?**, Proposition 16], and all the intermediate results lie in the extension generated by $\sqrt{\beta}$ over $\mathbb{Q}$, the algorithm is pseudopolynomial even for games with symbolically represented discounts. We note that in our game G, the optimal value resulting from this algorithm is rational even if $\sqrt{\beta}$ is irrational, because it corresponds to the minimal threshold achievable with probability 1 in a discounted MDP with rational discount β. Rationality of this minimal threshold can be shown by devising a suitable Bellman functional (we omit this argument, because it is not essential for our paper).

B Proofs for Section 4

B.1 Proof of Proposition 9

Here we present an extended version of Proposition 9.

▶ **Proposition 17.** *For every state s of the solvency MDP M there are rational numbers*

$U(M,s)$ and $L(M,s)$, such that

$$U(M,s) \overset{\text{def}}{=} \arg\inf_{x\in\mathbb{R}} \forall\sigma \,.\, \mathbb{P}^{\sigma}_{M,(s,x)}(\mathit{Win}) = 1,$$

$$L(M,s) \overset{\text{def}}{=} \arg\sup_{x\in\mathbb{R}} \forall\sigma \,.\, \mathbb{P}^{\sigma}_{M,(s,x)}(\mathit{Win}) = 0,$$

of encoding size polynomial in $||M||$, and they are solutions of the following linear programs:

$$\begin{array}{ll} \mathit{max} & \sum_{s\in S} L(M,s) \\ \mathit{s.t.} & L(M,s) \le \frac{1}{\varrho}(L(M,t) - F(s,a)) \\ & \textit{(for all } s \in S,\ a \in A(s), \textit{ and } t \in \mathit{Succ}(s,a)\textit{)} \end{array}$$

and

$$\begin{array}{ll} \mathit{min} & \sum_{s\in S} U(M,s) \\ \mathit{s.t.} & U(M,s) \ge \frac{1}{\varrho}(U(M,t) - F(s,a)) \\ & \textit{(for all } s \in S,\ a \in A(s), \textit{ and } t \in \mathit{Succ}(s,a)\textit{).} \end{array}$$

Moreover, we have $\mathbb{P}^{\sigma}_{M,(s,U(M,s))}(\mathit{Win}) = 1$ for every strategy σ.

Proof. First we show that these are actually real numbers and not $\pm\infty$. Let $g_{\max} = \max_{(s,a)\in S\times A} F(s,a)$ be the maximal gain that occurs in the solvency MDP, and fix arbitrary x such that $g_{\max} + \varrho \cdot x < 0$, denoting $\tau := g_{\max} + \varrho \cdot x$. Then *any* run starting in (s,x) is of the form $(s_0,x_0) \cdot a_1 \cdot (s_1,x_1) \cdot a_2 \cdots$ where $x_i \le x + i \cdot \tau$. Hence we get that $L(M,s) > x$. For $U(M,s)$ we take the minimal gain $g_{\min} = \min_{(s,a)\in S\times A} F(s,a)$ and proceed similarly.

We proceed by proving that the above values satisfy the optimality conditions. We first present the proof for value $U(M,s)$.

Assume that the initial wealth is $x_0 \ge U(M,s)$. From LP it follows that for any action a for all $t \in \mathit{Succ}(s,a)$ we have $\varrho \cdot x_0 + F(s,a) \ge U(M,t)$. So no matter how the strategy picks actions, in the following states the accumulated wealth never falls below $\min_{s\in S} U(M,s)$, and because we know that $U(M,s) \in \mathbb{R}$, this ensures that any strategy wins almost surely.

For the other direction assume that $x_0 < U(M,s)$ and let $\delta = U(M,s) - x_0$. We construct a strategy σ which loses with positive probability. Let σ pick an action $a = \arg\max_{a\in A(s)} \max_{t\in \mathit{Succ}(s,a)} \frac{1}{\varrho}(U(M,t) - F(s,a))$, and let $t = \arg\max_{t\in \mathit{Succ}(s,a)} \frac{1}{\varrho}(U(M,t) - F(s,a))$, i.e., such that action a and state t is a bounding constraint in the LP. It follows that $U(M,t) - (\varrho \cdot x_0 + F(s,a)) = \varrho \cdot \delta$. The strategy continues by picking actions the same way as in s and ensures that after k steps, there exists a run ending in some state t which has a nonzero measure and the difference between $U(M,t)$ and wealth is equal to $\varrho^k \cdot \delta$, and so for any value $X > -\infty$ we can find a k such that the wealth $< X$ will be accumulated on some finite path having nonzero probability. This implies that wealth $< \min_{s\in S} L(M,s)$ will be eventually reached and thus the strategy will lose with positive probability.

Now we prove that the values $L(M,s)$ satisfy the optimality conditions. First, assume that the initial wealth x_0 satisfies $x_0 < L(M,s_0)$ and let $\delta = L(M,s_0) - x_0$. From the linear program we know that for all actions a and successors $t \in \mathit{Succ}(s_0,a)$ we have that $L(M,t) - (\varrho \cdot x_0 + F(s_0,a)) \ge \varrho \cdot \delta$. Hence, no matter what is the choice of the strategy, in the next step the difference between wealth and $L(M,t)$ will be at least $\varrho \cdot \delta$ for any successor t. We can show by induction that after k steps the difference between the wealth and $L(M,t)$ is at least $\varrho^k \cdot \delta$; and because $\varrho > 1$ we have that as $k \to \infty$ we have wealth going to $-\infty$.

Now let the initial wealth be $x_0 > L(M,s_0)$ and let $\delta = x_0 - L(M,s_0)$. We construct a strategy, which is winning with positive probability. Consider the strategy σ which picks an action $a = \arg\min_{a\in A(s_0)} \min_{t\in \mathit{Succ}(s_0,a)} \frac{1}{\varrho}(L(M,t) - F(s_0,a))$, and let

$t = \arg\min_{t \in \mathit{Succ}(s_0,a)} \frac{1}{\varrho}(L(M,t) - F(s_0,a))$, i.e., such that action a and state t is a bounding constraint in the LP. Hence, it follows that $(\varrho \cdot x_0 + F(s_0,a)) - L(M,t) = \varrho \cdot \delta$. The strategy continues by picking actions the same way as in s_0 and ensures that after k steps, there exists a run ending in some state t which has a nonzero measure and the difference between wealth and the $L(M,t)$ is equal to $\varrho^k \cdot \delta$, and so for any value $X < \infty$ we can find a k such that the wealth $> X$ will be accumulated on some finite path having nonzero probability. This implies that wealth $> \max_{s \in S} U(M,s)$ will be eventually reached and thus the strategy will win with positive probability.

The bound on the encoding size of the numbers follows from the standard results on linear programming. ◀

B.2 Supplement to Example 10

We distinguish two cases. Firstly, if $3k + 2^n < 2^{n+1}$, we put $x_n = 3k + 2^n$. Secondly, if $3k + 2^n \geq 2^{n+1}$, we have $3k - 2^n = 3k + 2^n - 2^{n+1} \geq 0$. We argue that $3k - 2^n < 2^{n+1}$, which allows us to put $x_{n+1} = 3k - 2^n$. Suppose the opposite, i.e. $3k - 2^n \geq 2^{n+1}$. This gives us $3k \geq 2^{n+1} + 2^n = 3 \cdot 2^n$ and thus $k \geq 2^n$, again a contradiction.

B.3 Proof of Proposition 13

Let us recall the fact that thanks to the Proposition 9 we are now restricted to deterministic strategies.

Proposition 13. *For every initial configuration (s,x) and every $\varepsilon > 0$ there is a strategy σ_ε such that starting in $(s, x+\varepsilon/2)$, σ_ε ensures hitting of a rentier configuration in at most $n = \lceil \frac{\log(U(M)-L(M))+\log\varepsilon^{-1}+2}{\log\varrho} \rceil$ steps with probability at least $\mathit{Val}_M(s,x)$. In particular, $\mathbb{P}^{\sigma_\varepsilon}_{(s,x+\varepsilon/2)}(\mathit{Win}) \geq \mathit{Val}_M(s,x)$.*

In the proof of Proposition 13 we proceed by a series of lemmas. First we show that the vector $\mathbf{W} = (\mathbf{W}(s,p))^{s\in S}_{p\in[0,1]} \in \mathbb{R}^{S\times[0,1]}$ is a unique fixed point of a suitable Bellman operator $\mathcal{L}$.

Let s be any state of M. For an action $a \in A(s)$ and number $p \in [0,1]$ we denote by $B(s,a,p)$ the set of all vectors $\mathbf{q} \in [0,1]^{\mathit{Succ}(s,a)}$ that satisfy $\sum_{s' \in \mathit{Succ}(s,a)} \mathbf{q}(s') \cdot (T(s,a)(s')) \geq p$. The intuition behind the $B(s,a,p)$ vectors is that if a strategy σ is p-winning in (s,x) and it picks an action a, then there must be a vector $\mathbf{q} \in B(s,a,p)$ such that for all $s' \in \mathit{Succ}(s,a)$, the probability of winning from the successor of (s,x) that is of the form (s',x') for some x' must be at least $\mathbf{q}(s')$. Consider now the Bellman operator $\mathcal{L}$ defined on the uncountably-dimensional space $\mathbb{R}^{S\times[0,1]}$ as follows:

$$\mathcal{L}(\mathbf{V})(s,p) \quad = \quad \min_{a\in A(s)} \inf_{\mathbf{q}\in B(s,a,p)} \max_{s'\in \mathit{Succ}(s,a)} \frac{1}{\varrho} \cdot (\mathbf{V}(s', \mathbf{q}(s')) - F(s,a)),$$

for all vectors $\mathbf{V} \in \mathbb{R}^{S\times[0,1]}$ and all $(s,p) \in S \times [0,1]$.

▶ **Lemma 18.** *The vector $\mathbf{W}$ is a fixed point of the operator $\mathcal{L}$.*

Proof. Assume, for the sake of contradiction, that there are $s \in S$, $p \in [0,1]$ such that $\mathcal{L}(\mathbf{W})(s,p) < \mathbf{W}(s,p)$. Pick an arbitrary $\delta > 0$ such that $\mathcal{L}(\mathbf{W})(s,p) + \delta < \mathbf{W}(s,p)$, and denote by x the left-hand side of this inequality. From the definition of $\mathcal{L}$ it follows, that there are $a^* \in A(s)$ and $\mathbf{q}^* \in B(s,a^*,p)$ such that for all $s' \in \mathit{Succ}(s,a^*)$ we have $\frac{1}{\varrho} \cdot (\mathbf{W}(s', \mathbf{q}^*(s')) - F(s,a^*)) \leq \mathcal{L}(\mathbf{W})(s,p) + \delta = x$, or in other words,

$$\varrho \cdot x + F(s,a^*) > \mathbf{W}(s', \mathbf{q}^*(s')). \tag{1}$$

Now, starting in (s,x) the strategy can choose the action a^* in the first step. If in the second step the current vertex is s' (where $s' \in \mathit{Succ}(s,a^*)$), we switch to a strategy that ensures winning from $(s', \varrho \cdot x + F(s,a^*))$ with probability at least $\mathbf{q}^*(s')$ (such a strategy must exist, due to (1)). Using this approach, the probability of winning from (s,x) is at least $\sum_{s' \in \mathit{Succ}(s,a^*)} \mathbf{q}(s') \cdot (T(s,a^*)(s')) \geq p$, where the last inequality holds because $\mathbf{q}^* \in B(s,a^*,p)$. Since $x < \mathbf{W}(s,p)$, we get a contradiction with the definition of $\mathbf{W}$.

It remains to show that $\mathcal{L}(\mathbf{W})(s,p) \leq \mathbf{W}(s,p)$, for an arbitrary fixed $(s,p) \in S \times [0,1]$. It suffices to show that for every $\varepsilon > 0$ there is a strategy σ such that $\mathbb{P}^{\sigma}_{(s,\mathcal{L}(\mathbf{W})(s,p)+\varepsilon)} \geq p$. Similarly to the previous paragraph, there must be $a^* \in A(s)$ and $\mathbf{q}^* \in B(s,a^*,p)$ such that $\varrho \cdot (\mathcal{L}(\mathbf{W})(s,p) + \varepsilon) + F(s,a^*) \geq \mathbf{W}(s', \mathbf{q}^*(s'))$, for all $s' \in \mathit{Succ}(s,a^*)$. So if the strategy σ chooses a^* in the first step, then in the second step the play will be in some configuration $(s', \varrho \cdot (\mathcal{L}(\mathbf{W})(s,p) + \varepsilon) + F(s,a^*))$, from which a strategy winning with probability at least $\mathbf{q}^*(s')$ exists, and σ will behave as this strategy from the second step onwards. Since $\mathbf{q}^* \in B(s,a^*,p)$, it follows that indeed $\mathbb{P}^{\sigma}_{(s,\mathcal{L}(\mathbf{W})(s,p)+\varepsilon)}(\mathit{Win}) \geq p$. ◀

We denote by LU the set of all vectors $\mathbf{V} \in \mathbb{R}^{S \times [0,1]}$ that satisfy $L(M,s) \leq \mathbf{V}(s,p) \leq U(M,s)$, for all $(s,p) \in S \times [0,1]$. We also denote $||\mathbf{V}||_\infty = \sup_{(s,p) \in S \times [0,1]} |\mathbf{V}(i)|$.

▶ **Lemma 19.** *If $\mathbf{V} \in LU$, then also $\mathcal{L}(\mathbf{V}) \in LU$. Moreover, for every pair of vectors $\mathbf{V}, \mathbf{V}'$ we have $||\mathcal{L}(\mathbf{V}) - \mathcal{L}(\mathbf{V}')||_\infty \leq \frac{1}{\varrho} ||\mathbf{V} - \mathbf{V}'||_\infty$.*

Proof. Let $\mathbf{V} \in LU$, $s \in S$ and $p \in [0,1]$ be arbitrary. Assume, for the sake of contradiction, that $\mathcal{L}(\mathbf{V})(s,p) > U(M,s)$. By definition, any strategy wins from $(s, \mathcal{L}(\mathbf{V})(s,p))$ with probability 1. Thus, for every $a \in A(s)$ end every $s' \in \mathit{Succ}(s,a)$ we have $\varrho\mathcal{L}(\mathbf{V})(s,p) + F(s,a) \geq U(M,s')$. But at the same time, by definition of $\mathcal{L}$ we have $\mathcal{L}(\mathbf{V})(s,p) \leq \varrho^{-1}(\mathbf{V}(s',p') - F(s,a))$ for suitable $p' \in [0,1]$. Combining these two inequalities we get $\mathbf{V}(s',p') \geq U(M,s')$, a contradiction with $\mathbf{V} \in LU$. The inequality $\mathcal{L}(\mathbf{V})(s,p) \geq L(M,s)$ can be established in a similar way.

For the second part, fix arbitrary vectors $\mathbf{V}$, $\mathbf{V}' \in \mathbb{R}^{S \times [0,1]}$ and some $(s,p) \in S \times [0,1]$. We have to show that $|\mathbf{V}(s,p) - \mathbf{V}'(s,p)| \leq \varrho^{-1} ||\mathbf{V} - \mathbf{V}'||_\infty$. Let us choose an arbitrary $\varepsilon > 0$. From the definition of $\mathcal{L}$ it follows that there are $a, b \in A(s)$, $\mathbf{q} \in B(s,a,p)$ and $\mathbf{r} \in B(s,b,p)$ such that

$$y_1 \stackrel{\text{def}}{=} \max_{s' \in \mathit{Succ}(s,a)} \frac{1}{\varrho}\big(\mathbf{V}(s', \mathbf{q}(s')) - F(s,a)\big) \leq \mathbf{V}(s,p) + \frac{\varepsilon}{2} \tag{2}$$

$$y_2 \stackrel{\text{def}}{=} \max_{s' \in \mathit{Succ}(s,b)} \frac{1}{\varrho}\big(\mathbf{V}'(s', \mathbf{r}(s')) - F(s,b)\big) \leq \mathbf{V}'(s,p) + \frac{\varepsilon}{2}. \tag{3}$$

Assume that $y_1 \geq y_2$, the other case can be handled in a symmetric way. We have

$$\begin{aligned} 0 \leq y_1 - y_2 &\leq \max_{s' \in \mathit{Succ}(s,b)} \frac{1}{\varrho}\big(\mathbf{V}(s', \mathbf{r}(s')) - F(s,b)\big) + \frac{\varepsilon}{2} - y_2 \\ &= \frac{1}{\varrho} \max_{s' \in \mathit{Succ}(s,b)} (\mathbf{V}(s', \mathbf{r}(s')) - \mathbf{V}'(s', \mathbf{r}(s'))) + \frac{\varepsilon}{2} \leq \frac{1}{\varrho} ||\mathbf{V} - \mathbf{V}'||_\infty + \frac{\varepsilon}{2}, \end{aligned} \tag{4}$$

where the second inequality follows from the facts that $\mathbf{V}(s,p) \leq y_1 \leq \mathbf{V}(s,p) + \varepsilon/2$ and $\mathbf{V}(s,p) \leq \max_{s' \in \mathit{Succ}(s,b)} \frac{1}{\varrho}\big(\mathbf{V}(s', \mathbf{r}(s')) - F(s,b)\big)$. Putting (2), (3) and (4) together, and using the fact that $y_2 \geq \mathbf{V}'(s,p)$, we obtain

$$|\mathbf{V}(s,p) - \mathbf{V}'(s,p)| \leq |y_1 - y_2| + \frac{\varepsilon}{2} \leq \frac{1}{\varrho} ||\mathbf{V} - \mathbf{V}'||_\infty + \varepsilon.$$

Since $\varepsilon > 0$ was chosen arbitrarily, the result follows. ◀

The previous lemma shows, that the operator $\mathcal{L}$ is a contraction mapping on LU. Now the set LU equipped with the supremum norm $||\cdot||_\infty$ is a Banach space. By the Banach fixed-point theorem, $\mathcal{L}$ has a unique fixed point in LU, which must be equal to $\mathbf{W}$ by Lemma 18. Moreover, for every vector $\mathbf{V} \in LU$ the sequence $\mathcal{L}^n(\mathbf{V})$ converges to this fixed point as n approaches infinity.

Consider now a vector $\mathbf{V}^0 \in LU$ such that $\mathbf{V}^0(s,p) = U(M,s)$ for every (s,p).

▶ **Lemma 20.** *Consider any $(s,p) \in S \times [0,1]$ and any $y > \mathcal{L}^n(\mathbf{V}^0)(s,p)$. Then there is a strategy σ such that starting in (s,y), the strategy σ ensures hitting a rentier configuration in at most n steps with probability at least p.*

Proof. We proceed by induction on n. The case $n = 0$ is trivial. So assume that and that the lemma holds for some $n \in \mathbb{N}_0$. Consider any $(s,p) \in S \times [0,1]$. Then there must be an action $a \in A(s)$ and vector $\mathbf{r} \in B(s,a,p)$ such that

$$y > \max_{s' \in Succ(s,a)} \frac{1}{\varrho}\big(\mathcal{L}^n(\mathbf{V}^0)(s',\mathbf{r}(s')) - F(s,a)\big). \tag{5}$$

Then, in order to reach a rentier configuration in at most $n+1$ steps with probability at least p, the strategy can proceed as follows: in the first step, it chooses the aforementioned action a. In the second step, the play is in some state s' with probability $T(s,a)(s')$ and the current wealth is greater than $\mathcal{L}^n(\mathbf{V}^0)(s',\mathbf{r}(s'))$ (by (5)). By induction, the strategy can then switch to a strategy that reaches a rentier configuration in at most n steps with probability at least $\mathbf{r}(s')$. Because $\mathbf{r} \in B(s,a,p)$, it follows that this strategy ensures reaching a rentier configuration from (s,y) in at most $n+1$ steps with probability at least p. ◀

We can now finish the proof of Proposition 13. We have $||\mathbf{W} - \mathcal{L}^n(\mathbf{V}^0)||_\infty \leq \frac{1}{\varrho^n} \cdot ||\mathbf{W} - \mathbf{V}^0||_\infty \leq \frac{1}{\varrho^n} \cdot (U(M) - L(M))$ (where the first inequality follows from Lemmas 18 and 19). It follows that for $n = \lceil \frac{\log(U(M)-L(M)) + \log \varepsilon^{-1} + 2}{\log \varrho} \rceil$ it holds $||\mathbf{W} - \mathcal{L}^n(\mathbf{V}^0)||_\infty \leq \varepsilon/4$. In particular, $x + \varepsilon/2 \geq \mathbf{W}(s, Val_M(s,x)) + \varepsilon/2 > \mathcal{L}^n(\mathbf{V}^0)(s, Val_M(s,x))$. Thus, the strategy σ_ε can be chosen to be the strategy σ from Lemma 20 for $p = Val(s,x)$, n and $y = x + \varepsilon/2$.

B.4 Proof of Lemma 15

Lemma 15. *Let (s,y) be an arbitrary configuration of M. Then the following holds:*

1. *For every $\sigma \in \Sigma_M$ there is $\pi \in \Sigma_{M_{\lambda,n}}$ such that*

$$\mathbb{P}^{\pi}_{M_{\lambda,n},([s,y]_\lambda,0)}(Hit) \geq \mathbb{P}^{\sigma}_{M,(s,y)}(Ar(0)).$$

2. *There is $\sigma \in \Sigma_M$ such that*

$$\mathbb{P}^{\sigma}_{M,(s,y)}(Ar(n \cdot \lambda \cdot \varrho^n)) \geq \sup_{\pi} \mathbb{P}^{\pi}_{M_{\lambda,n},([s,y]_\lambda,0)}(Hit) \stackrel{\text{def}}{=} v,$$

where the supremum is taken over $\Sigma_{M_{\lambda,n}}$. Moreover, the number v and a finite representation of the strategy σ can be computed in time $||M_{\lambda,n}||^{\mathcal{O}(1)}$.

Proof. Again, we remind the reader that we are now restricted to deterministic strategies (because of Proposition 9). For the purpose of this proof, let us define the *absorbing vertices* of $M_{\lambda,n}$ to be all vertices of this MDP that are not of the form (s,D,i) with D bounded interval and $i < n-1$. Moreover, we denote by B the set of all histories in which the last vertex is of the form $(t,(U(M,t),\infty),i)$ while all the previous vertices are non-absorbing. We also denote by C the set of all histories in M that contain exactly one rentier configuration (and thus, their every proper prefix does not contain any such configuration). Note that $\mathit{Hit} = \bigcup_{u \in B} \mathsf{Cone}(u)$ and $Ar(0) = \bigcup_{v \in C} \mathsf{Cone}(v)$. In the following we assume that the initial configuration (t_0, x_0) of all runs in M is equal to (s,y) and that the initial vertex $(s_0, D_0, 0)$ of all runs in $M_{\lambda,n}$ is equal to $([s,y]_\lambda, 0)$.

First we describe certain natural correspondence between runs in M and $M_{\lambda,n}$, which will be used throughout the proof. Let X be a set of all histories in $M_{\lambda,n}$ that do contain at most one absorbing vertex. For any history $X \ni u = (s_0,D_0,0)a_1 \cdots a_k(s_k,D_k,k)$ in $M_{\lambda,n}$ there is exactly one history $f(u) = (t_0,x_0)b_1 \cdots b_k(t_k,x_k)$ in M such that, $s_i = t_i$ for all $0 \le i \le k$ and $a_i = b_i$ for all $1 \le i \le k$. Note that f, viewed as a function from X to the histories of M, is injective.

On the other hand, for every history $v = (t_0,x_0)b_1 \cdots b_k(t_k,x_k)$ in M there is a unique history $g(v) = (s_0,D_0,0)a_1 \cdots a_k(s_k,D_k,k)$ in $M_{\lambda,n}$ such that for all $0 \le i \le k$ the vertex (s_i,D_i,i) is either absorbing, or $s_i = t_i$ and $a_{i+1} = b_{i+1}$.[21] A straightforward induction reveals, that for all $0 \le i \le k$ we have $w_{D_i} \ge x_i$ (we always round the numbers *up* to the nearest multiple of λ). It follows that $\mathit{Hit} \supseteq \bigcup_{v \in C} \mathsf{Cone}(g(v))$. The function g viewed as a mapping from C to histories in M_λ, n, may not be injective. Below, we use $\mathit{ker}\, g$ to denote the kernel of g, and $[v]_g$ to denote the equivalence class of v in $C/\mathit{ker}\, g$.

(1.) Fix some strategy σ in M. We define strategy π as follows: for every history u that does not contain an absorbing vertex we set $\pi(u) = \sigma(f(u))$. For other histories (i.e. those that reach an absorbing vertex from which there is no escape), π can choose any action. It is easy to verify that for every history v in M we have

$$\sum_{u \in [v]_g} \mathbb{P}^{\sigma}_{M,(s,y)}\big(\mathsf{Cone}(u)\big) = \mathbb{P}^{\pi}_{M_{\lambda,n},[s,y]_\lambda}\big(\mathsf{Cone}(g(v))\big).$$

It follows that

$$\begin{aligned}
\mathbb{P}^{\pi}_{[s,y]_\lambda}(\mathit{Hit}) &\ge \mathbb{P}^{\pi}_{[s,y]_\lambda}\left(\bigcup_{v \in C} \mathsf{Cone}(g(v))\right) = \mathbb{P}^{\pi}_{[s,y]_\lambda}\left(\bigcup_{[v]_g \in C/\mathit{ker} g} \mathsf{Cone}(g(v))\right) \\
&= \sum_{[v]_g \in C/\mathit{ker} g} \mathbb{P}^{\pi}_{[s,y]_\lambda}\big(\mathsf{Cone}(g(v))\big) = \sum_{v \in C} \sum_{u \in [v]_g} \mathbb{P}^{\sigma}_{(s,y)}\big(\mathsf{Cone}(u)\big) \\
&= \sum_{v \in C} \mathbb{P}^{\sigma}_{(s,y)}\big(\mathsf{Cone}(v)\big) = \mathbb{P}^{\sigma}_{(s,y)}(Ar(0)),
\end{aligned}$$

where the first equality on the second line follows from the fact that $\mathsf{Cone}(h)$ and $\mathsf{Cone}(h')$ are disjoint events for $h \ne h'$ when h is not a prefix of h' and vice versa.

(2.) By standard results on MDPs with reachability objectives there is a memoryless strategy π^* such that $\mathbb{P}^{\pi^*}_{[s,y]_\lambda}(\mathit{Hit}) = \sup_\pi \mathbb{P}^{\pi}_{[s,y]_\lambda}(\mathit{Hit})$, where the maximum is taken over all strategies. It thus suffices to prove that there is a strategy σ in M satisfying $\mathbb{P}^{\sigma}_{(s,y)}(Ar(n \cdot \lambda \cdot \varrho^{n+1})) \ge \mathbb{P}^{\pi^*}_{[s,y]_\lambda}(\mathit{Hit})$.

Let $u = (s_0,D_0,0)a_1 \cdots a_k(s_k,D_k,k)$, $u \in B$ be a history in $M_{\lambda,n}$ and let $f(u) = (t_0,x_0)b_1 \cdots b_k(t_k,x_k)$ be the corresponding history in M. We prove by induction on i that for every $0 \le i < k$ we have $w_{D_k} \le x_k + (i+1)\lambda\varrho^i$. The case $i = 0$ is trivial, since $(s_0,D_0) = [t,y]_\lambda$ and $(t_0,x_0) = (t,y)$, so $w_{D_0} - x_0 \le \lambda$. Suppose now that $i > 0$ and that

[21] The equality of actions is considered only if $i \ne k$

$w_{D_{i-1}} - x_{i-1} \leq i\lambda\varrho^{i-1}$.

$$\begin{aligned} w_{D_i} - x_i &\leq \varrho w_{D_{i-1}} + F(s_{i-1}, a_{i-1}) + \lambda - x_i \\ &= \varrho w_{D_{i-1}} + F(s_{i-1}, a_{i-1}) + \lambda - \varrho x_{i-1} - F(t_{i-1}, b_{i-1}) \\ &= \varrho(w_{D_{i-1}} - x_{i-1}) + \lambda \leq i\lambda\varrho^{i} + \lambda \leq (i+1)\lambda\varrho^{i}. \end{aligned} \tag{6}$$

Here, the first inequality follows from the fact that D_i is an interval of length λ containing $\varrho w_{D_{i-1}} + F(s_{i-1}, a_{i-1})$, the first equality on the third line follows from $t_{i-1} = s_{i-1}$ and $b_{i-1} = a_{i-1}$, while the next inequality follows from the induction hypothesis. As a consequence, we have

$$\begin{aligned} x_k &= \varrho x_{k-1} + F(t_{k-1}, b_k) = \varrho x_{k-1} + F(s_{k-1}, a_{k-1}) \\ &\geq \varrho w_{D_{k-1}} + F(s_{k-1}, a_{k-1}) - k\lambda\varrho^{k+1} \geq U(M, s_k) - k\lambda\varrho^{k}, \end{aligned}$$

where the first inequality on the second line follows from (6), while the next inequality follows from the fact that the last vertex of history u is of the form $(t, (U(M,t), \infty), k)$.

Thus, $Ar(n\lambda\varrho^n) \supseteq \bigcup_{u \in B} \mathsf{Cone}(f(u))$. We now proceed similarly as in (1.). We define strategy σ as follows: for a given history v, if $g(v) \in X$, $\sigma(v) = \pi^*(g(v))$, while for the other histories, σ chooses an arbitrary action. Note that this representation of σ can be computed in time polynomial in $||M_{\lambda,n}||$, since we can use standard polynomial-time algorithm for MDPs with reachability objectives to compute π^*, and $g(v)$ can be computed in time linear in length of v.[22] It is easy to verify, that for every history $u \in X$ we have $\mathbb{P}^{\pi^*}_{M_{\lambda,n},[s,y]_\lambda}(\mathsf{Runs}(u)) = \mathbb{P}^{\sigma}_{M,(s,y)}\big(\mathsf{Runs}(f(u))\big)$. Combining the previous observations we get

$$\begin{aligned} \mathbb{P}^{\sigma}_{(s,y)}\big(Ar(n\lambda\varrho^n)\big) &\geq \mathbb{P}^{\sigma}_{(s,y)}\left(\bigcup_{u \in B} \mathsf{Cone}(f(u))\right) = \sum_{u \in B} \mathbb{P}^{\sigma}_{(s,y)}\big(\mathsf{Cone}(f(u))\big) \\ &= \sum_{u \in B} \mathbb{P}^{\pi^*}_{[s,y]_\lambda}(\mathsf{Cone}(u)) = \mathbb{P}^{\pi^*}_{[s,y]_\lambda}(\mathit{Hit}), \end{aligned}$$

where in the last equality on the second line we use the fact that f is injective. ◀

B.5 Proof that $n \cdot \lambda \cdot \varrho^n \leq \frac{\varepsilon}{2}$

We have

$$\begin{aligned} n \cdot \lambda \cdot \varrho^n &\leq \frac{\varepsilon^3}{64(U(M) - L(M))^2} \cdot \varrho^{\frac{\log(U(M)-L(M)) + \log\varepsilon^{-1} + 2}{\log\varrho} + 1} \\ &\leq \frac{\varepsilon^3}{64(U(M) - L(M))^2} \cdot 2^{\log(U(M)-L(M)) + \log\varepsilon^{-1} + 2} \cdot \varrho \\ &\leq \frac{\varrho \cdot \varepsilon}{4(U(M) - L(M))} \cdot \frac{\varepsilon}{2} \\ &\leq \frac{\varepsilon}{2} \end{aligned}$$

where the last inequality holds because we assumed that $\log\varrho < \log(U(M) - L(M)) + \log(\varepsilon^{-1}) + 2$.

[22] It can be actually computed online as new configurations are visited during the play.

B.6 Complexity Analysis for Theorem 11

Here we conclude the proof of Theorem 11. From the previous observations we have that the complexity is $||M_{\lambda,n}||^{\mathcal{O}(1)}$, which can be rewritten as

$$\left(|S| \cdot |A| \cdot \log\left(p_{\min}^{-1}\right) \cdot n \cdot \lambda^{-1}\right)^{\mathcal{O}(1)} = \left(|S| \cdot |A| \cdot \log\left(p_{\min}^{-1}\right) \cdot \frac{U(M) - L(M)}{\varepsilon \cdot \log \varrho}\right)^{\mathcal{O}(1)}.$$

Noting that $U(M) - L(M) \leq 2r_{\max}/(\varrho - 1)$ and $1/\log \varrho \leq (1 + 1/(\varrho - 1))$, we conclude that the complexity is indeed polynomial in $|S| \cdot |A| \cdot \log\left(p_{\min}^{-1}\right)$ and exponential in $\log(|r_{\max}|/(\varrho - 1))$ and $\log(1/\varepsilon)$.

B.7 Proof of Theorem 16

Theorem 16. *The problem of deciding whether $\mathbf{W}(s, p) \leq x$ for a given x is NP-hard. Furthermore, existence of any of the following algorithms is not possible unless P=NP:*

1. *An algorithm approximating $\mathbf{W}(s, p)$ up to the absolute error δ in time polynomial in $|S| \cdot |A| \cdot \log\left(p_{\min}^{-1}\right)$ and $\log(|r_{\max}|/(\varrho - 1))$ and exponential in $\log(1/\delta)$.*
2. *An algorithm approximating $\mathbf{W}(s, p)$ up to the absolute error δ in time polynomial in $|S| \cdot |A| \cdot \log\left(p_{\min}^{-1}\right)$ and $\log(1/\delta)$ and exponential in $\log(|r_{\max}|/(\varrho - 1))$.*

Above, the numbers $r_{\max}$ and $p_{\min}$ are as in Theorem 11.

Proof. We begin with part (1.), the second part is very similar. We give the proof by reduction from the Knapsack problem. Let us have an instance of a Knapsack problem with items $1, \ldots, n$ (we assume $n \geq 2$), where the weight and value of the ith item is w_i and v_i, respectively, and where the bound on the weight and value of the items to be put in the knapsack are W and V, respectively. We denote $w_{tot} = \sum_{1 \leq i \leq n} w_i$ and $v_{tot} = \sum_{1 \leq i \leq n} v_i$. Without loss of generality we assume that: all the numbers v_i, w_i are nonzero, and that the item weights are integers (this restriction of Knapsack is still NP-hard); and that $v_{tot} < 1/n^2$ (otherwise we can transform the instance by dividing all the numbers v_i and number V by $v_{tot} \cdot n^2$, without influencing the existence of a solution).

We show how to compute, in time polynomial in the encoding size of the Knapsack instance, a solvency MDP $M = (S, A, T, F, \varrho)$ with an interest rate $\varrho = 1 + \frac{1}{4n^2}$, and a number p such that there is a solution to the instance of Knapsack if and only if $\mathbf{W}(s_1, p) = 0$ (for some distinguished state s_1 of M). The interest rate ϱ is chosen in such a way that the inequality $\frac{\varrho^{2n}}{4} \leq \frac{1}{2}$ holds.

First, we put $S = \{s_i, s_i^+, s_i^- \mid 1 \leq i \leq n\} \cup \{s_{n+1}, t_1, t_2, t_3\}$ and $A = \{a_i^+, a_i^- \mid 1 \leq i \leq n\} \cup \{b\}$.

Next, we set $\alpha = 1/n^2$ and define the transitions as follows:

- $T(s_i, a_i^+) = [s_i^+ \mapsto 1]$, for all $1 \leq i \leq n$;
- $T(s_i, a_i^-) = [s_i^- \mapsto 1]$, for all $1 \leq i \leq n$;
- $T(s_i^+, b) = [t_1 \mapsto \frac{v_i}{1-(i-1)\alpha}, t_2 \mapsto \frac{\alpha - v_i}{1-(i-1)\alpha}, s_{i+1} \mapsto 1 - \frac{\alpha}{1-(i-1)\alpha}]$ for all $1 \leq i \leq n$ (note that this indeed defines a probability distribution, since $v_{tot} \leq \alpha < 1 - n\alpha$);
- $T(s_i^-, b) = [t_3 \mapsto \frac{\alpha}{1-(i-1)\alpha}, s_{i+1} \mapsto 1 - \frac{\alpha}{1-(i-1)\alpha}]$ for all $1 \leq i \leq n$;
- $T(t_1, b) = [t_1 \mapsto 1]$, $M(t_2, b) = [t_2 \mapsto 1]$, and $M(t_3, b) = [t_3 \mapsto 1]$;
- $T(s_{n+1}, b) = [s_{n+1} \mapsto 1]$.

The rewards are defined in the following way:

- $F(s_i^-, b) = w_i \cdot \varrho^{-2(n-i)}$, for all $1 \leq i \leq n$;
- $F(t_1, b) = 1$, $F(t_2, b) = F(t_3, b) = -2(w_{tot} + 1)$;

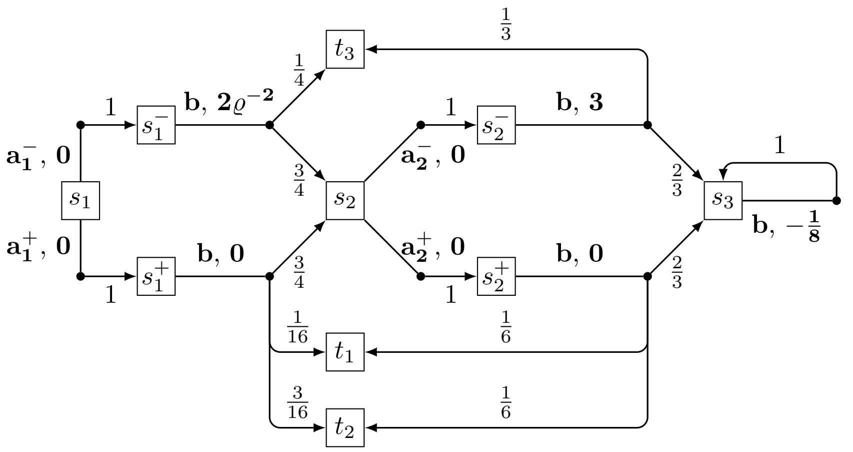

$\frac{1}{3}$
t_3
$\frac{1}{4}$
$\mathbf{b}$, $\mathbf{2}\varrho^{\mathbf{-2}}$
1
s_1^-
$\mathbf{a_1^-}$, $\mathbf{0}$
s_1
$\mathbf{a_1^+}$, $\mathbf{0}$
s_1^+
1
$\mathbf{b}$, $\mathbf{0}$
$\frac{3}{4}$
s_2
$\frac{3}{4}$
1
s_2^-
$\mathbf{b}$, $\mathbf{3}$
$\mathbf{a_2^-}$, $\mathbf{0}$
$\mathbf{a_2^+}$, $\mathbf{0}$
s_2^+
1
$\mathbf{b}$, $\mathbf{0}$
$\frac{2}{3}$
$\frac{2}{3}$
s_3
1
$\mathbf{b}$, $-\frac{\mathbf{1}}{\mathbf{8}}$
$\frac{1}{16}$
t_1
$\frac{1}{6}$
$\frac{3}{16}$
t_2
$\frac{1}{6}$

Figure 1 Solvency MDP M constructed for a simple Knapsack instance with items $1, 2$ such that $w_1 = 2$, $w_2 = 3$, $v_1 = 1/16$, $v_2 = 1/8$, $W = 3$ and $V = 1/8$. We have $\varrho = 1 + 1/16$ and $p = 7/8$. To achieve greater compactness the picture omits loops under action b on states t_1, t_2 and t_3. Action b is rewarded with -12 in states t_2 and t_3.

- $F(s_{n+1}, b) = -(w_{tot} - W)/4n^2$. This ensures that $U(M, s_{n+1}) = w_{tot} - W$, so a run that visits s_{n+1} is winning if and only if the current wealth upon entering s_{n+1} is at least $w_{tot} - W$.
- All other rewards are zero.

Figure 1 illustrates the construction on a simple example.

Finally, we set $p = 1 + V - 1/n$. Clearly, the MDP M and number p can be computed in polynomial time. Note that $p > 1$ would imply $V > v_{tot}$, so in this case the Knapsack does not admit a solution. Thus, we assume that $p \in [0, 1]$.

Under any strategy the state s_{i+1} is reached with probability $(1 - i\alpha)$. This can be shown by induction on i, since

$$\begin{aligned}(1 - i\alpha)(1 - \frac{\alpha}{1 - i\alpha}) &= 1 - \frac{\alpha}{1 - i\alpha} - i\alpha + \frac{i\alpha^2}{1 - i\alpha} = 1 - \frac{\alpha + i\alpha - i^2\alpha^2 - i\alpha^2}{1 - i\alpha} \\ &= 1 - \frac{(i + 1)\alpha(1 - i\alpha)}{1 - i\alpha} = 1 - (i + 1)\alpha\end{aligned}$$

Hence we get that the state s_{n+1} is reached with probability $(1 - 1/n)$. Further, the probability that a run contains both s_i and t_1 is v_i if the strategy picks a_i^+ in the state s_i, and 0 otherwise. So let us choose any (deterministic) strategy σ and an arbitrary initial configuration (s_1, x) with $x \in [0, 1/4]$. Obviously, we interpret the choice of action a_i^+ by σ in s_i as choosing the item i to be packed into the knapsack. Denote $I_\sigma^+ \subseteq \{1, \ldots, n\}$ the set of all indexes i such that σ chooses action a_i^+ in s_i (i.e. the set of items chosen to be packed into the knapsack). A straightforward induction on i reveals that conditional on reaching the state s_i, the current wealth upon reaching this state is

$$\sum_{\substack{j \in \{1, \ldots, n\} \setminus I_\sigma^+ \\ j < i}} w_j \cdot \varrho^{-2(n-i+1)} + \varrho^{2(i-1)} \cdot x.$$

This has two consequences. First, until the run reaches one of the states t_1, t_2 or s_{n+1} the current wealth is always bounded by 0 from below and by $w_{tot} + 1$ from above. Thus, a run starting in (s_1, x) is winning if it reaches t_1 and losing if it reaches t_2 or t_3. Second, if the run reaches the state s_{n+1} (this happens with probability $(1 - 1/n)$), then the current wealth upon reaching this state is equal to

$$\sum_{i \in \{1,\dots,n\} \setminus I^+_\sigma} w_i + \varrho^{2n} \cdot x. \tag{7}$$

Thus, under σ the conditional probability of winning on condition that the play reaches s_{n+1} is 1 if and only if the items in $\{1,\dots,n\} \setminus I^+_\sigma$ have total weight at least $w_{tot} - W - \varrho^{2n} \cdot x$, or in other words, if the items in I^+_σhave total weight at most $W + \varrho^{2n} \cdot x \le W + 1/2$. Otherwise, this conditional probability is 0. Since we assume that the item weights are integers, the aforementioned probability is 1 iff the items contained in I^+_σ have total weight at most W.

We claim that for any strategy σ the set I^+_σ is a solution of the original instance of Knapsack (i.e. set of items selected for inclusion into the knapsack, which satisfies the usual constraints on weight and value) if and only if $\mathbb{P}^\sigma_{(s_1,x)}(\mathit{Win}) \ge p$.

First suppose that I^+_σ is a solution to the Knapsack instance. Since the total weight of items in I^+_σ is at most W, from (7) it follows that under σ, once the play reaches s_{n+1} (this happens with probability $1 - 1/n$) the current wealth is at least $w_{tot} - W = U(M, s_{n+1})$, so conditional on reaching s_{n+1} the probability of winning is 1. Moreover, we know that every run that reaches t_1 is winning as well. We know that this happens with probability $\sum_{i \in I^+_\sigma} v_i \ge V$. We conclude that by using strategy σ we win from (s_1, x) with probability at least $1 - 1/n + V = p$.

On the other hand, if I^+_σ is not a solution of the original instance, there are two cases to consider. Either $\sum_{i \in \{1,\dots,n\} \setminus I^+_\sigma} w_i < w_{tot} - W$, in which case the probability of winning under σ is at most v_{tot}, because the conditional probability of winning upon reaching s_{n+1} is 0, and the only other way to win is to reach t_1. Or $\sum_{i \in I^+_\sigma} v_i < V$, in which case the probability of winning is strictly smaller than $1 - 1/n + V$ (probability of reaching s_{n+1}, where the conditional probability of winning can be 1, plus the probability of visiting t_1 which is $\sum_{i \in I^+_\sigma} v_i < V$). In either case, the probability of winning is less than p.

It follows that the instance of Knapsack admits a solution if and only if there is σ such that $\mathbb{P}^\sigma_{(s_1,x)}(\mathit{Win}) \ge p$, or in other words, iff $\mathit{Val}_M(s_1, x) \ge p$ for every $x \in [0, 1/4]$. To recognize whether this is the case, it suffices to approximate the value $\mathbf{W}(s_1, p)$ up to the absolute error $\delta = 1/8$. Furthermore, for the constructed MDP M we have $\log(|r_{\max}|/(\varrho - 1)) = \log(\mathrm{poly}(w_{tot} \cdot n^2))$ for some polynomial poly. Thus, the existence of an algorithm approximating $\mathbf{W}(s_1, p)$ in time polynomial in $|S| \cdot |A| \cdot \log\left(p_{\min}^{-1}\right) \cdot \log(|r_{\max}|/(\varrho - 1)) \cdot \delta^{-1}$ would imply the existence of an polynomial-time algorithm for Knapsack.

(2.) For the second part of the theorem, the proof is almost the same, we divide all the rewards in M by some sufficiently large number, forcing $r_{\max}$ to be polynomial in the encoding size of the Knapsack instance. We then show that in order to decide whether $\mathbf{W}(s, p) \le 0$ (and thus, whether the instance admits a solution), it suffices to approximate $\mathbf{W}(s, p)$ up to the absolute error $\mathcal{O}(1/w_{tot})$, where w_{tot} is the sum of weights over all items in the instance.

Formally, the only difference is in the gain function F of the constructed MDP M. We put

- $F(s_i^-, b) = w_i \cdot \varrho^{-2(n-i)}/w_{tot}$, for all $1 \le i \le n$;
- $F(t_1, b) = 1$, $F(t_2, b) = F(t_3, b) = -2$;
- $F(s_{n+1}, b) = -4(1 - W/w_{tot}) \cdot n^2$. This ensures that a run that hits s_{n+1} is winning if and only if the current wealth upon entering s_{n+1} is at least $1 - W/w_{tot}$.
- Other rewards are zero.

It is again easy to verify that the states t_1 and t_2 are always winning and losing, respectively, and that for a given strategy σ the current wealth upon entering s_{n+1} is equal to $\sum_{i\in\{1,\dots,n\}\setminus I_\sigma^+} w_i/w_{tot} + \varrho^{2n}\cdot x$, where x is the initial wealth. Now let x be any number from the interval $[0, 1/(4w_{tot})]$ and σ be any strategy. Then the conditional probability of winning on condition that s_{n+1} is reached is 1 iff $\sum_{i\in\{1,\dots,n\}\setminus I_\sigma^+} w_i/w_{tot} + \varrho^{2n}\cdot x \geq (w_{tot} - W)/w_{tot}$, and 0 otherwise. Since $\varrho^{2n}\cdot x \leq 1/(2w_{tot})$, this conditional probability is 1 if and only if $\sum_{i\in\{1,\dots,n\}\setminus I_\sigma^+} w_i/w_{tot} \geq (w_{tot} - W - 1/2)/w_{tot}$, i.e. iff I_σ^+ contains items of weight at most W (since the item weights are integers). Now we can argue in exactly the same way as in the previous part, that I_+^σ is a solution to the Knapsack instance iff σ ensures winning with probability at least $p = 1 - 1/n + V$ from (s_1, x). It follows that the Knapsack instance has a solution if and only if it is possible to win with probability at least p from s_1 with any initial wealth between 0 and $1/(4w_{tot})$. To check whether this is the case it suffices to approximate $\mathbf{W}(s_1, p)$ up to the absolute error $\delta = 1/(8w_{tot})$. The constructed MDP M has $\log(|r_{\max}|/(\varrho - 1)) = \log(\mathrm{poly}(n))$ for some polynomial poly (since $r_{\max} \leq 4n^2$). Thus, an approximation algorithm running in time polynomial in $|S|\cdot|A|\cdot\log\left(p_{\min}^{-1}\right)\cdot\log(\delta^{-1})$ and exponential in $\log(|r_{\max}|/(\varrho - 1))$ would yield a polynomial-time algorithm for knapsack. ◀