Solvency Markov Decision Processes with Interest

Tomáš Brázdil*[1], Taolue Chen[2], Vojtěch Forejt†[3], Petr Novotný[1], and Aistis Simaitis[3]

1 **Faculty of Informatics, Masaryk University, Czech Republic**
2 **Department of Computer Science, Middlesex University London, UK**
3 **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 ∩ 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. [16, 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 [13, 14].

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

* 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.

(real) reward to every action which is collected whenever the action is chosen. The states 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. [16]). 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., [17, 11, 18, 19]. 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

[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.

management under uncertainty. We show that for every solvency MDP there is a bound on 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 [20] 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. Proofs omitted due to the space constraints can be found in [7].

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. [8, 9, 10]), 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 [17, 11, 18, 19], where the authors studied finite- and infinite-horizon cases. In the finite-horizon case, in particular [19] 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 [19], 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 [12] 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 $dist(X)$ 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 $T : V \times A \to dist(V)$ 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 $Succ(v,a) = \{u \mid T(v,a)(u) > 0\}$ the *support* of $T(v,a)$.

An *infinite path* (or *run*) is a sequence $v_0a_1v_1a_2v_2\cdots \in (V\times A)^\omega$ such that $a_{i+1}\in A(v_i)$ and $v_{i+1}\in Succ(v_i,a_{i+1})$ 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 $\omega = v_0a_1v_1a_2v_2\ldots$ and $i\in\mathbb{N}$ we denote by ω_i the finite path $v_0a_1\cdots a_iv_i$.

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.

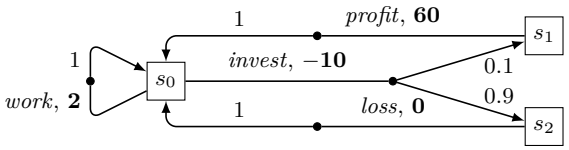
Each history $w\in\mathsf{Hist}_M$ determines the set $\mathsf{Cone}(w)$ consisting of all runs having w as a prefix. To an MDP M, its vertex v and strategy σ we associate the probability space $(\mathsf{Runs}_M,\mathcal{F},\mathbb{P}^\sigma_{M,v})$, where $\mathcal{F}$ is the σ-field generated by all $\mathsf{Cone}(w)$, and $\mathbb{P}^\sigma_{M,v}$ is the unique probability measure such that for every history $w=v_0a_1\ldots a_kv_k$ 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 $1\le i\le k$ (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 $Win = \mathsf{Runs}_M\setminus\{(s_0,x_0)a_1(s_1,x_1)\cdots\in\mathsf{Runs}_M \mid \liminf_{n\to\infty}x_n=-\infty\}$. 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 $Val_M(s_0,x_0)=\sup_\sigma\mathbb{P}^\sigma_{M,(s_0,x_0)}(Win)$ the maximal probability of winning with a given wealth, and $\mathbf{W}_M(s,p)=\inf\{x\mid Val_M(s,x)\ge p\}$ 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 $\mathbb{P}^\sigma_{M,(s_0,x_0)}(Win)\ge p$. A 1-winning strategy is called *almost surely winning*, and strategy σ with $\mathbb{P}^\sigma_{M,(s_0,x_0)}(Win)=0$ is called *almost surely losing*.

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



Here $S = \{s_0, s_1, s_2\}$, $A = \{work, invest, profit, loss\}$, T is depicted by the arrows in the figure, for example $T(s_0, invest) = [s_1 \mapsto 0.1, s_2 \mapsto 0.9]$, the function F is given by the bold numbers next to the actions, e.g. $F(s, work) = 2$, 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 $\mathbb{P}^{\sigma}_{M,(s_0,-10)}(Win) = 0$ 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 $Val_M(s_0, -10) = 0.1$. 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 $Val_M(s_0, -2) = 1$ since a strategy that always chooses *work* is 1-winning in $(s_0, -2)$. This demonstrates that the function $Val(s, \cdot)$ 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 $disc(\cdot)$ which to every run $\omega = s_0 a_1 s_1 a_2 \dots$ in D^{β} assigns its *total discounted reward* $disc(\omega) = \sum_{i=1}^{\infty} F(s_{i-1}, a_i) \cdot \beta^i$. The threshold objective asks the controller to maximize, for a given threshold $t \in \mathbb{Q}$, the probability of the event $Thr(t) = \{\omega \in Run(D^{\beta}) \mid disc(\omega) \geq t\}$.

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 $Thr(-x_0)$. 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 $(s_0, x_0)a_1(s_1, x_1)a_2 \dots$ in M_{ϱ} with a run $s_0 a_1 s_1 a_2 \dots$ 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 $\mathbb{P}^{\sigma}_{M,(s,x)}(Win) = \mathbb{P}^{\sigma}_{D,s}(Thr(-x))$ for all $\sigma \in \Sigma$.*

Proof. It suffices to show that for every run ω we have $\omega \in Win \Leftrightarrow disc(\omega) \geq -x$. Fix a run $\omega = (s_0, x_0)a_1(s_1, x_1)a_2 \dots$, and define, for every $n \geq 0$, $disc_n(\omega) \stackrel{\text{def}}{=} \sum_{i=1}^{n} F(s_{i-1}, a_i) \cdot \frac{1}{\varrho^i}$ (an empty sum is assumed to be equal to 0). Obviously, for every $n \geq 0$ we have $x_n = \varrho^n \cdot (disc_n(\omega) + x_0)$. Thus, if $disc(\omega) = \lim_{n\to\infty} disc_n(\omega) > -x_0$, then $\lim_{n\to\infty} x_n$ exists and it is equal to $+\infty$. Similarly, if $disc(\omega) < -x_0$, then $\lim_{n\to\infty} x_n = -\infty$. If $disc(\omega) = -x_0$, the infimum wealth x_n along ω is finite (see [7]), and so $\omega \in Win$. ◀

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 [7].

▶ **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 [7]. 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 [20]. Also, when one of the players controls no states in a game, the problem can be solved in polynomial time [20]. 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 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.

[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.

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 [15], 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 [7].

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

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

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 that is carried out in [7].

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 [7].

▶ **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 $n=\left\lceil\frac{\log(U(M)-L(M))+\log\varepsilon^{-1}+2}{\log\varrho}\right\rceil$ 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 [7]) 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 $\mathbb{P}^\pi_{M_{\lambda,n},([s,y]_\lambda,0)}(Hit) \geq \mathbb{P}^\sigma_{M,(s,y)}(Ar(0))$.*
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 (see [7]) 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.

The complexity analysis of the reduction is a mere technicality and it is shown in [7]. ◀*(Thm. 11)*

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. In [7] 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.

A crucial component of these reductions is the fact that $Val_M(t, \cdot)$ 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. ◀

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 $Val(s, x)$. The usual approaches of discretising the state space do not work in this case since the function $Val(s, \cdot)$ 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 T. Brázdil, T. Chen, V. Forejt, P. Novotný, and A. Simaitis. Solvency Markov Decision Processes with Interest. *CoRR*, abs/1310.3119, 2013.

8 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.

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

10 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.

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

12 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.

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

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

15 D. Martin. The Determinacy of Blackwell Games. *J. of Symb. Logic*, 63(4):1565–1581, 1998.

16 M. Schäl. Markov Decision Processes in Finance and Dynamic Options. *International Series in Operations Research & Management Science*, 40:461–487, 2002.

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

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

19 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.

20 U. Zwick and M. Paterson. The Complexity of Mean Payoff Games on Graphs. *TCS*, 158(1–2):343–359, 1996.