



# Risk-Averse Optimization of Total Rewards in Markovian Models Using Deviation Measures

Christel Baier  

Technische Universität Dresden, Germany

Jakob Piribauer  

Technische Universität Dresden, Germany; Universität Leipzig, Germany

Maximilian Starke

Technische Universität Dresden, Germany

---

## Abstract

This paper addresses objectives tailored to the *risk-averse* optimization of accumulated rewards in Markov decision processes (MDPs). The studied objectives require maximizing the expected value of the accumulated rewards minus a penalty factor times a deviation measure of the resulting distribution of rewards. Using the variance in this penalty mechanism leads to the variance-penalized expectation (VPE) for which it is known that optimal schedulers have to minimize future expected rewards when a high amount of rewards has been accumulated. This behavior is undesirable as risk-averse behavior should keep the probability of particularly low outcomes low, but not discourage the accumulation of additional rewards on already good executions.

The paper investigates the semi-variance, which only takes outcomes below the expected value into account, the mean absolute deviation (MAD), and the semi-MAD as alternative deviation measures. Furthermore, a penalty mechanism that penalizes outcomes below a fixed threshold is studied. For all of these objectives, the properties of optimal schedulers are specified and in particular the question whether these objectives overcome the problem observed for the VPE is answered. Further, the resulting algorithmic problems on MDPs and Markov chains are investigated.

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

**Keywords and phrases** Markov decision processes, risk-aversion, deviation measures, total reward

**Digital Object Identifier** 10.4230/LIPIcs.CONCUR.2024.9

**Related Version** *Full Version*: <https://arxiv.org/abs/2407.06887> [6]

**Supplementary Material** *Software (Source Code)*:

<https://github.com/experiments-collection/risk-averse-stochastic-shortest-paths>  
archived at `swh:1:dir:eddcf497fe105ca58ea2b4f67171e814a6d35f29`

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

## 1 Introduction

Markov decision processes (MDPs) are a prominent model for systems whose behavior is subject to *non-determinism* and *probabilism*. Non-deterministic behavior might arise, e.g., if a system is employed in an unknown environment, can be controlled by a user, or works concurrently. On the other hand, if, e.g., sufficiently much data on the failure of components is available or randomized algorithms make use of randomization explicitly, it is reasonable to model these aspects of the system as probabilistic.

In order to model quantitative aspects of a system, such as energy consumption, execution time, or utility, MDPs are often equipped with a *reward function* that specifies how much reward is received in each step of an execution. A typical task is then to resolve the non-deterministic choices by specifying a *scheduler*, a.k.a. *policy*, such that the expected value of



© Christel Baier, Jakob Piribauer, and Maximilian Starke;  
licensed under Creative Commons License CC-BY 4.0

35th International Conference on Concurrency Theory (CONCUR 2024).

Editors: Rupak Majumdar and Alexandra Silva; Article No. 9; pp. 9:1–9:20

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the total accumulated reward is maximal (or minimal). In verification, such optimization problems naturally occur when investigating the worst- or best-case expected value of the accumulated reward where worst- and best-case range over all resolutions of the non-deterministic choices. If additionally a target state has to be reached almost surely, this problem is known as the *stochastic shortest path problem* [7, 12].

### Risk-averse optimization

If the objective is the maximization of the expected value of the accumulated rewards, all other aspects of the probability distribution of accumulated rewards are disregarded. This might lead to undesirable behavior as the optimal scheduler might receive low rewards with high probability as long as the expected value is optimal. In many situations, however, a slightly lower expected reward is preferable if it is obtained by a more “stable” behavior in which the risk of encountering low rewards is reduced. E.g., in a traffic control scenarios, it might be important to reduce the risk of congestions while ensuring a reasonable average throughput instead of solely optimizing the average throughput.

In order to define objectives incentivizing such risk-averse behavior, it is worth taking a look at finance and in particular portfolio optimization. Here, Markowitz proclaimed that a portfolio of financial positions should be chosen such that it is Pareto optimal with respect to the expected return and the variance of the return [23]. One way extensively studied in finance to obtain Pareto optimal portfolios is to maximize the *variance-penalized expectation (VPE)*, which is the expected value minus a penalty factor  $\lambda$  times the variance. The parameter  $\lambda$  can be used to obtain different levels of risk-aversion.

Besides the variance, further deviation measures have been investigated to reduce risk in portfolio optimization: The use of the *semi-variance*, which – in contrast to the variance – only takes the deviation of outcomes below the expected value into account, as a penalty mechanism has been introduced in this context by Markowitz [24]. Furthermore, instead of considering quadratic deviations from the expected value as in the case of variance and semi-variance, the *mean absolute deviation (MAD)* can be used to obtain the MAD-penalized expectation (MADPE) studied for portfolio optimization in [18]. The MAD measures the expected absolute deviation from the expected value.

In this paper, we investigate these different deviation measure based penalty mechanisms in the context of the maximization of rewards in MDPs.

### Variance-penalized expectation in MDPs (VPE)

Recently, the maximization of the VPE of accumulated rewards in MDPs was studied in [25]: On the positive side, it is shown that optimal schedulers for the VPE can be chosen to be deterministic finite-memory schedulers. Nevertheless, the optimization of the VPE is shown to be computationally hard: The threshold problem whether the optimal VPE exceeds a given threshold  $\vartheta$  is EXPTIME-hard. An optimal scheduler can be computed in exponential space.

A main drawback of the VPE, however, is of conceptual nature: In [25], it is shown that VPE-optimal schedulers have to *minimize* the future expected rewards as soon as a high amount of rewards (above a computable bound  $B$ ) has been accumulated. We call such schedulers *eventually reward-minimizing schedulers (ERMin-schedulers)*. Intuitively, the reason is that a further accumulation of additional rewards after a high amount of rewards has already been accumulated has a stronger effect on the variance than on the expected value due to the quadratic nature of the variance. Conceptually, this can be considered to be a flaw in the use of the VPE as an objective to yield risk-averse behavior.

■ **Table 1** Overview of the complexity results and the types of schedulers needed for the optimization of the studied objectives and the VPE. The entries “-” indicate that the problem was not studied further as the scheduler needed for the optimization are the undersirable ERMin-schedulers.

	hardness of threshold problem	computation of optimum	optimal schedulers
VPE [25]	EXPTIME-hard; in P for Markov chains	in exponential space	deterministic, finite-memory ERMin-schedulers
SVPE	–	–	randomized, ERMin-schedulers can be necessary
MADPE ( $\lambda \leq 1/2$ ), SMADPE ( $\lambda \leq 1$ )	PP-hard for acyclic Markov chains	quadratic program of exponential size	randomized, finite-memory ERMax-schedulers
MADPE ( $\lambda > 1/2$ ), SMADPE ( $\lambda > 1$ )	–	–	randomized, ERMin-schedulers can be necessary
TBPE	PP-hard for acyclic Markov chains	in pseudo-polynomial time	deterministic, finite-memory ERMax-schedulers

The desired behaviour a suitable objective should induce is that a scheduler achieves a high expected accumulated reward, while keeping the probability of particularly bad outcomes low. Improving on already good outcomes should not have a negative effect. So, we want optimal schedulers to be *eventually reward-maximizing (ERMax-schedulers)*, i.e., that they maximize the expected reward once the accumulated reward exceeds some bound  $B$ .

### Deviation-measure-penalized expectation

Towards this goal, we investigate objectives in the spirit of the VPE, which are of the form  $\mathbb{E}^{\mathfrak{S}}(\text{rew}) - \lambda \mathbb{DEV}^{\mathfrak{S}}(\text{rew})$  where a penalty factor  $\lambda$  times a deviation measure  $\mathbb{DEV}^{\mathfrak{S}}(\text{rew})$  of the probability distribution of accumulated rewards under a scheduler  $\mathfrak{S}$  is subtracted from the expected accumulated reward  $\mathbb{E}^{\mathfrak{S}}(\text{rew})$ .

The first deviation measure we investigate is the MAD. In contrast to the variance, the contribution of an outcome to the MAD only grows linearly with its distance to the expected value. For the MAD and the variance, we also study one-sided variants in which only outcomes below the expected value are considered: The semi-MAD (SMAD) and semi-variance quantify the average absolute or squared deviation below the expected value by assigning deviation 0 to all outcomes above the expected value. Finally, we investigate a simpler alternative to the MADPE: Instead of measuring the deviation from the expected value of accumulated rewards, which itself depends on the chosen scheduler, we consider a threshold-based penalized expectation (TBPE), where outcomes below a threshold  $t$  that can be chosen externally are penalized either linearly or according to more complicated functions.

### Contributions

The main contributions, also summarized in Table 1, are as follows.

- We show that optimal schedulers for the MADPE can be chosen to be ERMax-schedulers, as desired, if the risk-aversion parameter  $\lambda$  is sufficiently small, i.e. if  $\lambda \leq 1/2$ . This bound on the parameter is shown to be tight. Furthermore, we show that randomized schedulers are necessary for the optimization.

We formulate the optimization problem as a quadratic program and obtain a EXPSPACE-upper complexity bound for the threshold problem for the MADPE. On the other hand, we show that already in acyclic Markov chains the threshold problems for the MADPE and the MAD are PP-hard under polynomial-time Turing reductions.

As the semi-MAD is always half of the MAD, the results transfer to the semi-MADPE.

- We investigate the semivariance-penalized expectation (SVPE) and show – somewhat surprisingly – that, for any risk-aversion parameter  $\lambda$ , there are MDPs in which optimal schedulers are ERMin-schedulers. Hence, the SVPE as objective does not overcome the undesirable effects observed for the VPE. Furthermore, we show that, in contrast to the VPE, randomization is necessary for the optimization of the SVPE.
- We show that the TBPE can be optimized in pseudo-polynomial time and that deciding if the TBPE exceeds a bound for linear penalty functions even in acyclic Markov chains is PP-hard under polynomial-time Turing reductions.

As a proof-of-concept, we analyze our algorithms for the optimization of the MADPE and for the TBPE in a small series of experiments.

### Related work

The above mentioned work on the VPE for accumulated rewards in MDPs [25] is the closest related work to our paper. Earlier work on the VPE in MDPs addressed the finite-horizon setting with terminal rewards [11] or applied the notion to mean payoff and discounted rewards [13]. Further, [31] presents a policy iteration algorithm converging against *local* optima for a similar measure. The computation of the variance of accumulated rewards has been studied in Markov chains [30] and in MDPs [21, 22]. In [8], the satisfiability of constraints on the expected mean payoff in conjunction with constraints on the variance or related notions such as a local variability are studied for MDPs.

For MDPs, the SVPE of random variables defined in terms of the stationary distribution has been studied via the use of reinforcement learning algorithms [20]. Conceptually and methodologically this work is nevertheless not closely related to our work. We are not aware of investigations of the MADPE on MDPs.

Furthermore, several approaches to formalize various other risk-averse optimization problems for accumulated rewards in MDPs have been proposed and studied in the literature. This includes the computation of worst- or best-case quantiles [29, 4, 16, 27], also called *values-at-risk*: Given a probability  $p$ , quantiles on the accumulated rewards are the best bound  $C$  such that the accumulated rewards stays below  $C$  with probability at most  $p$  under all or under some scheduler. While quantiles still disregard the distribution below, the *conditional value-at-risk* and the *entropic value-at-risk* are more involved measures that quantify how far the probability mass of the tail of the probability distribution lies below a given quantile. In the context of risk-averse optimization in MDPs, these measures have been studied in [19] and [1]. A further approach, the *entropic risk* measure, reweighs outcomes by an exponential utility function. Optimizing this entropic risk measure leads to schedulers that tend to still achieve a high expected value while keeping the probability of low outcomes small. The entropic risk measure applied to accumulated rewards have been studied in [3] for stochastic games that extend MDPs with an adversarial player.

## 2 Preliminaries

### Notations for Markov decision processes

A *Markov decision process* (MDP) is a tuple  $\mathcal{M} = (S, Act, P, s_{init}, rew)$  where  $S$  is a finite set of states,  $Act$  a finite set of actions,  $P: S \times Act \times S \rightarrow [0, 1] \cap \mathbb{Q}$  the transition probability function,  $s_{init} \in S$  the initial state, and  $rew: S \times Act \rightarrow \mathbb{N}$  the reward function. Note that we only allow non-negative rewards and that rational rewards can be transformed to integral rewards by multiplying all rewards with the least common multiple of all denominators of the rational rewards. We require that  $\sum_{t \in S} P(s, \alpha, t) \in \{0, 1\}$  for all  $(s, \alpha) \in S \times Act$ . We say that action  $\alpha$  is enabled in state  $s$  iff  $\sum_{t \in S} P(s, \alpha, t) = 1$  and denote the set of all actions that are enabled in state  $s$  by  $Act(s)$ . If  $Act(s) = \emptyset$ , we say that  $s$  is a *trap* state. The paths of  $\mathcal{M}$  are finite or infinite sequences  $s_0 \alpha_0 s_1 \alpha_1 \dots$  where states and actions alternate such that  $P(s_i, \alpha_i, s_{i+1}) > 0$  for all  $i \geq 0$ . For  $\pi = s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{k-1} s_k$ ,  $rew(\pi) = rew(s_0, \alpha_0) + \dots + rew(s_{k-1}, \alpha_{k-1})$  – and analogously for infinite paths – denotes the accumulated reward of  $\pi$ ,  $P(\pi) = P(s_0, \alpha_0, s_1) \cdot \dots \cdot P(s_{k-1}, \alpha_{k-1}, s_k)$  its probability, and  $last(\pi) = s_k$  its last state. A path is called *maximal* if it is infinite or ends in the trap state *goal*. The *size* of  $\mathcal{M}$  is the sum of the number of states plus the total sum of the logarithmic lengths of the non-zero probability values  $P(s, \alpha, s')$  as fractions of co-prime integers and the weight values  $rew(s, \alpha)$ .

A *Markov chain* is an MDP in which the set of actions is a singleton. In this case, we can drop the set of actions and consider a Markov chain as a tuple  $\mathcal{M} = (S, P, s_{init}, rew)$  where  $P$  now is a function from  $S \times S$  to  $[0, 1]$  and  $rew$  a function from  $S$  to  $\mathbb{N}$ .

An *end component* of  $\mathcal{M}$  is a strongly connected sub-MDP formalized by a subset  $S' \subseteq S$  of states and a non-empty subset  $\mathfrak{A}(s) \subseteq Act(s)$  for each state  $s \in S'$  such that for each  $s \in S'$ ,  $t \in S$  and  $\alpha \in \mathfrak{A}(s)$  with  $P(s, \alpha, t) > 0$ , we have  $t \in S'$  and such that in the resulting sub-MDP all states are reachable from each other. An end-component is a 0-end-component if it only contains state-action-pairs with reward 0.

### Scheduler

A *scheduler* for  $\mathcal{M}$  is a function  $\mathfrak{S}$  that assigns to each non-maximal path  $\pi$  a probability distribution over  $Act(last(\pi))$ . If the choice of a scheduler  $\mathfrak{S}$  depends only on the current state, i.e., if  $\mathfrak{S}(\pi) = \mathfrak{S}(\pi')$  for all non-maximal paths  $\pi$  and  $\pi'$  with  $last(\pi) = last(\pi')$ , we say that  $\mathfrak{S}$  is *memoryless* and also view it as functions mapping states  $s \in S$  to probability distributions over  $Act(s)$ . A scheduler  $\mathfrak{S}$  that satisfies  $\mathfrak{S}(\pi) = \mathfrak{S}(\pi')$  for all pairs of finite paths  $\pi$  and  $\pi'$  with  $last(\pi) = last(\pi')$  and  $rew(\pi) = rew(\pi')$  is called *reward-based* and can be viewed as a function from state-reward pairs  $S \times \mathbb{N}$  to probability distributions over actions. If there is a finite set  $X$  of memory modes and a memory update function  $U: S \times Act \times S \times X \rightarrow X$  such that the choice of  $\mathfrak{S}$  only depends on the current state after a finite path and the memory mode obtained from updating the memory mode according to  $U$  in each step, we say that  $\mathfrak{S}$  is a *finite-memory scheduler*. A scheduler  $\mathfrak{S}$  is called *deterministic* if  $\mathfrak{S}(\pi)$  is a Dirac distribution for each path  $\pi$  in which case we also view the scheduler as a mapping to actions in  $Act(last(\pi))$ .

### Probability measure

We write  $\Pr_{\mathcal{M}, s}^{\mathfrak{S}}$  to denote the probability measure induced by a scheduler  $\mathfrak{S}$  and a state  $s$  of an MDP  $\mathcal{M}$ . It is defined on the  $\sigma$ -algebra generated by the cylinder sets  $Cyl(\pi)$  of all maximal extensions of a finite path  $\pi = s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{k-1} s_k$  with  $s_0 = s$  by assigning

to  $Cyl(\pi)$  the probability that  $\pi$  is realized under  $\mathfrak{S}$ , which is  $\mathfrak{S}(s_0)(\alpha_0) \cdot P(s_0, \alpha_0, s_1) \cdot \dots \cdot \mathfrak{S}(s_0 \alpha_0 \dots s_{k-1})(\alpha_{k-1}) \cdot P(s_{k-1}, \alpha_{k-1}, s_k)$ . For a set of states  $T$ , we use  $\diamond T$  to denote the event that a state in  $T$  is reached. For details, see [26].

For a random variable  $X$  that is defined on (some of the) maximal paths in  $\mathcal{M}$ , we denote the expected value of  $X$  under the probability measure induced by a scheduler  $\mathfrak{S}$  and state  $s$  by  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X)$ . We define  $\mathbb{E}_{\mathcal{M},s}^{\min}(X) = \inf_{\mathfrak{S}} \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X)$  and  $\mathbb{E}_{\mathcal{M},s}^{\max}(X) = \sup_{\mathfrak{S}} \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X)$  where  $\mathfrak{S}$  ranges over all schedulers for  $\mathcal{M}$  under which  $X$  is defined almost surely. The variance of  $X$  under the probability measure determined by  $\mathfrak{S}$  and  $s$  in  $\mathcal{M}$  is denoted by  $\mathbb{V}_{\mathcal{M},s}^{\mathfrak{S}}(X)$  and defined by  $\mathbb{V}_{\mathcal{M},s}^{\mathfrak{S}}(X) \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}((X - \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X))^2) = \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X^2) - \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X)^2$ . Furthermore, for a measurable set of paths  $\psi$  with positive probability,  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X|\psi)$  denotes the conditional expectation of  $X$  under  $\psi$ . If  $s = s_{init}$ , we sometimes drop the subscript  $s$ .

### Accumulated rewards

Given an MDP  $\mathcal{M} = (S, Act, P, s_{init}, rew)$ , the total accumulated reward is given by the extension of the function  $rew$  to maximal paths. We can check whether  $\mathbb{E}_{\mathcal{M}}^{\max}(rew) = \infty$  by checking whether all (maximal) end components are 0-end components in polynomial time [12]. For our purposes, only MDPs  $\mathcal{M}$  with  $\mathbb{E}_{\mathcal{M}}^{\max}(rew) < \infty$  are interesting. In these MDPs, we can collapse all end components  $\mathcal{E}$ , which are all 0-end components, to single states  $s_{\mathcal{E}}$  while adding a transition with reward 0 to a new trap state. This does not affect the possible distributions of the random variable  $rew$  that can be realized by a scheduler [12]. Furthermore, the behavior of the MDP starting from a state  $s$  with  $\mathbb{E}_{\mathcal{M},s}^{\max}(rew) = 0$ , i.e., from a state  $s$  from which no positive reward is reachable, is irrelevant. So, we can collapse all these states  $s$  with  $\mathbb{E}_{\mathcal{M},s}^{\max}(rew) = 0$  (together with the new trap state) to a single trap state that we call *goal*. By these constructions, we obtain a new MDP  $\mathcal{M}'$  in which exactly the same distributions of the total reward can be realized by schedulers as in  $\mathcal{M}$ . As  $\mathcal{M}'$  does not contain any end components anymore and *goal* is the only trap state in  $\mathcal{M}'$ , the state *goal* is now reached with probability 1 under any scheduler. In the light of the described constructions, we work under the following assumption:

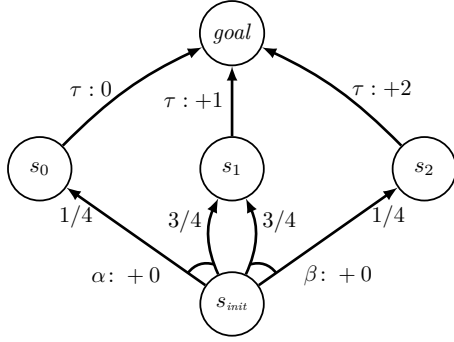
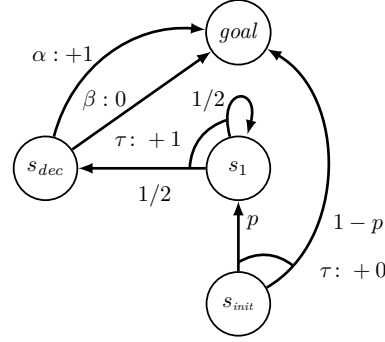
► **Assumption 1.** *W.l.o.g., we assume that all MDPs have a trap state *goal*, which is reached with probability 1 under all schedulers. We add this trap state to the signature and hence denote MDPs  $\mathcal{M}$  as tuples  $\mathcal{M} = (S, Act, P, s_{init}, rew, goal)$ .*

All objectives studied in this paper depend only on the distribution of the random variable  $rew$ . By the following lemma, which is folklore and follows from the formulation in [25, Lemma 2] (see also the full version [6]), we can restrict ourselves to reward-based schedulers.

► **Lemma 2.1.** *Let  $\mathcal{M} = (S, Act, P, s_{init}, rew, goal)$  be an MDP satisfying Assumption 1. Then, for any scheduler  $\mathfrak{S}$  there is a reward-based scheduler  $\mathfrak{T}$  such that the distribution of the random variable  $rew$  is the same under the probability measures  $\Pr_{\mathcal{M}}^{\mathfrak{S}}$  and  $\Pr_{\mathcal{M}}^{\mathfrak{T}}$ .*

## 3 Mean absolute deviation-penalized expectation

As described in the introduction, the VPE suffers from the drawback that optimal schedulers are ERMin-schedulers, which is an undesirable behavior. Intuitively, the reason for this behavior in the case of VPE lies in the fact that the variance grows quadratically with the distance to the expected value. A natural alternative is choosing the absolute distance rather than the quadratic distance from the expected value as the measure for the penalty. So, we define the *mean absolute deviation* (MAD) of a random variable  $X$  as the probability-weighted sum of the distance to the expected value:  $\text{MAD}(X) \stackrel{\text{def}}{=} \mathbb{E}(|X - \mathbb{E}(X)|)$ .

(a) The MDP  $\mathcal{M}$  used in Example 3.1.(b) The MDP  $\mathcal{M}$  used in Example 3.2.

■ **Figure 1** Two example MDPs.

We consider the MAD-penalized expectation (MADPE) of the accumulated weight in an MDP  $\mathcal{M} = (S, Act, P, s_{init}, rew, goal)$  analogously to the VPE: We define the MAD of the accumulated reward  $rew$  under scheduler  $\mathfrak{S}$  as  $\text{MAD}_{\mathcal{M}}^{\mathfrak{S}}(rew) \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(|rew - \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(rew)|)$ . The MAD-penalized expectation with parameter  $\lambda \in \mathbb{R}$  is now  $\text{MADPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}}(rew) \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(rew) - \lambda \text{MAD}_{\mathcal{M}}^{\mathfrak{S}}(rew)$  analogously to the VPE. Our goal is to find

$$\text{MADPE}[\lambda]_{\mathcal{M}}^{\max}(rew) \stackrel{\text{def}}{=} \sup_{\mathfrak{S}} \text{MADPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}}(rew)$$

as well as an optimal scheduler. In the sequel, we will prove the following results. Omitted proofs can be found in [6].

1. In general, randomization is necessary to optimize the MADPE.
2. If  $\lambda > \frac{1}{2}$ , then there is an MDP  $\mathcal{M}$  such that any optimal scheduler for the MADPE is an ERMin-scheduler.
3. If  $\lambda \leq \frac{1}{2}$ , for any MDP  $\mathcal{M}$ , optimal schedulers can be chosen to be reward-based ERMax-schedulers.
4. If  $\lambda \leq \frac{1}{2}$ , the optimal MADPE can be computed in exponential time.
5. Even for acyclic Markov chains, deciding whether the MADPE exceeds a given threshold  $\vartheta$  is PP-hard under polynomial-time Turing reductions.

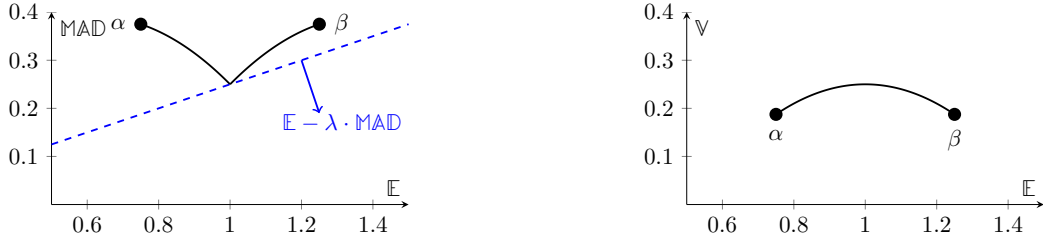
### 3.1 Randomization and optimality of ERMin-schedulers

We work with MDPs  $\mathcal{M} = (S, Act, P, s_{init}, rew, goal)$  satisfying Assumption 1. First, we show that randomization is necessary for the optimization of the MADPE in the following example.

► **Example 3.1.** Consider the MDP  $\mathcal{M}$  in Figure 1a. We consider the schedulers  $\mathfrak{S}_{\alpha}$  choosing  $\alpha$  in  $s_{init}$ ,  $\mathfrak{S}_{\beta}$  choosing  $\beta$ , and  $\mathfrak{S}_{1/2}$  choosing  $\alpha$  and  $\beta$  with probability  $1/2$  each and obtain:  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_{\alpha}}(rew) = 3/4$ ,  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_{1/2}}(rew) = 1$ , and  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_{\beta}}(rew) = 5/4$ . The MADs are  $\text{MAD}_{\mathcal{M}}^{\mathfrak{S}_{\alpha}}(rew) = 3/8$ ,  $\text{MAD}_{\mathcal{M}}^{\mathfrak{S}_{1/2}}(rew) = 1/4 \cdot 1 = 1/4$ , and  $\text{MAD}_{\mathcal{M}}^{\mathfrak{S}_{\beta}}(rew) = 3/8$ . Clearly, the MADPE under  $\mathfrak{S}_{\beta}$  is better than under  $\mathfrak{S}_{\alpha}$  for any  $\lambda > 0$ . For the MADPE of  $\mathfrak{S}_{1/2}$  and  $\mathfrak{S}_{\beta}$  with  $\lambda = 4$ , we obtain

$$\text{MADPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}_{1/2}}(rew) = 1 - \frac{1}{4}\lambda = 0, \quad \text{MADPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}_{\beta}}(rew) = \frac{5}{4} - \frac{3}{8}\lambda = -\frac{1}{4}.$$

So, the randomized scheduler  $\mathfrak{S}_{1/2}$  is better than the deterministic schedulers  $\mathfrak{S}_{\alpha}$  and  $\mathfrak{S}_{\beta}$ . In Figure 2, we depict the MAD in comparison to the expected value of any randomized scheduler for  $\mathcal{M}$ . The kink in the graph at expected value 1 can be explained by the fact



■ **Figure 2** Plot of MAD and variance over the expected value for schedulers obtained by choosing  $\alpha$  with probability  $p \in [0, 1]$  in the MDP  $\mathcal{M}$  depicted in Figure 1a.

that the MAD contains a summand for  $|1 - \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(\text{rew})|$ . The dotted blue line consists of all points in the MAD-E-plane with the same MADPE as the scheduler  $\mathfrak{S}_{1/2}$  illustrating that this scheduler is in fact optimal as the MADPE increases in the direction of the arrow. For comparison, we also depict the variances of randomized schedulers over the expectation. Clearly, for any  $\lambda$  the deterministic scheduler choosing  $\beta$  will always be VPE-optimal.

In the next example, we will illustrate that the MADPE fails to guarantee in general that optimal schedulers are eventually reward-maximizing.

► **Example 3.2.** Consider the MDP  $\mathcal{M}$  depicted in Figure 1b for  $p \in (0, 1/3]$ . Always choosing  $\alpha$  in state  $s_{dec}$  maximizes the expected value. Under this scheduler, the expected value is  $3p \leq 1$  as moving from state  $s_1$  to state  $s_{dec}$  takes two steps in expectation. So, under any scheduler, the expected value lies between 0 and 1. So, all paths leading via  $s_1$  yield a reward above the expected value, while only the path going directly to *goal* from  $s_{init}$  yields a reward below the expected value. For the MAD under a scheduler  $\mathfrak{S}$ , we obtain  $\text{MAD}_{\mathcal{M}}^{\mathfrak{S}}(\text{rew}) = 2 \cdot (1 - p) \cdot \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(\text{rew})$  (see the full version [6] for the calculations).

For a given  $\lambda > \frac{1}{2}$ , we can choose  $p \in (0, 1/3]$  such that  $\lambda > \frac{1}{2(1-p)}$  and hence  $\lambda \cdot 2 \cdot (1-p) > 1$ . Now, under any scheduler  $\mathfrak{S}$ , the MADPE for parameter  $\lambda$  is

$$\text{MADPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}}(\text{rew}) = \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(\text{rew}) - \lambda \cdot 2 \cdot (1-p) \cdot \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(\text{rew}) = (1 - \lambda \cdot 2 \cdot (1-p)) \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(\text{rew}).$$

As  $1 - \lambda \cdot 2 \cdot (1-p) < 0$ , a scheduler maximizing the MADPE has to minimize the expected value of *rew*. In  $\mathcal{M}_p$ , this means always choosing  $\beta$ . So, for any  $\lambda > \frac{1}{2}$ , there is an MDP in which optimal schedulers have to minimize the future expected rewards no matter how large the accumulated reward already is.

### 3.2 Sufficiently small parameters $\lambda$

As we have seen, the MADPE as an objective does not in general guarantee that optimal schedulers are ERMax-schedulers. In this section, we now show that this desirable property is guaranteed if the risk-aversion parameter  $\lambda$  is at most  $\frac{1}{2}$ .

By Lemma 2.1, we already know that we can restrict ourselves to reward-based schedulers when optimizing the MADPE. For two reward-based schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  and a natural number  $k$ , we define the reward-based scheduler  $\mathfrak{S} \uparrow_k \mathfrak{T}$  on state-reward-pairs  $(s, w) \in S \times \mathbb{N}$  by  $(\mathfrak{S} \uparrow_k \mathfrak{T})(s, w) = \begin{cases} \mathfrak{S}(s, w) & \text{if } w < k, \\ \mathfrak{T}(s, w) & \text{if } w \geq k \end{cases}$ , where we view  $\mathfrak{S}$  and  $\mathfrak{T}$  as functions from  $S \times \mathbb{N}$  to distributions over actions.

For risk-aversion parameters  $\lambda$  of at most  $1/2$ , the following theorem implies that optimal schedulers for the MADPE can be chosen to be ERMax-schedulers.



► **Theorem 3.3.** *Let  $\mathcal{M} = (S, Act, P, s_{init}, rew, goal)$  be an MDP satisfying Assumption 1 and let  $\lambda \in (0, \frac{1}{2}]$  be a parameter for the MADPE. Further, let  $\mathfrak{T}$  be a memoryless deterministic scheduler with  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{T}}(rew) = \mathbb{E}_{\mathcal{M},s}^{\max}(rew)$ . Let  $k = \lceil \mathbb{E}_{\mathcal{M}}^{\max}(rew) \rceil$ . Then, for any reward-based scheduler  $\mathfrak{S}$ , we have  $\text{MADPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}}(rew) \leq \text{MADPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S} \uparrow k \mathfrak{T}}(rew)$ .*

The theorem is shown by expressing the MADPE using conditional expectations under the condition that the reward exceeds the bound  $k$ . Note that the theorem implies that it does not matter which expectation optimal scheduler  $\mathfrak{T}$  is chosen after a reward of at least  $\mathbb{E}_{\mathcal{M}}^{\max}(rew)$  has been accumulated.

### 3.3 Computing the maximal MADPE

Theorem 3.3 tells us that the value  $\text{MADPE}[\lambda]_{\mathcal{M}}^{\max}$  in an MDP  $\mathcal{M}$  for  $\lambda \in (0, 1/2]$  is the supremum of  $\text{MADPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}}$  over all reward-based schedulers  $\mathfrak{S}$  that behave according to a fixed memoryless deterministic scheduler  $\mathfrak{T}$  maximizing the expected reward as soon as a reward of more than  $\mathbb{E}_{\mathcal{M}}^{\max}(rew)$  has been accumulated. Let us denote the set of such schedulers by  $\text{Sched}_{\mathcal{M}}^{\mathfrak{T}}$ .

The result shares some similarity with the results in [5] on the computation of maximal conditional expected rewards under the condition that a set of target states is reached. In both cases, a reward-based scheduler that has to keep track of the accumulated reward up to some bound  $B$  has to be computed. The bound  $B$ , however, is obtained quite differently. Here, the maximal expected accumulated reward can be used as this bound. The bound in [5] is in general much larger (although also exponential). Similar reward-based schedulers are also necessary for the model-checking of temporal formulas with certain reward operators [10] and for the optimization of the variance-penalized expectation [25].

We are now in the position to provide a model transformation such that afterwards we can restrict ourselves to memoryless schedulers. Given the MDP  $\mathcal{M} = (S, Act, P, s_{init}, rew, goal)$ , let  $k = \lceil \mathbb{E}_{\mathcal{M}}^{\max}(rew) \rceil$  and let  $\ell$  be the largest reward of a state-weight pair in  $\mathcal{M}$ . We now define the MDP  $\mathcal{N} = (S', Act', P', s'_{init}, rew', goal')$ .

The state space  $S' = S \times \{0, \dots, k + \ell - 1\} \cup \{goal'\}$  and represents states together with the reward that has been accumulated so far, as well as a new trap state  $goal'$ . The initial state is  $s'_{init} = (s_{init}, 0)$ . The set of actions is extended by one new action  $\tau$ . The transition probability function  $P'$  for  $(s, w) \in S \times \{0, \dots, k + \ell - 1\}$  and  $\alpha \in Act$  is given by  $P'((s, w), \alpha, (t, v)) = P(s, \alpha, t)$  if  $w \leq k - 1$  and  $v = w + rew(s, \alpha)$ , and is set to 0 otherwise. So, in all states in  $S \times \{k, \dots, k + \ell - 1\}$  and in  $\{goal'\} \times \{0, \dots, k - 1\}$  none of the actions in  $Act$  are enabled. Instead in these states the new action  $\tau$  is enabled and leads to the trap state  $goal'$  with probability 1. The reward function is 0 on all state-action pairs containing an action from  $Act$ . Only the new action  $\tau$  gets assigned a reward by

$$\begin{aligned} rew'((goal, w)) &= w && \text{for all } w \in \{0, \dots, k + \ell - 1\} \quad \text{and} \\ rew'((s, w)) &= w + \mathbb{E}_{\mathcal{M},s}^{\max}(rew) && \text{for } s \in S \setminus \{goal\} \text{ and } w \in \{k, \dots, k + \ell - 1\}. \end{aligned}$$

So, in  $\mathcal{N}$ , rewards are only received in the very last step when entering the trap state  $goal'$ .

Now, a scheduler  $\mathfrak{S} \in \text{Sched}_{\mathcal{M}}^{\mathfrak{T}}$  for  $\mathcal{M}$  can be seen as a memoryless scheduler for  $\mathcal{N}$  and vice versa: The scheduler  $\mathfrak{S}$  makes decision for all state-reward pairs  $(s, w)$  with  $s \neq goal$  and  $w < \mathbb{E}_{\mathcal{M}}^{\max}(rew)$ . For higher values of accumulated reward, it switches to the behavior of the memoryless scheduler  $\mathfrak{T}$ . A memoryless scheduler for  $\mathcal{N}$  has to choose a probability distribution over  $Act$  on the same pairs  $(s, w)$ . For higher values of  $w$  or for pairs  $(goal, w)$ , only action  $\tau$  is enabled in  $\mathcal{N}$ . So, with a slight abuse of notation, we interpret schedulers in  $\text{Sched}_{\mathcal{M}}^{\mathfrak{T}}$  for  $\mathcal{M}$  also as memoryless schedulers for  $\mathcal{N}$  and vice versa.

► **Remark 3.4.** As reward-based schedulers are sufficient to maximize the MADPE and in  $\mathcal{N}$  rewards are only received in the last step, we can conclude that memoryless schedulers are sufficient to maximize the MADPE in  $\mathcal{N}$ .

► **Lemma 3.5.** *Given  $\mathcal{M}$  and  $\mathcal{N}$  as above, a scheduler  $\mathfrak{S} \in \text{Sched}_{\mathcal{M}}^{\mathfrak{X}}$  and  $\lambda \in (0, 1/2]$ , we have  $\text{MADPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}}(\text{rew}) = \text{MADPE}[\lambda]_{\mathcal{N}}^{\mathfrak{S}}(\text{rew}')$ .*

We utilize the MDP  $\mathcal{N}$  to compute the maximal MADPE via a quadratic program:

► **Theorem 3.6.** *Let  $\mathcal{M}$  be an MDP with non-negative rewards and  $\lambda \in (0, 1/2]$ . Then,  $\text{MADPE}[\lambda]_{\mathcal{M}}^{\max}$  is the optimal solution to a linearly-constrained quadratic program that can be constructed from  $\mathcal{M}$  and  $\lambda$  in exponential time.*

Note that the MDP  $\mathcal{N}$  can be constructed in exponential time from  $\mathcal{M}$  as the numerical value of the maximal expected value  $\mathbb{E}_{\mathcal{M}}^{\max}(\text{rew})$  is at most exponentially large in the size of  $\mathcal{M}$ . So, it is sufficient to construct a quadratic program from  $\mathcal{N}$  in polynomial time. In the sequel, we provide the construction of the quadratic program and prove its correctness.

We start by providing linear constraints that specify the possible combinations of expected frequencies of state-action-pairs under some scheduler. We use variables  $x_{s,w,\alpha}$  for all  $s \in S$ ,  $w \in \{0, 1, \dots, k + \ell - 1\}$ , and  $\alpha \in \text{Act}'((s, w))$ . For these variables, we require

$$x_{s,w,\alpha} \geq 0, \quad \text{and} \quad (1)$$

$$\sum_{\alpha \in \text{Act}(s)} x_{s,w,\alpha} = \sum_{t \in S, \beta \in \text{Act}(t)} x_{t,w-\text{rew}(t,\beta),\beta} \cdot P(t, \beta, s) + \mathbb{1}_{(s,w)=(s_{\text{init}},0)} \quad (2)$$

where  $\mathbb{1}_{(s,w)=(s_{\text{init}},0)} = 1$  iff  $s = s_{\text{init}}$  and  $w = 0$ , and  $\mathbb{1}_{(s,w)=(s_{\text{init}},0)} = 0$  otherwise. In any solution to these two constraints, the variables  $x_{s,w,\alpha}$  represent the expected frequency with which action  $\alpha$  is chosen in state  $(s, w)$  under some scheduler. This is made precise below.

Rewards are only accumulated on the final transitions from a state  $(s, w)$  to  $\text{goal}'$  via action  $\tau$  for  $s = \text{goal}$  or  $w \geq k$ . As these transitions lead to the absorbing state with probability 1, the expected frequency with which the action  $\tau$  is chosen is the probability with which the respective transition is taken. So, we can encode the expected value in an auxiliary variable  $e$  defined via the constraint

$$e = \sum_{w=0}^{k-1} x_{\text{goal},w,\tau} \cdot w + \sum_{w=k}^{k+\ell-1} \sum_{s \in S} x_{s,w,\tau} \cdot (w + \mathbb{E}_{\mathcal{M},s}^{\max}(\text{rew})). \quad (3)$$

► **Lemma 3.7.** *For any solution vector to constraints (1) – (3), there is a scheduler  $\mathfrak{S}$  for  $\mathcal{N}$  such that  $\Pr_{\mathcal{N}}^{\mathfrak{S}}(\diamond(s, w)) = x_{s,w,\tau}$  for all  $(s, w)$  with  $s = \text{goal}$  or  $w \geq k$  and such that  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(\text{rew}) = e$ ; and vice versa.*

Now, we can use these auxiliary variables to encode the MADPE as an objective function:

$$\text{maximize } e - \lambda \left( \sum_{w=0}^{k-1} x_{\text{goal},w,\tau} \cdot |w - e| + \sum_{w=k}^{k+\ell-1} \sum_{s \in S} x_{s,w,\tau} \cdot |w + \mathbb{E}_{\mathcal{M},s}^{\max}(\text{rew}) - e| \right) \quad (4)$$

This function still contains the absolute value operator. However, all absolute value terms occur with a negative sign. Therefore, we can use further variables  $g_i$  for  $i \in \{0, \dots, k - 1\}$  and  $h_{s,w}$  for  $(s, w) \in S \times \{k, \dots, k + \ell - 1\}$  to capture the absolute value. The following constraints state that these variables are at least as big as the respective absolute value terms. For  $w \in \{0, \dots, k - 1\}$ , we require

$$g_w \geq w - e \quad \text{and} \quad -g_w \leq w - e. \quad (5)$$

For  $(s, w) \in S \times \{k, \dots, k + \ell - 1\}$ , we require

$$h_{s,w} \geq w + \mathbb{E}_{\mathcal{M},s}^{\max}(rew) - e \quad \text{and} \quad -h_{s,w} \leq w + \mathbb{E}_{\mathcal{M},s}^{\max}(rew) - e. \quad (6)$$

The new objective function can now be written as

$$\text{maximize} \quad e - \lambda \left( \sum_{w=0}^{k-1} x_{goal,w,\tau} \cdot g_w + \sum_{w=k}^{k+\ell-1} \sum_{s \in S} x_{s,w,\tau} \cdot h_{s,w} \right). \quad (7)$$

► **Theorem 3.8.** *The optimal solution to (7) under constraints (1) - (3), (5), and (6) is the maximal MADPE  $\text{MADPE}[\lambda]_{\mathcal{N}}^{\max}$ .*

**Proof.** As all variables are non-negative, the variables  $g_w$  with  $0 \leq w \leq k - 1$  and  $h_{s,w}$  with  $w \geq k$  in the objective function (7) occur under a negative sign. To maximize the objective function, these variables hence have to be set to the minimal possible values given the value of the variable  $e$ . By constraints (5) and (6), these minimal possible values are the values  $|w - e|$  and  $|w + \mathbb{E}_{\mathcal{M},s}^{\max}(rew) - e|$ , respectively. So, the optimal value of this quadratic objective function is the same as of the objective function (4), which directly encodes the MADPE. ◀

### 3.4 Computational hardness of the MADPE

The complexity class PP [14] is characterized as the class of languages  $\mathcal{L}$  that have a probabilistic polynomial-time bounded Turing machine  $M_{\mathcal{L}}$  such that  $\tau \in \mathcal{L}$  if and only if  $M_{\mathcal{L}}$  accepts  $\tau$  with probability at least  $1/2$  for all words  $\tau$ . We will show PP-hardness under polynomial-time Turing reductions. So, for the reduction, we allow querying an oracle for the problem we reduce to. A polynomial time algorithm for a problem that is PP-hard under polynomial Turing reductions would imply that the polynomial hierarchy collapses [28].

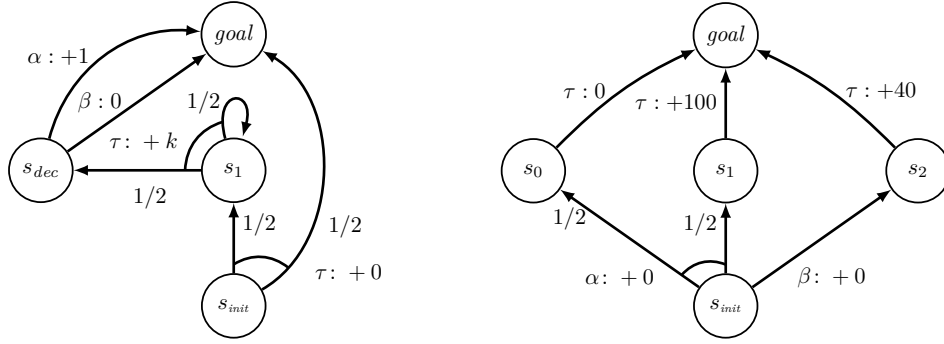
► **Theorem 3.9.** *Deciding for an acyclic Markov chain  $\mathcal{M}$  and a threshold  $\vartheta \in \mathbb{Q}$  whether  $\text{MAD}_{\mathcal{M}}(rew) \geq \vartheta$  is PP-hard under polynomial-time Turing reductions.*

**Proof sketch.** We reduce from the following problem that is shown to be PP-hard in [16]: Given an acyclic Markov chain  $\mathcal{M} = (S, P, s_{init}, rew)$ , and a natural number  $t$ , decide whether  $\Pr_{\mathcal{M}}(rew > t) \geq 1/2$ . We first show that the exact value  $\text{MAD}_{\mathcal{M}}(rew)$  can be computed in acyclic Markov chains via a binary search using polynomially many calls to an oracle for the threshold problem. Then, we prove that  $\Pr_{\mathcal{M}}(rew > t)$  can be computed by comparing the MAD in two variations of  $\mathcal{M}$  that ensure that the expected value of  $rew$  in these variations is  $t$  and  $t + 1/2$ , respectively. ◀

► **Corollary 3.10.** *Deciding for an acyclic Markov chain  $\mathcal{M}$ ,  $\lambda \in \mathbb{Q}_+$  and  $\vartheta \in \mathbb{Q}$  if  $\text{MADPE}[\lambda]_{\mathcal{M}}(rew) \geq \vartheta$  is PP-hard under polynomial-time Turing reductions.*

## 4 Semi-deviation measure-penalized expectation

To overcome the restrictions on the parameter  $\lambda$  for the MADPE or to overcome the undesirable behavior observed for the VPE, one might be tempted to consider the semi-MAD (SMAD) or the semi-variance as a deviation measure that only considers outcomes below the expected value as a measure for the penalty.



(a) The MDP  $\mathcal{M}$  used in Example 4.1. (b) The MDP  $\mathcal{M}$  used in Example 4.2.

■ **Figure 3** Two example MDPs for phenomena of the SVPE.

### Semi-MAD-penalized expectation

We define  $\text{SMAD}(X) = \mathbb{E}(\max(0, \mathbb{E}(X) - X))$  for a random variable  $X$ . So, all outcomes above the expected value do not contribute to the SMAD. However, the SMAD is always half the MAD, i.e.,  $\text{SMAD}(X) = \text{MAD}(X)/2$ , as one can easily compute (see [6]). So, using the SMAD as a penalty term is the same as using the MAD besides a rescaling of the penalty factor  $\lambda$  by a factor of 2.

### Semi-variance-penalized expectation (SVPE)

We now define the semi-variance, to only treat outliers below the expected value with a quadratic penalty. However we will see that SVPE-optimal schedulers might still have to be ERMin-schedulers. We define the semi-variance by ignoring outliers above the expected value as follows  $\text{SV}(X) := \mathbb{E}((\min(X - \mathbb{E}(X), 0))^2)$ . Applied to the accumulated reward in an MDP  $\mathcal{M} = (S, Act, P, s_{init}, rew, goal)$ , we define  $\text{SV}_{\mathcal{M}}^{\mathfrak{S}}(rew) := \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}((\min(rew - \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(rew), 0))^2)$  for schedulers  $\mathfrak{S}$ . Using this as a penalty, we obtain the SVPE for a parameter  $\lambda$

$$\text{SVPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}}(rew) = \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(rew) - \lambda \cdot \text{SV}_{\mathcal{M}}^{\mathfrak{S}}(rew)$$

and define the optimal value  $\text{SVPE}[\lambda]_{\mathcal{M}}^{\max}(rew)$  as usual. Besides the possible necessity of ERMin-schedulers, we will see that randomization is necessary to optimize the SVPE in contrast to the VPE, for which optimal deterministic (finite-memory) schedulers exist [25].

► **Example 4.1** (ERMin-schedulers). Let  $\lambda > 0$  be a parameter for the SVPE. Consider the MDP  $\mathcal{M}$  depicted in Figure 3a where the weight  $k$  is some natural number  $k > 1/\lambda$ . First, observe that under any scheduler  $\mathfrak{S}$ , we have  $k \leq \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(rew) \leq k + 1/2$ . Now, let  $\ell \geq 2$  be a natural number and let  $\mathfrak{S}_p$  be a family of schedulers for  $p \in [0, 1]$  that behaves exactly the same on all paths except for the path that reaches  $s_{dec}$  with accumulated reward exactly  $\ell \cdot k$ . In this state,  $\mathfrak{S}_p$  chooses  $\alpha$  with probability  $p$  and  $\beta$  with probability  $1 - p$ .

We now want to compare the SVPE of  $\mathfrak{S}_p$  for  $p > 0$  to the SVPE of  $\mathfrak{S}_0$ . So, let  $\lambda > 0$  be given. First, we define  $E := \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_0}(rew)$  and observe  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) = E + \frac{p}{2^{\ell+1}}$  as the path on which  $\mathfrak{S}_p$  and  $\mathfrak{S}_0$  differ has probability  $\frac{1}{2^{\ell+1}}$ . Furthermore, both schedulers differ only on a path with a reward higher than the maximal possible expected accumulated reward, which is  $k + 1/2$ . This means that the semivariance under  $\mathfrak{S}_p$  will be larger as under  $\mathfrak{S}_0$ . Note that exactly the outcomes with reward at most  $k$  contribute to the semivariance and these outcomes have exactly the same probability under  $\mathfrak{S}_p$  and  $\mathfrak{S}_0$ . However, the expected value under  $\mathfrak{S}_p$  is

higher. We estimate  $\mathbb{S}\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) - \mathbb{S}\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_0}(rew) \geq \frac{1}{2}(E + \frac{p}{2^{\ell+1}})^2 - \frac{1}{2}E^2$  by only considering the increase in the squared distance from the mean for the outcome 0 that occurs with probability  $1/2$  under both schedulers. So, we can conclude  $\mathbb{S}\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) - \mathbb{S}\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_0}(rew) \geq \frac{1}{2}(\frac{2Ep}{2^{\ell+1}} + (\frac{p}{2^{\ell+1}})^2) \geq \frac{Ep}{2^{\ell+1}}$ . For the SVPE, this implies  $\mathbb{S}\mathbb{V}\mathbb{P}\mathbb{E}[\lambda]_{\mathcal{M}}^{\mathfrak{S}_p}(rew) - \mathbb{S}\mathbb{V}\mathbb{P}\mathbb{E}[\lambda]_{\mathcal{M}}^{\mathfrak{S}_0}(rew) \leq \frac{p}{2^{\ell+1}} - \lambda \frac{Ep}{2^{\ell+1}}$ . As  $E \geq k$  and  $\lambda > 1/k$ , the SVPE under scheduler  $\mathfrak{S}_0$  is higher than under  $\mathfrak{S}_p$ . Note that  $\ell \geq 2$  was chosen arbitrarily. So, this argument shows that any scheduler can be improved by always scheduling  $\beta$  in  $s_{dec}$  as soon as the accumulated reward is at least  $2k$ .

For each  $\lambda > 0$ , we have provided an MDP in which optimal schedulers are necessarily ERMin-schedulers. This is exactly the undesirable behavior as for the VPE we aim to overcome. So, the SVPE is not a suitable alternative.

► **Example 4.2.** To conclude, we show that randomization is necessary to maximize the SVPE. Consider the MDP  $\mathcal{M}$  depicted in Figure 3b. Let  $\mathfrak{S}_p$  be the scheduler that chooses action  $\alpha$  with probability  $p$ . Further, let  $\lambda = \frac{1}{100}$ . We compute  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) = 40 + 10p$ . Under  $\mathfrak{S}_p$ , reward 40 is accumulated with probability  $1 - p$  and reward 0 with probability  $p/2$ . So, we obtain  $\mathbb{S}\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) = (1 - p) \cdot (10p)^2 + \frac{p}{2} \cdot (40 + 10p)^2 = 800p + 500p^2 - 50p^3$ . Finally, we compute  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) - \lambda \mathbb{S}\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) = 40 + 2p - 5p^2 + \frac{1}{2}p^3$ . We determine the unique maximum of this expression on the interval  $[0, 1]$  at the zero of its derivative, which lies at  $p \approx 0.206$ . So, randomization is necessary in order to maximize the SVPE in this MDP.

To conclude, let us compute the variance to illustrate that randomization is not increasing the VPE. We obtain  $\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) = \mathbb{S}\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) + \frac{p}{2}(60 - 10p)^2 = 2600p - 100p^2$ . For the VPE for an arbitrary parameter  $\lambda > 0$ , this results in  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) - \lambda \mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_p}(rew) = 40 + 10p - 2600\lambda p + 100\lambda p^2$ . Due to the positive coefficient in front of  $p^2$  this is a parabola opened upwards. So, for any  $\lambda$ , one of the deterministic schedulers with  $p = 0$  or  $p = 1$  is optimal.

## 5 Threshold-based penalty

The MADPE penalizes outcomes below the expected value of the accumulated reward. The computation of the optimal MADPE via a quadratic program of exponential size, however, might not be feasible on large models. A conceptually simpler alternative, for which we will be able to provide a pseudo-polynomial optimization algorithm, is to externally fix a threshold  $t$  and to penalize outcomes below this threshold  $t$ . To this end, we define a threshold-based penalty function  $TBP_t^\lambda: \mathbb{R} \rightarrow \mathbb{R}$  for parameters  $\lambda, t > 0$  by  $TBP_t^\lambda(x) = x - \lambda \cdot \max(t - x, 0)$ .

This function returns  $x$  if  $x$  is at least  $t$  and otherwise penalizes the deviation below the value  $t$  linearly with the penalty factor  $\lambda$ . In an MDP  $\mathcal{M}$ , our goal is now to maximize – by choosing a scheduler  $\mathfrak{S}$  – the threshold-based-penalized expectation (TBPE)

$$\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(TBP_t^\lambda(rew)) = \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(rew) - \lambda \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(\max(t - rew, 0))$$

Note that in a Markov chain  $\mathcal{N}$ , the TBPE agrees with the SMADPE if we set  $t = \mathbb{E}_{\mathcal{N}}(rew)$ .

The main theorem is the following. Omitted proofs can be found in the full version [6].

► **Theorem 5.1.** *Let  $\mathcal{M} = (S, Act, P, s_{init}, rew, goal)$  be an MDP satisfying Assumption 1 and let  $t, \lambda > 0$  be rationals. Then,  $\mathbb{E}_{\mathcal{M}}^{\max}(TBP_t^\lambda(rew))$  and an optimal scheduler can be computed in time polynomial in the size of  $\mathcal{M}$  and in the numerical value of  $t$ .*

The theorem follows from the following lemma:

► **Lemma 5.2.** *Given  $\mathcal{M}$ ,  $t$ , and  $\lambda$  as in Theorem 5.1, we can construct an MDP  $\mathcal{M}'$  with reward function  $rew'$  (that takes rational rewards that may be negative) and with  $|S| \cdot \lceil t \rceil$  many states in time polynomial in  $|S| \cdot \lceil t \rceil$  such that  $\mathbb{E}_{\mathcal{M}}^{\max}(TBP_t^\lambda(rew)) = \mathbb{E}_{\mathcal{M}'}^{\max}(rew')$ .*

**Proof sketch.** The MDP  $\mathcal{M}'$  is an unfolding of the MDP  $\mathcal{M}$  that keeps track of the accumulated reward until it exceeds  $t$ . So, states are extended with a second component specifying the reward accumulated so far. This second component does not change anymore once it reaches  $t$ . For a state action pair  $((s, w), \alpha)$ , the new reward function is defined as  $rew'((s, w), \alpha) = TBP_t^\lambda(w + rew(s, \alpha)) - TBP_t^\lambda(w)$ . The initial state  $(s_{init}, 0)$  is reached via one additional new transition with reward  $TBP_t^\lambda(0)$  (which is negative). ◀

While  $\mathcal{M}'$  constructed in this proof has a rational reward function that may be negative, the MDP  $\mathcal{M}'$  does not contain end components. Hence, the maximization of the expected accumulated reward in  $\mathcal{M}'$  can be carried out in polynomial time [7] leading to Theorem 5.1. Furthermore, memoryless deterministic schedulers for  $\mathcal{M}'$  are sufficient for the maximization. These schedulers correspond to deterministic, finite-memory ERMax-schedulers for  $\mathcal{M}$ .

► **Remark 5.3.** The proof of Lemma 5.2 (and Thm. 5.1) works analogously for any penalty function that penalizes outcomes below  $t$ : for any function  $m$  such that  $m(x) = x$  for  $x \geq t$  that is computable in polynomial time on natural numbers, we can construct  $\mathcal{M}'$  with a reward function  $rew'$  with  $|S| \cdot \lceil t \rceil$  many states in time polynomial in  $|S| \cdot \lceil t \rceil$  such that  $\mathbb{E}_{\mathcal{M}}^{\max}(m(rew)) = \mathbb{E}_{\mathcal{M}'}^{\max}(rew')$  (for more details, see [6]). Again,  $\mathcal{M}'$  has no end components and the maximal expected reward in  $\mathcal{M}'$  can be computed in time polynomial in the size of  $\mathcal{M}'$  [7].

Finally, we show a hardness result similar as for the MADPE.

► **Theorem 5.4.** *Given an acyclic Markov chain  $\mathcal{M} = (S, P, s_{init}, rew)$  and  $\vartheta, t \in \mathbb{Q}$ , deciding whether  $\mathbb{E}_{\mathcal{M}}(TBP_t^1(rew)) \geq \vartheta$  is PP-hard under polynomial-time Turing reductions.*

Note that this hardness result holds for a fixed parameter. The choice of this parameter  $\lambda = 1$  is arbitrary. The proof works analogously for any positive parameter  $\lambda > 0$ .

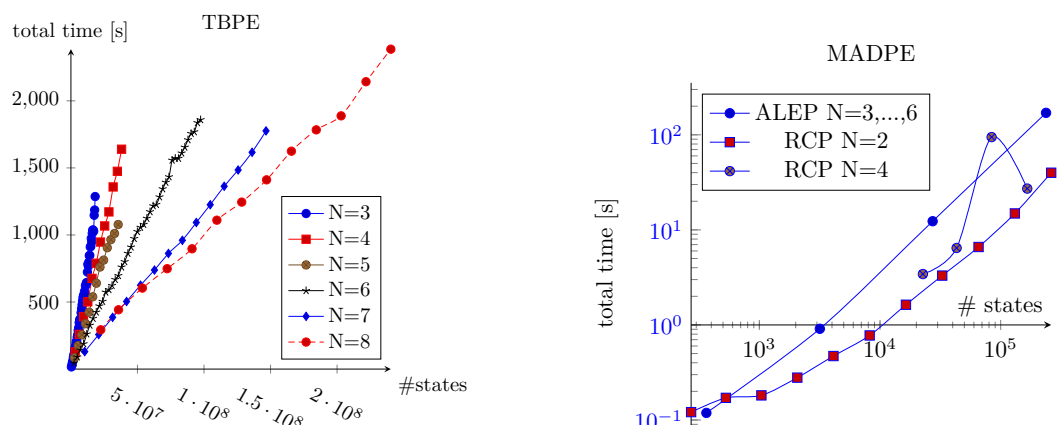
## 6 Prototypical implementation and first experiments

To give a prototypical proof-of-concept for the application of the MADPE and TBPE in practice, we run experiments using the model-checker PRISM [9] and the optimization problem solver Gurobi [15]. The source code for the experiments is available on github<sup>1</sup>. All measurements were done on a machine running Windows 10 Pro 22H2 with an Intel Core i9-9900K CPU and 32GB RAM. We use MDP models written in the PRISM input language (available on the PRISM website<sup>2</sup>) for the asynchronous leader election protocol (ALEP) [17] and, in the case of the MADPE, also for the randomized consensus protocol (RCP) [2]. For both protocols, parameters can be chosen leading to models of different sizes and in the models for both protocols non-negative rewards are specified.

To test our algorithm for the TBPE, for each PRISM model for the ALEP with number of processes  $N = 3, \dots, 8$ , we added a single module which implements the reward counter until reaching the threshold and a new reward definition as in the construction used in the proof of Lemma 5.2. We used the penalty factor  $\lambda = \frac{3}{2}$  in our examples and varied the threshold  $t$ . In Figure 4a, the sizes of the unfolded MDPs for varying values of  $t$ , which are proportional to  $t$ , and the time needed to compute the maximal TBPE are shown. We observe that for this example the required time grows approximately linearly with the size of the unfolded MDP and consequently with the numerical value of  $t$ . For the model with  $N = 8$ , which

<sup>1</sup> <https://github.com/experiments-collection/risk-averse-stochastic-shortest-paths>

<sup>2</sup> <https://www.prismmodelchecker.org/>



(a) The number of states of the unfolded MDPs and the time to compute the optimal TBPE for different parameter choices for the ALEP.

(b) Time to build and solve the quadratic program for the maximization of the MADPE.

■ **Figure 4** Experimental evaluation of the algorithms for TBPE and MADPE.

has approximately  $1.8 \cdot 10^7$  many states, and  $t = 13$ , the unfolded MDP has approximately  $2.4 \cdot 10^8$  many states and the computation of the optimal TBPE takes approximately 2385 seconds. More detailed plots for different values for  $N$  can be found in Appendix A.

To test our algorithm for the MADPE using quadratic programs, we use the ALEP and the RCP models with various parameter choices. The parameter  $\lambda$  is set to 0.4. First we run PRISM to obtain a model representation with all states, transitions, rewards and the maximal expected total reward from each state. Second, we run a python script which constructs all the constraints as described in Section 3 to obtain a linearly constrained program with a quadratic objective. The script uses Gurobi [15] to then solve the optimization problem. The diagram in Figure 4b shows the total time for running the toolchain over the number of reachable states of each model according to PRISM's output. For the largest tested models with approximately  $2 \cdot 10^5$  many states, the maximal MADPE could be computed in less than 200 seconds.

## 7 Conclusion

For various deviation measures, we investigated the deviation-measure-penalized expectation as risk-averse objective applied to the maximization of accumulated rewards in MDPs. As known from the literature, the VPE suffers from the fact that optimal schedulers have to be ERMin-schedulers. Surprisingly, this can still be the case for the SVPE. For the MADPE, a different picture arises: If the penalty factor  $\lambda$  is at most  $1/2$ , optimal schedulers can be chosen to be ERMax-schedulers. If  $\lambda > 1/2$ , ERMin-schedulers can be necessary. Finally, the threshold-based penalty mechanism in the TBPE ensures that optimal schedulers are ERMax-schedulers. For an overview of the further results regarding computational complexity and the structure of optimal schedulers see Table 1.

Despite the PP-hardness results for acyclic Markov chains, the first experimental evaluation of the two cases that ensure the existence of optimal ERMax-schedulers, namely the TBPE in general and the MADPE for small penalty factors  $\lambda \leq 1/2$ , indicates that the optimization seems to be possible in reasonable time on models of considerable size. Further experiments on the scalability of the algorithms, however, are left as future work. In addition, future experiments should examine whether the optimal schedulers for the different measures show a reasonable risk-averse behavior in case studies.

We addressed the maximization of accumulated rewards here. As we work with non-negative rewards, the case of minimization is not symmetric and is subject to future investigations. Finally, the studied objectives can be transferred to other random variables such as the mean payoff, which is a further interesting direction for future work.

---

## References

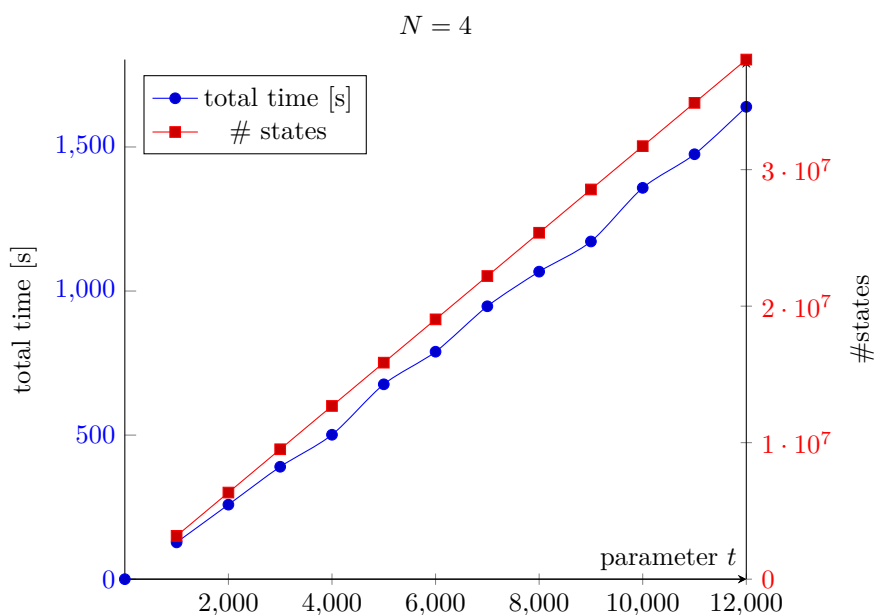
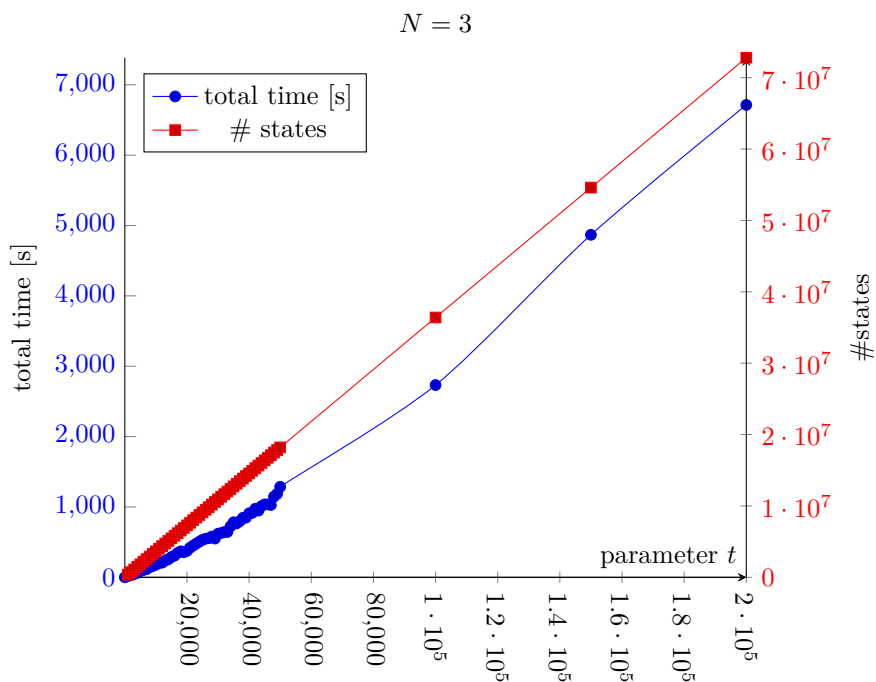
- 1 Mohamadreza Ahmadi, Anushri Dixit, Joel W. Burdick, and Aaron D. Ames. Risk-averse stochastic shortest path planning. In *2021 60th IEEE Conference on Decision and Control (CDC), Austin, TX, USA, December 14-17, 2021*, pages 5199–5204. IEEE, 2021. doi:10.1109/CDC45484.2021.9683527.
- 2 James Aspnes and Maurice Herlihy. Fast randomized consensus using shared memory. *Journal of Algorithms*, 11(3):441–461, 1990. doi:10.1016/0196-6774(90)90021-6.
- 3 Christel Baier, Krishnendu Chatterjee, Tobias Meggendorfer, and Jakob Piribauer. Entropic risk for turn-based stochastic games. In Jérôme Leroux, Sylvain Lombardy, and David Peleg, editors, *48th International Symposium on Mathematical Foundations of Computer Science, MFCS 2023, August 28 to September 1, 2023, Bordeaux, France*, volume 272 of *LIPICs*, pages 15:1–15:16. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPICs.MFCS.2023.15.
- 4 Christel Baier, Marcus Daum, Clemens Dubsclaff, Joachim Klein, and Sascha Klüppelholz. Energy-utility quantiles. In Julia M. Badger and Kristin Yvonne Rozier, editors, *NASA Formal Methods - 6th International Symposium, NFM 2014, Houston, TX, USA, April 29 - May 1, 2014. Proceedings*, volume 8430 of *Lecture Notes in Computer Science*, pages 285–299. Springer, 2014. doi:10.1007/978-3-319-06200-6\_24.
- 5 Christel Baier, Joachim Klein, Sascha Klüppelholz, and Sascha Wunderlich. Maximizing the conditional expected reward for reaching the goal. In Axel Legay and Tiziana Margaria, editors, *23rd International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, volume 10206 of *Lecture Notes in Computer Science*, pages 269–285. Springer, 2017. doi:10.1007/978-3-662-54580-5\_16.
- 6 Christel Baier, Jakob Piribauer, and Maximilian Starke. Risk-averse optimization of total rewards in markovian models using deviation measures, 2024. arXiv:2407.06887.
- 7 Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Math. Oper. Res.*, 16:580–595, 1991. doi:10.1287/moor.16.3.580.
- 8 Tomás Brázdil, Krishnendu Chatterjee, Vojtech Forejt, and Antonín Kucera. Trading performance for stability in markov decision processes. *J. Comput. Syst. Sci.*, 84:144–170, 2017. doi:10.1016/j.jcss.2016.09.009.
- 9 T. Chen, V. Forejt, M. Kwiatkowska, D. Parker, and A. Simaitis. PRISM-games: A model checker for stochastic multi-player games. In N. Piterman and S. Smolka, editors, *Proc. 19th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'13)*, volume 7795 of *LNCS*, pages 185–191. Springer, 2013. doi:10.1007/978-3-642-36742-7\_13.
- 10 Taolue Chen, Vojtech Forejt, Marta Z. Kwiatkowska, David Parker, and Aistis Simaitis. Automatic verification of competitive stochastic systems. *Formal Methods Syst. Des.*, 43(1):61–92, 2013. doi:10.1007/S10703-013-0183-7.
- 11 EJ Collins. Finite-horizon variance penalised Markov decision processes. *Operations-Research-Spektrum*, 19(1):35–39, 1997.
- 12 Luca de Alfaro. Computing minimum and maximum reachability times in probabilistic systems. In *10th International Conference on Concurrency Theory (CONCUR)*, volume 1664 of *Lecture Notes in Computer Science*, pages 66–81, 1999. doi:10.1007/3-540-48320-9\_7.
- 13 Jerzy A Filar, Lodewijk CM Kallenberg, and Huey-Miin Lee. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989. doi:10.1287/moor.14.1.147.
- 14 John Gill. Computational complexity of probabilistic Turing machines. *SIAM Journal on Computing*, 6(4):675–695, 1977. doi:10.1137/0206049.

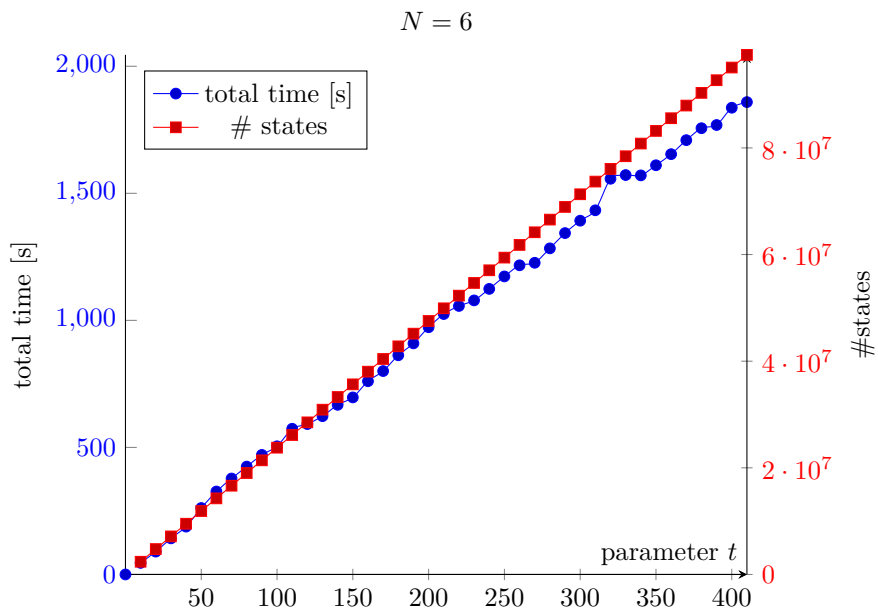
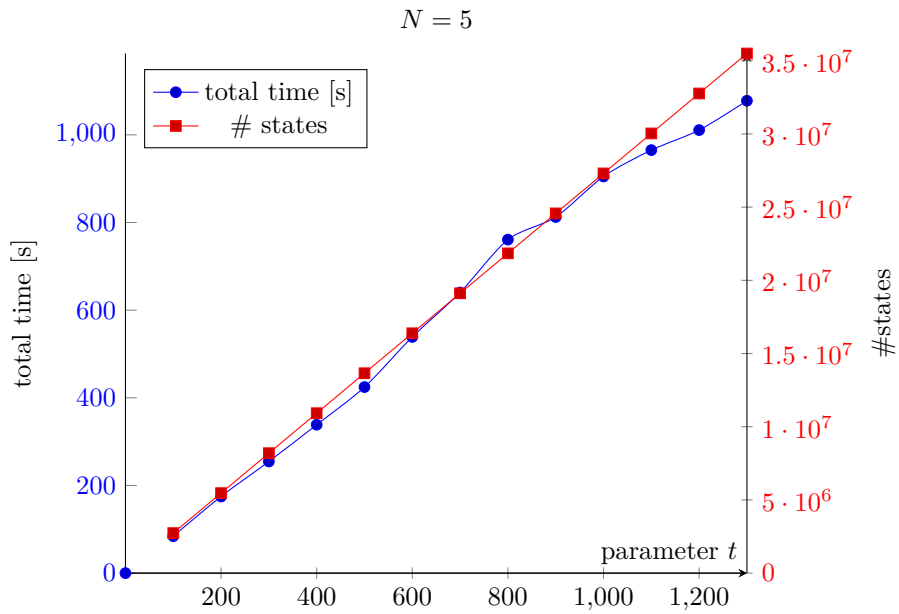


- 15 Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL: <https://www.gurobi.com>.
- 16 Christoph Haase and Stefan Kiefer. The odds of staying on budget. In Magnús M. Halldórsson, Kazuo Iwama, Naoki Kobayashi, and Bettina Speckmann, editors, *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part II*, volume 9135 of *Lecture Notes in Computer Science*, pages 234–246. Springer, 2015. doi:10.1007/978-3-662-47666-6\_19.
- 17 A. Itai and M. Rodeh. Symmetry breaking in distributed networks. *Information and Computation*, 88(1), 1990. doi:10.1016/0890-5401(90)90004-2.
- 18 Hiroshi Konno and Hiroaki Yamazaki. Mean-absolute deviation portfolio optimization model and its applications to tokyo stock market. *Management Science*, 37(5):519–531, 1991. URL: <http://www.jstor.org/stable/2632458>.
- 19 Jan Kretínský and Tobias Meggendorfer. Conditional value-at-risk for reachability and mean payoff in Markov decision processes. In *33rd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 609–618. ACM, 2018. doi:10.1145/3209108.3209176.
- 20 Xiaoteng Ma, Shuai Ma, Li Xia, and Qianchuan Zhao. Mean-semivariance policy optimization via risk-averse reinforcement learning. *J. Artif. Intell. Res.*, 75:569–595, 2022. doi:10.1613/jair.1.13833.
- 21 Petr Mandl. On the variance in controlled Markov chains. *Kybernetika*, 7(1):1–12, 1971. URL: <http://www.kybernetika.cz/content/1971/1/1>.
- 22 Shie Mannor and John N. Tsitsiklis. Mean-variance optimization in Markov decision processes. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML'11*, pages 177–184, Madison, WI, USA, 2011. Omnipress. URL: [https://icml.cc/2011/papers/156\\_icmlpaper.pdf](https://icml.cc/2011/papers/156_icmlpaper.pdf).
- 23 Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. URL: <http://www.jstor.org/stable/2975974>.
- 24 Harry M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Yale University Press, 1959. URL: <http://www.jstor.org/stable/j.ctt1bh4c8h>.
- 25 Jakob Piribauer, Ocan Sankur, and Christel Baier. The variance-penalized stochastic shortest path problem. In Mikolaj Bojanczyk, Emanuela Merelli, and David P. Woodruff, editors, *49th International Colloquium on Automata, Languages, and Programming, ICALP 2022, July 4-8, 2022, Paris, France*, volume 229 of *LIPICs*, pages 129:1–129:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022. doi:10.4230/LIPICs.ICALP.2022.129.
- 26 Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994. doi:10.1002/9780470316887.
- 27 Mickael Randour, Jean-François Raskin, and Ocan Sankur. Percentile queries in multi-dimensional markov decision processes. *Formal Methods Syst. Des.*, 50(2-3):207–248, 2017. doi:10.1007/s10703-016-0262-7.
- 28 Seinosuke Toda. PP is as Hard as the Polynomial-Time Hierarchy. *SIAM Journal on Computing*, 20(5):865–877, 1991. doi:10.1137/0220053.
- 29 Michael Ummels and Christel Baier. Computing quantiles in Markov reward models. In Frank Pfenning, editor, *16th International Conference on Foundations of Software Science and Computation Structures (FoSSaCS)*, volume 7794 of *Lecture Notes in Computer Science*, pages 353–368. Springer, 2013. doi:10.1007/978-3-642-37075-5\_23.
- 30 Tom Verhoeff. Reward variance in Markov chains: A calculational approach. In *Proceedings of Eindhoven FASTAR Days*. Technische Universiteit Eindhoven, 2004.
- 31 Li Xia. Risk-sensitive Markov decision processes with combined metrics of mean and variance. *Production and Operations Management*, 29(12):2808–2827, 2020. doi:10.1111/poms.13252.

**A** Experimental evaluation

In the sequel, the number of states of the unfolded MDPs as well as the time to compute the maximal TBPE as described in Section 6 are depicted for the Asynchronous Leader Election Protocol with parameter  $N = 3, \dots, 8$  and varying values of the parameter  $t$ . The number of states of the unfolded MDPs grows linearly in  $t$  as expected. Interestingly, also the required times seem to grow linearly in  $t$ .





9:20 Risk-Averse Optimization of Total Rewards in Markovian Models

