



# Demonic Variance and a Non-Determinism Score for Markov Decision Processes

Jakob Piribauer  

Technische Universität Dresden, Germany  
Universität Leipzig, Germany

---

## Abstract

This paper studies the influence of probabilism and non-determinism on some quantitative aspect  $X$  of the execution of a system modeled as a Markov decision process (MDP). To this end, the novel notion of *demonic variance* is introduced: For a random variable  $X$  in an MDP  $\mathcal{M}$ , it is defined as  $1/2$  times the maximal expected squared distance of the values of  $X$  in two independent execution of  $\mathcal{M}$  in which also the non-deterministic choices are resolved independently by two distinct schedulers.

It is shown that the demonic variance is between 1 and 2 times as large as the maximal variance of  $X$  in  $\mathcal{M}$  that can be achieved by a single scheduler. This allows defining a non-determinism score for  $\mathcal{M}$  and  $X$  measuring how strongly the difference of  $X$  in two executions of  $\mathcal{M}$  can be influenced by the non-deterministic choices. Properties of MDPs  $\mathcal{M}$  with extremal values of the non-determinism score are established. Further, the algorithmic problems of computing the maximal variance and the demonic variance are investigated for two random variables, namely weighted reachability and accumulated rewards. In the process, also the structure of schedulers maximizing the variance and of scheduler pairs realizing the demonic variance is analyzed.

**2012 ACM Subject Classification** Theory of computation  $\rightarrow$  Logic and verification

**Keywords and phrases** Markov decision processes, variance, non-determinism, probabilism

**Digital Object Identifier** 10.4230/LIPIcs.MFCS.2024.79

**Related Version** *Full Version*: <https://arxiv.org/abs/2406.18727> [26]

**Funding** This work was partly funded by the DFG Grant 389792660 as part of TRR 248 (Foundations of Perspicuous Software Systems).

## 1 Introduction

In software and hardware systems, uncertainty manifests in two distinct forms: *non-determinism* and *probabilism*. Non-determinism emerges from, e.g., unknown operating environments, user interactions, or concurrent processes. Probabilistic behavior arises through deliberate randomization in algorithms or can be inferred, e.g., from probabilities of component failures. In this paper, we investigate the uncertainty in the value  $X$  of some quantitative aspect of a system whose behavior is subject to non-determinism *and* probabilism. On the one hand, we aim to quantify this uncertainty. In the spirit of the variance that quantifies uncertainty in purely probabilistic settings, we introduce the notion of *demonic variance* that generalizes the variance in the presence of non-determinism. On the other hand, we provide a *non-determinism score* (NDS) based on this demonic variance that measures the extent to which the uncertainty of  $X$  can be ascribed to the non-determinism.

As formal models, we use Markov decision processes (MDPs, see, e.g., [29]), one of the most prominent models combining non-determinism and probabilism, heavily used in verification, operations research, and artificial intelligence. The non-deterministic choices in an MDP are resolved by a *scheduler*. Once a scheduler is fixed, the system behaves purely probabilistically.



© Jakob Piribauer;

licensed under Creative Commons License CC-BY 4.0

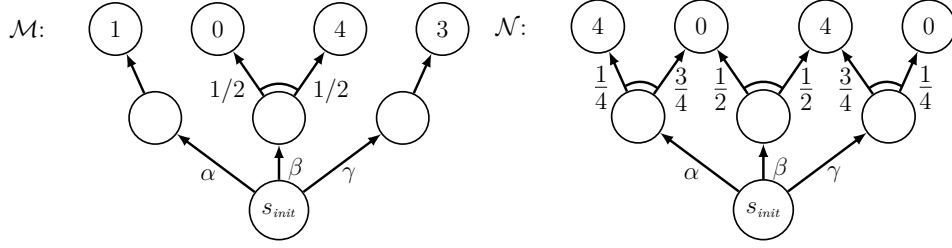
49th International Symposium on Mathematical Foundations of Computer Science (MFCS 2024).

Editors: Rastislav Kráľovič and Antonín Kučera; Article No. 79; pp. 79:1–79:15

Leibniz International Proceedings in Informatics



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



■ **Figure 1** MDPs modeling a communication protocol and the time required to process a message.

**Demonic variance.** For a random variable  $Y$ , the variance is equal to half the expected squared deviation of two independent copies  $Y_1$  and  $Y_2$  of  $Y$ :

$$\mathbb{V}(Y) \stackrel{\text{def}}{=} \mathbb{E}((Y - \mathbb{E}(Y))^2) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \frac{1}{2}\mathbb{E}(Y_1^2 - 2Y_1Y_2 + Y_2^2) = \frac{1}{2}\mathbb{E}((Y_1 - Y_2)^2).$$

For a quantity  $X$  in an MDP  $\mathcal{M}$ , we obtain a random variable  $X_{\mathcal{M}}^{\mathfrak{S}}$  for each scheduler  $\mathfrak{S}$ .<sup>1</sup> The maximal variance  $\mathbb{V}_{\mathcal{M}}^{\max}(X) \stackrel{\text{def}}{=} \sup_{\mathfrak{S}} \mathbb{V}(X_{\mathcal{M}}^{\mathfrak{S}})$  can serve as a measure for the “amount of probabilistic uncertainty” regarding  $X$  present in the MDP. However, in the presence of non-determinism, quantifying the spread of outcomes in terms of the squared deviation of two independent executions of a system gives rise to a whole new meaning: We can allow the non-determinism to be resolved independently as well. To this end, we consider two different scheduler  $\mathfrak{S}_1$  and  $\mathfrak{S}_2$  in two independent copies  $\mathcal{M}_1$  and  $\mathcal{M}_2$  of  $\mathcal{M}$  and define

$$\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_1, \mathfrak{S}_2}(X) \stackrel{\text{def}}{=} \frac{1}{2}\mathbb{E}((X_{\mathcal{M}_1}^{\mathfrak{S}_1} - X_{\mathcal{M}_2}^{\mathfrak{S}_2})^2).$$

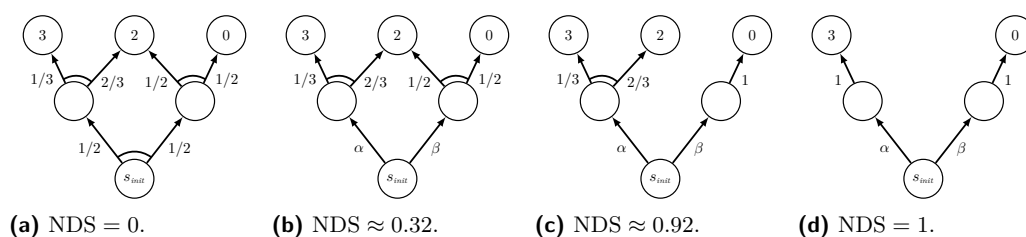
If we now allow for a *demonic* choice of the two schedulers making this uncertainty as large as possible, we arrive at the *demonic variance*  $\mathbb{V}_{\mathcal{M}}^{\text{dem}}(X) \stackrel{\text{def}}{=} \sup_{\mathfrak{S}_1, \mathfrak{S}_2} \mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_1, \mathfrak{S}_2}(X)$  of  $X$  in  $\mathcal{M}$ .

► **Example 1.1.** To illustrate a potential use case, consider a communication network in which messages are processed according to a randomized protocol employed on different hardware at the different nodes of the network. A low worst-case expected processing time of the protocol is clearly desirable. In addition, however, large differences in the processing time at different nodes make buffering necessary and increase the risk of package losses.

Consider the MDPs  $\mathcal{M}$  and  $\mathcal{N}$  in Fig. 1 modeling such a communication protocol. Initially, a non-deterministic choice between  $\alpha$ ,  $\beta$ , and  $\gamma$  is made. Then, a final node containing the processing time  $X$  is reached according to the depicted distributions. In both MDPs, the expected value of  $X$  lies between 1 and 3 for all schedulers  $\mathfrak{S}$  – with the values 1 and 3 being realized by  $\alpha$  and  $\gamma$ . Furthermore, as the outcomes lie between 0 and 4, the distribution over outcomes leading to the highest possible variance of 4 is the one that takes value 0 and 4 with probability  $\frac{1}{2}$  each, which is realized by a scheduler choosing  $\beta$ . So,  $\mathbb{V}_{\mathcal{M}}^{\max}(X) = \mathbb{V}_{\mathcal{N}}^{\max}(X) = 4$ .

However, the demonic variances are different: Our results will show that the demonic variance is obtained by a pair of *deterministic* schedulers that do not randomize over the non-deterministic choices. In  $\mathcal{M}$ , we can easily check that no combination of such schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  leads to a value  $\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}, \mathfrak{T}}(X)$  of more than  $4 = \mathbb{V}_{\mathcal{M}}^{\beta, \beta}(X)$  where  $\beta$  denotes the scheduler that chooses  $\beta$  with probability 1. In  $\mathcal{N}$ , on the other hand, the demonic variance is  $\mathbb{V}_{\mathcal{N}}^{\text{dem}}(X) = \mathbb{V}_{\mathcal{N}}^{\alpha, \gamma}(X) = \frac{1}{2}\mathbb{E}((X_{\mathcal{N}_1}^{\alpha} - X_{\mathcal{N}_2}^{\gamma})^2) = \frac{1}{2}(\frac{10}{16} \cdot 16) = 5$ .

<sup>1</sup> Note that the notation  $X_{\mathcal{M}}^{\mathfrak{S}}$  here differs from the notation used in the technical part of the paper.



■ **Figure 2** Example MDPs with different non-determinism scores (NDSs).

So, despite the same maximal variance and range of expected values, the worst-case expected squared deviation between two values of  $X$  in independent executions is worse in  $\mathcal{N}$  than in  $\mathcal{M}$ . Hence, we argue that the protocol modeled by  $\mathcal{M}$  should be preferred.

**Non-determinism score (NDS).** By the definition of the demonic variance, it is clear that  $\mathbb{V}_{\mathcal{M}}^{dem}(X) \geq \mathbb{V}_{\mathcal{M}}^{\max}(X)$ . Under mild assumptions ensuring the well-definedness, we will prove that  $\mathbb{V}_{\mathcal{M}}^{dem}(X) \leq 2\mathbb{V}_{\mathcal{M}}^{\max}(X)$ , too. So, the demonic variance is between 1 and 2 times as large as the maximal variance. We use this to define the *non-determinism score (NDS)*

$$\text{NDS}(\mathcal{M}, X) \stackrel{\text{def}}{=} \frac{\mathbb{V}_{\mathcal{M}}^{dem}(X) - \mathbb{V}_{\mathcal{M}}^{\max}(X)}{\mathbb{V}_{\mathcal{M}}^{\max}(X)} \in [0, 1].$$

The NDS captures how much larger the expected squared deviation of two outcomes can be made by resolving the non-determinism in two executions independently compared to how large it can be solely due to the probabilism under a single resolution of the non-determinism.

► **Example 1.2.** For an illustration of the NDS, four simple MDPs and their NDSs are depicted in Figure 2. In all of the MDPs except for the first one, a scheduler has to make a (randomized) choice over actions  $\alpha$  and  $\beta$  in the initial state  $s_{init}$ . Afterwards one of the terminal states is reached according to the specified probabilities. The terminal states are equipped with a weight that specifies the value of  $X$  at the end of the execution. For all of these MDPs, the maximal variance can be computed by expressing the variance in terms of the probability  $p$  that  $\alpha$  is chosen and maximizing the resulting quadratic function. In the interest of brevity, we do not present these computations. The pair of (deterministic) schedulers realizing the demonic variance always consists of the scheduler choosing  $\alpha$  and the scheduler choosing  $\beta$  making it easy to compute the demonic variance in these examples.

**Potential applications.** First of all, the demonic variance might serve as the basis for refined guarantees on the behavior of systems, in particular, when employed in different environments. As a first result in this direction, we will prove an analogue to Chebyshev's Inequality using the demonic variance. Further, as illustrated in Example 1.1, achieving a low demonic variance or NDS can be desirable when designing systems. Hence, a reasonable synthesis task could be to design a system ensuring a high expected value of a quantity  $X$  while keeping the demonic variance of  $X$  below a threshold.

Secondly, the demonic variance and the NDS can serve to enhance the explainability of a system's behavior, a topic of growing importance in the area of formal verification (see, e.g., [4] for an overview). More concretely, the NDS can be understood as a measure assigning *responsibility* for the scattering of a quantity  $X$  in different executions to the non-determinism and the probabilism present in the system, respectively. Further, considering the NDS for different starting states makes it possible to pinpoint regions of the state space in which the

non-determinism has a particularly high influence. Notions of responsibility that quantify to which extent certain facets of the behavior of a system can be ascribed to certain components, states, or events have been studied in various settings [10, 33, 6, 25, 5].

Finally, the NDS can also be understood as a measure for the power of control when non-determinism models controllable aspects of a system. This interpretation could be useful, e.g., when designing exploration strategies in reinforcement learning. Here, the task is to learn good strategies as fast as possible by interacting with a system. One of the main challenges is to decide which regions of the state space to explore (see [21] for a recent survey). Estimations for the NDS starting from different states could be useful here: States from which the NDS is high might be more promising to explore than states from which the NDS is low as the difference in received rewards from such a state is largely subject to randomness.

**Contributions.** Besides establishing general results for the demonic variance and the NDS, we investigate the two notions for weighted reachability and accumulated rewards. For weighted reachability, terminal states of an MDP are equipped with a weight that is received if an execution ends in this state. For accumulated rewards, all states are assigned rewards that are summed up along an execution. The main contributions of this paper are as follows.

- We introduce the novel notions of demonic variance and non-determinism score. For general random variables  $X$ , we prove that the demonic variance is at most twice as large as the maximal variance. Furthermore, we prove an analogue of Chebyshev’s inequality. For the non-determinism score, we establish consequences of a score of 0 or 1.
- In the process, we prove a result of independent interest using a topology on the space of schedulers that states that convergence with respect to this topology implies convergence of the corresponding probability measures.
- For weighted reachability, we show that the maximal and the demonic variance can be computed via quadratic programs. For the maximal variance, this results in a polynomial-time algorithm; for the demonic variance, in a separable bilinear program of polynomial size yielding an exponential time upper bound. Further, we establish that there is a memoryless scheduler maximizing the variance and a pair of memoryless deterministic schedulers realizing the demonic variance.
- For accumulated rewards, we prove that the maximal variance and an optimal finite-memory scheduler can be computed in exponential time. Further, we prove that the demonic variance is realized by a pair of deterministic finite-memory schedulers which can be computed via a bilinear program of exponential size.

**Related work.** We are not aware of investigations of notions similar to the demonic variance for MDPs. Previous work on the variance in MDPs usually focused on the minimization of the variance. In [24], the problem to find schedulers that ensure a certain expected value while keeping the variance below a threshold is investigated for accumulated rewards in the finite horizon setting. It is shown that deciding whether there is a scheduler ensuring variance 0 is NP-hard. In [22], the minimization of the variance of accumulated rewards and of the mean payoff is addressed with a focus on optimality equations and no algorithmic results. The variance of accumulated weights in Markov chains is shown to be computable in polynomial time in [32]. For the mean payoff, algorithms were given to compute schedulers that achieve given bounds on the expectation and notions of variance and variability in [9].

One objective incorporating the variance that has been studied on MDPs is the variance-penalized expectation (VPE) [16, 13, 28]. Here, the goal is to find a scheduler that maximizes the expected reward minus a penalty factor times the variance. In [28], the objective is studied for accumulated rewards. Methodically, our results for the maximal and demonic

variance of accumulated rewards share similarities with the techniques of [28] and we make use of some results proved there, such as the result that among expectation-optimal schedulers a variance-optimal memoryless deterministic scheduler can be computed in polynomial time. Nevertheless, the optimization of the VPE inherently requires the minimization of the variance. In particular, it is shown in [28] that deterministic schedulers are optimal for the VPE, while randomization is necessary for the maximization of the variance.

Besides the variance, several other notions that aim to bound the uncertainty of the outcome of some quantitative aspect in MDPs have been studied – in particular, in the context of risk-averse optimization: Given a probability  $p$ , quantiles for a quantity  $X$  are the best bound  $B$  such that  $X$  exceeds  $B$  with probability at most  $p$  in the worst or best case. For accumulated rewards in MDPs, quantiles have been studied in [31, 3, 17, 30]. The *conditional value-at-risk* is a more involved measure that quantifies how far the probability mass of the tail of the probability distribution lies above a quantile. In [20], this notion has been investigated for weighted reachability and mean payoff; in [27] for accumulated rewards. A further measure incentivizing a high expected value while keeping the probability of low outcomes small is the entropic risk measure. For accumulated rewards, this measure has been studied in [2] in stochastic games that extend MDPs with an adversarial player.

Finally, as the demonic variance is a measure that looks at a system across different executions, there is a conceptual similarity to hyperproperties [12, 11]. For probabilistic systems, logics expressing hyperproperties that allow to quantify over different executions or schedulers have been introduced in [1, 15].

## 2 Preliminaries

**Notations for Markov decision processes.** A *Markov decision process* (MDP) is a tuple  $\mathcal{M} = (S, Act, P, s_{init})$  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, and  $s_{init} \in S$  the initial state. 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  $Act(s)$ . We further require that  $Act(s) \neq \emptyset$  for all  $s \in S$ . If for a state  $s$  and all actions  $\alpha \in Act(s)$ , we have  $P(s, \alpha, s) = 1$ , we say that  $s$  is *absorbing*. 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  $\pi = s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{k-1} s_k$ ,  $P(\pi) = P(s_0, \alpha_0, s_1) \cdot \dots \cdot P(s_{k-1}, \alpha_{k-1}, s_k)$  denotes the probability of  $\pi$  and  $last(\pi) = s_k$  its last state. Often, we equip MDPs with a reward function  $rew: S \times Act \rightarrow \mathbb{N}$ . The *size* of  $\mathcal{M}$  is the sum of the number of states plus the total sum of the encoding lengths in binary of the non-zero probability values  $P(s, \alpha, s')$  as fractions of co-prime integers as well as the encoding length in binary of the rewards if a reward function is used. 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. Given two MDPs  $\mathcal{M} = (S, Act, P, s_{init})$  and  $\mathcal{N} = (S', Act', P', s'_{init})$ , we define the (synchronous) product  $\mathcal{M} \otimes \mathcal{N}$  as the tuple  $(S \times S', Act \times Act', P^\otimes, (s_{init}, s'_{init}))$  where we define  $P^\otimes((s, s'), (\alpha, \beta), (t, t')) = P(s, \alpha, t) \cdot P(s', \beta, t')$  for all  $(s, s'), (t, t') \in S \times S'$  and  $(\alpha, \beta) \in Act \times Act'$ .

**Schedulers.** A *scheduler* (also called *policy*) for  $\mathcal{M}$  is a function  $\mathfrak{S}$  that assigns to each finite path  $\pi$  a probability distribution over  $Act(last(\pi))$ . If  $\mathfrak{S}(\pi) = \mathfrak{S}(\pi')$  for all finite paths  $\pi$  and  $\pi'$  with  $last(\pi) = last(\pi')$ , we say that  $\mathfrak{S}$  is *memoryless*. In this case, we also view schedulers as functions mapping states  $s \in S$  to probability distributions over  $Act(s)$ . A scheduler  $\mathfrak{S}$  is called *deterministic* if  $\mathfrak{S}(\pi)$  is a Dirac distribution for each finite path  $\pi$ , in which case we also view the scheduler as a mapping to actions in  $Act(last(\pi))$ . Given two MDPs  $\mathcal{M} = (S, Act, P, s_{init})$  and  $\mathcal{N} = (S', Act', P', s'_{init})$  and two schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  for  $\mathcal{M}$  and  $\mathcal{N}$ , respectively, we define the product scheduler  $\mathfrak{S} \otimes \mathfrak{T}$  for  $\mathcal{M} \otimes \mathcal{N}$  by defining for a finite path  $\pi = (s_0, t_0) (\alpha_0, \beta_0) (s_1, t_1) \dots (s_k, t_k)$ :  $\mathfrak{S} \otimes \mathfrak{T}(\pi)(\alpha, \beta) = \mathfrak{S}(s_0 \alpha_0 \dots s_k)(\alpha) \cdot \mathfrak{T}(t_0 \beta_0 \dots t_k)(\beta)$  for all  $(\alpha, \beta) \in Act \times Act'$ .

**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 infinite extensions of a finite path  $\pi = s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{k-1} s_k$  starting in state  $s$ , i.e.,  $s_0 = s$ , by assigning to  $Cyl(\pi)$  the probability that  $\pi$  is realized under  $\mathfrak{S}$ , which is  $P^{\mathfrak{S}}(\pi) \stackrel{\text{def}}{=} \prod_{i=0}^{k-1} \mathfrak{S}(s_0 \alpha_0 \dots s_i)(\alpha_i) \cdot P(s_i, \alpha_0, s_{i+1})$ . This can be extended to a unique probability measure on the mentioned  $\sigma$ -algebra. For details, see [29]. For a random variable  $X$ , i.e., a measurable function defined on infinite paths in  $\mathcal{M}$ , we denote the expected value of  $X$  under a scheduler  $\mathfrak{S}$  and state  $s$  by  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X)$ . We define  $\mathbb{E}_{\mathcal{M},s}^{\min}(X) \stackrel{\text{def}}{=} \inf_{\mathfrak{S}} \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X)$  and  $\mathbb{E}_{\mathcal{M},s}^{\max}(X) \stackrel{\text{def}}{=} \sup_{\mathfrak{S}} \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X)$ . 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  $\mathbb{V}_{\mathcal{M},s}^{\mathfrak{S}}(X) \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}((X - \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X))^2) = \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X^2) - \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X)^2$ . We define  $\mathbb{V}_{\mathcal{M},s}^{\max}(X) \stackrel{\text{def}}{=} \sup_{\mathfrak{S}} \mathbb{V}_{\mathcal{M},s}^{\mathfrak{S}}(X)$ . If  $s = s_{init}$ , we sometimes drop the subscript  $s$  in  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}$ ,  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}$  and  $\mathbb{V}_{\mathcal{M},s}^{\mathfrak{S}}(X)$ .

**Mixing schedulers.** Intuitively, we often want to use a scheduler that initially decides to behave like a scheduler  $\mathfrak{S}$  and then to stick to this scheduler with probability  $p$  and to behave like a scheduler  $\mathfrak{T}$  with probability  $1 - p$ . As this intuitive description does not match the definition of schedulers as functions from finite paths<sup>2</sup>, we provide a formal definition: For two schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  and  $p \in [0, 1]$ , we use  $p\mathfrak{S} \oplus (1 - p)\mathfrak{T}$  to denote the following scheduler. For a path  $\pi = s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{k-1} s_k$ , we define for an action  $\alpha$  enabled in  $s_k$

$$(p\mathfrak{S} \oplus (1 - p)\mathfrak{T})(\pi)(\alpha) \stackrel{\text{def}}{=} \frac{p \cdot P^{\mathfrak{S}}(\pi) \cdot \mathfrak{S}(\pi)(\alpha)}{p \cdot P^{\mathfrak{S}}(\pi) + (1 - p) \cdot P^{\mathfrak{T}}(\pi)} + \frac{(1 - p) \cdot P^{\mathfrak{T}}(\pi) \cdot \mathfrak{T}(\pi)(\alpha)}{p \cdot P^{\mathfrak{S}}(\pi) + (1 - p) \cdot P^{\mathfrak{T}}(\pi)}.$$

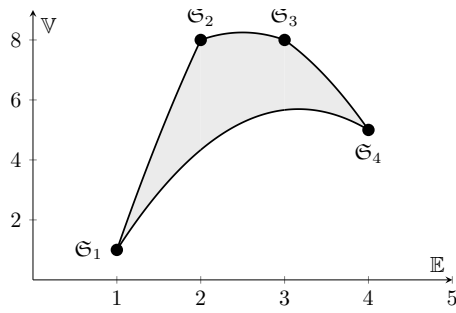
This is well-defined for any path that has positive probability under  $\mathfrak{S}$  or  $\mathfrak{T}$ . The following result is folklore; a proof is included in the full version [26].

► **Proposition 2.1.** *Let the schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  and the value  $p$  be as above. Then, for any path  $\pi = s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{k-1} s_k$ , we have  $P^{p\mathfrak{S} \oplus (1-p)\mathfrak{T}}(\pi) = pP^{\mathfrak{S}}(\pi) + (1 - p)P^{\mathfrak{T}}(\pi)$ .*

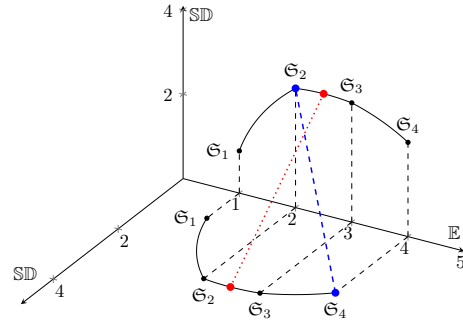
We conclude  $\Pr_{\mathcal{M},s}^{p\mathfrak{S} \oplus (1-p)\mathfrak{T}}(A) = p\Pr_{\mathcal{M},s}^{\mathfrak{S}}(A) + (1 - p)\Pr_{\mathcal{M},s}^{\mathfrak{T}}(A)$  for any measurable set of paths  $A$ . Hence, we can think of the scheduler  $p\mathfrak{S} \oplus (1 - p)\mathfrak{T}$  as behaving like  $\mathfrak{S}$  with probability  $p$  and like  $\mathfrak{T}$  with probability  $(1 - p)$ . In particular, we can also conclude that for a random variable  $X$ , we have  $\mathbb{E}_{\mathcal{M},s}^{p\mathfrak{S} \oplus (1-p)\mathfrak{T}}(X) = p\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(X) + (1 - p)\mathbb{E}_{\mathcal{M},s}^{\mathfrak{T}}(X)$ . For the variance, we obtain the following as shown in full version [26].

► **Lemma 2.2.** *Given  $\mathcal{M}$ ,  $X$ , and two schedulers  $\mathfrak{S}_1$  and  $\mathfrak{S}_2$ , as well as  $p \in [0, 1]$ , let  $\mathfrak{T} = p\mathfrak{S}_1 \oplus (1 - p)\mathfrak{S}_2$ . Then,  $\mathbb{V}_{\mathcal{M}}^{\mathfrak{T}}(X) = p\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_1}(X) + (1 - p)\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_2}(X) + p(1 - p)(\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_1}(X) - \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_2}(X))^2$ .*

<sup>2</sup> This description would be admissible if we allowed stochastic memory updates (see, e.g., [8]).



(a) Possible combinations of variance and expectation in Example 3.2.



(b) Plot of the standard deviation over the expectation on two orthogonal planes.

Figure 3 Graphical illustration of the task to find the demonic variance (see Example 3.2).

**Topology and convergence of measures.** Given a family of topological spaces  $((S_i, \tau_i))_{i \in I}$ , the product topology  $\tau$  on  $\prod_{i \in I} S_i$  is the coarsest topology such that the projections  $p_i: \prod_{i \in I} S_i \rightarrow S_i, (s_i)_{i \in I} \mapsto s_i$  are continuous for all  $i \in I$ . For measures  $(\mu_j)_{j \in \mathbb{N}}$  and  $\mu$  on a measure space  $(\Omega, \Sigma)$  where  $\Omega$  is a metrizable topological space and  $\Sigma$  the Borel  $\sigma$ -algebra on  $\Omega$ , we say that the sequence  $(\mu_j)_{j \in \mathbb{N}}$  *weakly converges* to  $\mu$  if for all bounded continuous functions  $f: \Omega \rightarrow \mathbb{R}$ , we have  $\lim_{j \rightarrow \infty} \int f d\mu_j = \int f d\mu$ . The set of infinite paths  $\Pi_{\mathcal{M}}$  of an MDP  $\mathcal{M}$  with the topology generated by the cylinder sets is metrizable as we can define the metric  $d(\pi, \pi') = 2^{-\ell}$  where  $\ell$  is the length of the longest common prefix of  $\pi$  and  $\pi'$ .

### 3 Demonic variance and non-determinism score

In this section, we formally define the demonic variance. After proving first auxiliary results, we prove an analogue of Chebyshev’s Inequality using the demonic variance. Then, we introduce the non-determinism score and investigate necessary and sufficient conditions for this score to be 0 or 1. Proofs omitted here can be found in the full version [26].

Throughout this section, let  $\mathcal{M} = (S, Act, P, s_{init})$  be an MDP and let  $X$  be a random variable, i.e., a Borel measurable function on the infinite paths of  $\mathcal{M}$ . We will work under two assumptions that ensure that all notions are well-defined: First, note that  $\mathbb{V}_{\mathcal{M}}^{\max}(X) = 0$  implies that there is a constant  $c$  such that under all schedulers  $\mathfrak{S}$ , we have  $\Pr_{\mathcal{M}}^{\mathfrak{S}}(X = c) = 1$  – an uninteresting case. Furthermore, for meaningful definitions of demonic variance and non-determinism score, we need that the expected value and the variance of  $X$  in  $\mathcal{M}$  are finite. Hence, we work under the following assumption:

► **Assumption 3.1.** We assume that  $0 < \mathbb{V}_{\mathcal{M}}^{\max}(X) < \infty$  and that  $\sup_{\mathfrak{S}} |\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(X)| < \infty$ .

#### 3.1 Demonic variance

As described in the introduction, the idea behind the demonic variance is to quantify the expected squared deviation of  $X$  in two independent executions of  $\mathcal{M}$ , in which the non-determinism is resolved independently as well. We use the following notation: Given a path in  $\mathcal{M} \otimes \mathcal{M}$  consisting of a sequence of pairs of states and pairs of actions, we denote by  $X_1$  and  $X_2$  the function  $X$  applied to the projection of the path on the first component and on the second component, respectively. Given two schedulers  $\mathfrak{S}_1$  and  $\mathfrak{S}_2$  for  $\mathcal{M}$ , we define

$$\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_1, \mathfrak{S}_2}(X) \stackrel{\text{def}}{=} \frac{1}{2} \mathbb{E}_{\mathcal{M} \otimes \mathcal{M}}^{\mathfrak{S}_1 \otimes \mathfrak{S}_2} ((X_1 - X_2)^2).$$

Intuitively, in this definition two independent executions of  $\mathcal{M}$  are run in parallel while the non-determinism is resolved by  $\mathfrak{S}_1$  in the first execution and by  $\mathfrak{S}_2$  in the second component. As the two components in the products  $\mathcal{M} \otimes \mathcal{M}$  and  $\mathfrak{S}_1 \otimes \mathfrak{S}_2$  are independent, the resulting distributions of  $X$  in the two components, i.e.,  $X_1$  and  $X_2$  are independent as well. The factor  $\frac{1}{2}$  is included as for a random variable  $Y$ , this factor also appears in the representation  $\mathbb{V}(Y) = \frac{1}{2}\mathbb{E}((Y_1 - Y_2)^2)$  for two independent copies  $Y_1$  and  $Y_2$  of  $Y$ .

The *demonic variance* is now the worst-case value when ranging over all pairs of schedulers:

$$\mathbb{V}_{\mathcal{M}}^{dem}(X) \stackrel{\text{def}}{=} \sup_{\mathfrak{S}_1, \mathfrak{S}_2} \frac{1}{2} \mathbb{E}_{\mathcal{M} \otimes \mathcal{M}}^{\mathfrak{S}_1 \otimes \mathfrak{S}_2} ((X_1 - X_2)^2).$$

A first simple, but useful, result allows us to express  $\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_1, \mathfrak{S}_2}(X)$  in terms of the expected values and variances of  $X$  under  $\mathfrak{S}_1$  and  $\mathfrak{S}_2$ .

► **Lemma 3.1.** *Given two schedulers  $\mathfrak{S}_1$  and  $\mathfrak{S}_2$  for  $\mathcal{M}$ , we have*

$$\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_1, \mathfrak{S}_2}(X) = \frac{1}{2} \left( \mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_1}(X) + \mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_2}(X) + (\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_1}(X) - \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_2}(X))^2 \right).$$

This lemma allows us to provide an insightful graphical interpretation of the demonic variance using the standard deviation  $\mathbb{SD}(X) \stackrel{\text{def}}{=} \sqrt{\mathbb{V}(X)}$  of a random variable  $X$ :

► **Example 3.2.** Suppose in an MDP  $\mathcal{M}$ , there are four deterministic scheduler  $\mathfrak{S}_1, \dots, \mathfrak{S}_4$  with expected values 1, 2, 3, and 4 and variances 1, 8, 8, and 5 for a random variable  $X$ . Lemma 2.2 allows us to compute the variances of schedulers obtained by randomization leading to parabolic line segments in the expectation-variance-plane as depicted in Figure 3a (see also [28]). Further randomizations also make it possible to realize any combination of expectation and variance in the interior of the resulting shape. When looking for the maximal variance and the demonic variance, only the upper bound of this shape is relevant.

In Figure 3b, we now depict the standard deviations of schedulers on this upper bound over the expectation twice on two orthogonal planes. Clearly, the highest standard deviation (and consequently variance) is obtained for the expected value 2.5 in this example. The red dotted line of length  $\sqrt{2\mathbb{V}_{\mathcal{M}}^{\max}(X)}$  connects the two points corresponding to this maximum on the two planes. Considering  $\mathfrak{S}_2$  and  $\mathfrak{S}_4$ , we can also find the value  $\sqrt{2\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_2, \mathfrak{S}_4}(X)}$ : The blue dashed line connects the point corresponding to  $\mathfrak{S}_2$  on one of the planes to the point corresponding to  $\mathfrak{S}_4$  on the other plane. By the Pythagorean theorem, its length is

$$\sqrt{\sqrt{\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_2}(X)}^2 + (\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_2}(X) - \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_4}(X))^2 + \sqrt{\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_4}(X)}^2} = \sqrt{2\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_2, \mathfrak{S}_4}(X)}.$$

So, finding  $\sqrt{2}$  times the “demonic standard deviation” and hence the demonic variance corresponds to finding two points on the two orthogonal graphs with maximal distance.

The relation between maximal and demonic variance is shown in the following proposition.

► **Proposition 3.3.** *We have  $\mathbb{V}_{\mathcal{M}}^{\max}(X) \leq \mathbb{V}_{\mathcal{M}}^{dem}(X) \leq 2\mathbb{V}_{\mathcal{M}}^{\max}(X)$ .*

By means of Chebyshev’s Inequality, the variance can be used to bound the probability that a random variable  $Y$  lies far from its expected value. Using the demonic variance, we can prove an analogous result providing bounds on the probability that the outcomes of  $X$  in two independent executions of the MDP  $\mathcal{M}$  lie far apart. This can be seen as a first step in the direction of using the demonic variance to provide guarantees on the behavior of a system.



► **Theorem 3.4.** *We have  $\Pr_{\mathcal{M} \otimes \mathcal{M}}^{\mathfrak{S} \otimes \mathfrak{T}} \left( |X_1 - X_2| \geq k \cdot \sqrt{\mathbb{V}_{\mathcal{M}}^{dem}(X)} \right) \leq \frac{2}{k^2}$  for any  $k \in \mathbb{R}_{>0}$  and schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  for  $\mathcal{M}$ .*

Using the result that  $\mathbb{V}_{\mathcal{M}}^{dem}(X) \leq 2\mathbb{V}_{\mathcal{M}}^{\max}(X)$ , we obtain the following variant of the inequality providing a weaker bound in terms of the maximal variance.

► **Corollary 3.5.** *We have  $\Pr_{\mathcal{M} \otimes \mathcal{M}}^{\mathfrak{S} \otimes \mathfrak{T}} (|X_1 - X_2| \geq k \cdot \sqrt{\mathbb{V}_{\mathcal{M}}^{\max}(X)}) \leq \frac{4}{k^2}$  for any  $k \in \mathbb{R}_{>0}$  and schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  for  $\mathcal{M}$ .*

### 3.2 Non-determinism score

We have seen that the demonic variance is larger than the maximal variance by a factor between 1 and 2. As described in the introduction, we use this insight as the basis for a score quantifying how much worse the “uncertainty” of  $X$  is when non-determinism can be resolved differently in two executions of an MDP compared to how bad it can be in a single execution. We define the non-determinism score (NDS)

$$\text{NDS}(\mathcal{M}, X) \stackrel{\text{def}}{=} \frac{\mathbb{V}_{\mathcal{M}}^{dem}(X) - \mathbb{V}_{\mathcal{M}}^{\max}(X)}{\mathbb{V}_{\mathcal{M}}^{\max}(X)}.$$

By Assumption 3.1, the NDS is well-defined. By Proposition 3.3, the NDS always returns a value in  $[0, 1]$ . Clearly, in Markov chains, the NDS is 0. A bit more general, we can show:

► **Proposition 3.6.** *If  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(X) = \mathbb{E}_{\mathcal{M}}^{\mathfrak{T}}(X)$  for all schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$ , then  $\text{NDS}(\mathcal{M}, X) = 0$ .*

In transition systems viewed as MDPs in which all transition probabilities are 0 or 1, the NDS is 1: Under Assumption 3.1 in a transition system the value of  $X$  must be bounded, i.e.,  $X \in [a, b]$  for some  $a, b \in \mathbb{R}$  such that  $\sup_{\pi} X(\pi) = b$  and  $\inf_{\pi} X(\pi) = a$  where  $\pi$  ranges over all paths. Any path can be realized by a scheduler with probability 1. So, for any  $\varepsilon > 0$ , there are schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  with  $\Pr_{\mathcal{M}}^{\mathfrak{S}}(X < a + \varepsilon) = 1$  and  $\Pr_{\mathcal{M}}^{\mathfrak{T}}(X > b - \varepsilon) = 1$ . Then,  $\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}, \mathfrak{T}}(X) \geq \frac{1}{2}(b - a - 2\varepsilon)^2$ . For  $\varepsilon \rightarrow 0$ , this converges to  $\frac{(a-b)^2}{2}$ . It is well-known that the variance of random variables taking values in  $[a, b]$  is maximal for the random variable taking values  $a$  and  $b$  with probability  $\frac{1}{2}$  each. The variance in this case is  $\frac{(a-b)^2}{4}$ . So, the maximal variance is (at most) half the demonic variance in this case. Consequently, the NDS is 1.

Of course, a NDS of 1 does not imply that there are no probabilistic transitions in  $\mathcal{M}$ . Nevertheless, a NDS of 1 has severe implications showing that the outcome of  $X$  can be heavily influenced by the non-determinism in this case as the following theorem shows:

► **Theorem 3.7.** *If  $\text{NDS}(\mathcal{M}, X) = 1$ , the following statements hold:*

1. *For every  $\varepsilon > 0$ , there are schedulers  $\mathfrak{Min}_{\varepsilon}$  and  $\mathfrak{Max}_{\varepsilon}$  with  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{Min}_{\varepsilon}}(X) \leq \mathbb{E}_{\mathcal{M}}^{\min}(X) + \varepsilon$  and  $\mathbb{V}_{\mathcal{M}}^{\mathfrak{Min}_{\varepsilon}}(X) \leq \varepsilon$ , and  $\mathbb{E}_{\mathcal{M}}^{\mathfrak{Max}_{\varepsilon}}(X) \geq \mathbb{E}_{\mathcal{M}}^{\max}(X) - \varepsilon$  and  $\mathbb{V}_{\mathcal{M}}^{\mathfrak{Max}_{\varepsilon}}(X) \leq \varepsilon$ .*
2. *If there are schedulers  $\mathfrak{S}_0$  and  $\mathfrak{S}_1$ , with  $\mathbb{V}_{\mathcal{M}}^{dem}(X) = \mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_0, \mathfrak{S}_1}(X)$ , then, for  $i = 0$  or  $i = 1$ ,  $\Pr_{\mathcal{M}}^{\mathfrak{S}_i}(X = \mathbb{E}_{\mathcal{M}}^{\min}(X)) = 1$  and  $\Pr_{\mathcal{M}}^{\mathfrak{S}_{1-i}}(X = \mathbb{E}_{\mathcal{M}}^{\max}(X)) = 1$ .*
3. *If  $X$  is bounded and continuous wrt the topology generated by the cylinder sets, there are schedulers  $\mathfrak{Min}$  and  $\mathfrak{Max}$  with  $\Pr_{\mathcal{M}}^{\mathfrak{Min}}(X = \mathbb{E}_{\mathcal{M}}^{\min}(X)) = 1$  and  $\Pr_{\mathcal{M}}^{\mathfrak{Max}}(X = \mathbb{E}_{\mathcal{M}}^{\max}(X)) = 1$ .*

The first two statements can be shown by elementary calculations. For the third statement, we use topological arguments. We view schedulers as elements of  $\prod_{k=0}^{\infty} \text{Distr}(\text{Act})^{\text{Paths}_{\mathcal{M}}^k}$  where  $\text{Paths}_{\mathcal{M}}^k$  is the set of paths of length  $k$  in  $\mathcal{M}$  and prove the following result:

► **Proposition 3.8.** *The space of schedulers  $\text{Sched}(\mathcal{M}) = \prod_{k=0}^{\infty} \text{Distr}(\text{Act})^{\text{Paths}_{\mathcal{M}}^k}$  with the product topology is compact. So, every sequence of schedulers has a converging subsequence in this space. Further, for a sequence  $(\mathfrak{S}_j)_{j \in \mathbb{N}}$  converging to a scheduler  $\mathfrak{S}$  in this space, the sequence of probability measures  $(\Pr_{\mathcal{M}}^{\mathfrak{S}_j})_{j \in \mathbb{N}}$  weakly converges to the probability measure  $\Pr_{\mathcal{M}}^{\mathfrak{S}}$ .*

An example for a random variable that is bounded and continuous wrt the topology generated by the cylinder sets is the discounted reward: Given a reward function  $rew: S \rightarrow \mathbb{R}$ , the discounted reward of a path  $\pi = s_0 a_0 s_1 \dots$  is defined as  $DR_\lambda(\pi) \stackrel{\text{def}}{=} \sum_{j=0}^{\infty} \lambda^j rew(s_j)$  for some discount factor  $\lambda \in (0, 1)$ . First,  $|DR_\lambda|$  is bounded by  $\max_{s \in S} |rew(s)| \cdot \frac{1}{1-\lambda}$ . Further, for any  $\varepsilon > 0$ , let  $N$  be a natural number such that  $\max_{s \in S} |rew(s)| \cdot \frac{\lambda^N}{1-\lambda} < \varepsilon$ . Then,  $|DR_\lambda(\pi) - DR_\lambda(\rho)| < \varepsilon$  for all paths  $\pi$  and  $\rho$  that share a prefix of length more than  $N$ .

#### 4 Weighted reachability

We now address the problems to compute the demonic and the maximal variance for weighted reachability where a weight is collected on a run depending on which absorbing state is reached. As the NDS is defined via these two quantities, we do not address it separately here. Throughout this section, let  $\mathcal{M} = (S, Act, P, s_{init})$  be an MDP with set of absorbing states  $T \subseteq S$  and let  $wgt: T \rightarrow \mathbb{Q}$  be a weight function. We define the random variable WR on infinite paths  $\pi$  by  $WR(\pi) = wgt(t)$  if  $\pi$  reaches the absorbing state  $t \in T$ , and  $WR(\pi) = 0$  if  $\pi$  does not reach  $T$ . The main result we are going to establish is the following:

**Main result.** *The maximal variance  $\mathbb{V}_{\mathcal{M}}^{\max}(\text{WR})$  and an optimal memoryless randomized scheduler can be computed in polynomial time.*

*The demonic variance  $\mathbb{V}_{\mathcal{M}}^{\text{dem}}(\text{WR})$  can be computed as the solution to a bilinear program that can be constructed in polynomial time. Furthermore, there is a pair of memoryless deterministic schedulers realizing the demonic variance.*

The following standard model transformation collapsing end components (see [14]) allows us to assume that  $T$  is reached almost surely under any scheduler: We add a new absorbing state  $t^*$  and set  $wgt(t^*) = 0$  and collapse all maximal end components  $\mathcal{E}$  in  $S \setminus T$  to single states  $s_{\mathcal{E}}$ . In  $s_{\mathcal{E}}$ , all actions that were enabled in some state in  $\mathcal{E}$  and that did not belong to  $\mathcal{E}$  as well as a new action  $\tau$  leading to  $t^*$  with probability 1 are enabled. In the resulting MDP  $\mathcal{N}$ , the set of absorbing states  $T \cup \{t^*\}$  is reached almost surely under any scheduler. Further, for any scheduler  $\mathfrak{S}$  for  $\mathcal{M}$ , there is a scheduler  $\mathfrak{T}$  for  $\mathcal{N}$  such that the distribution of WR is the same under  $\mathfrak{S}$  in  $\mathcal{M}$  and under  $\mathfrak{T}$  in  $\mathcal{N}$ , and vice versa. So, w.l.o.g., assume the following:

► **Assumption 4.1.** *The set  $T$  is reached almost surely under any scheduler  $\mathfrak{S}$  for  $\mathcal{M}$ .*

In the sequel, we first address the computation of the maximal variance and afterwards of the demonic variance of WR in  $\mathcal{M}$ . Omitted proofs can be found in the full version [26].

**Computation of the maximal variance.** It is well-known that the set of vectors  $(\text{Pr}_{\mathcal{M}}^{\mathfrak{S}}(\diamond q))_{q \in T}$  of combinations of reachability probabilities for states in  $T$  that can be realized by a scheduler  $\mathfrak{S}$  can be described by a system of linear inequalities (see, e.g., [18]). We provide such a system of inequalities below in equations (1) – (3). The equations use variables  $x_{s,\alpha}$  for all state-action pairs  $(s, \alpha)$  encoding the expected number of times action  $\alpha$  is taken in state  $s$ . Setting  $\mathbf{1}_{s=s_{init}} = 1$  if  $s = s_{init}$  and  $\mathbf{1}_{s=s_{init}} = 0$  otherwise, we require

$$x_{s,\alpha} \geq 0 \quad \text{for all } (s, \alpha), \quad (1)$$

$$\sum_{\alpha \in Act(s)} x_{s,\alpha} = \sum_{t \in S, \beta \in Act(t)} x_{t,\beta} \cdot P(t, \beta, s) + \mathbf{1}_{s=s_{init}} \quad \text{for all } s \in S \setminus T, \quad (2)$$

$$y_q = \sum_{t \in S, \beta \in Act(t)} x_{t,\beta} \cdot P(t, \beta, q) \quad \text{for all } q \in T. \quad (3)$$

The variables  $y_q$  for  $q \in T$  represent the probabilities that state  $q$  is reached. We can now express the expected value of WR and  $\text{WR}^2$  via variables  $e_1$  and  $e_2$  via the constraints:

$$e_1 = \sum_{q \in T} y_q \cdot \text{wgt}(q) \quad \text{and} \quad e_2 = \sum_{q \in T} y_q \cdot \text{wgt}(q)^2. \quad (4)$$

The variance can now be written as a quadratic objective function in  $e_1$  and  $e_2$ :

$$\text{maximize} \quad e_2 - e_1^2. \quad (5)$$

► **Theorem 4.1.** *The maximal value in objective (5) under constraints (1) – (4) is  $\mathbb{V}_{\mathcal{M}}^{\max}(\text{WR})$ .*

Due to the concavity of the objective function, we conclude:

► **Corollary 4.2.** *The maximal variance  $\mathbb{V}_{\mathcal{M}}^{\max}(\text{WR})$  can be computed in polynomial time. Furthermore, there is a memoryless randomized scheduler  $\mathfrak{S}$  with  $\mathbb{V}_{\mathcal{M}}^{\mathfrak{S}}(\text{WR}) = \mathbb{V}_{\mathcal{M}}^{\max}(\text{WR})$ , which can also be computed in polynomial time.*

**Computation of the demonic variance.** The demonic variance can also be expressed as the solution to a quadratic program. To encode the reachability probabilities for states in  $T$  under two distinct schedulers, we use variables  $x_{s,\alpha}$  for all state weight pairs  $(s, \alpha)$  and  $y_q$  for  $q \in T$  subject to constraints (1) – (3) as before. Additionally, we use variables  $x'_{s,\alpha}$  for all state weight pairs  $(s, \alpha)$  and  $y'_q$  for  $q \in T$  subject to the analogue constraints (1') – (3') using these primed variables. The maximization of the demonic variance can be expressed as

$$\text{maximize} \quad \frac{1}{2} \sum_{q,r \in T} y_q \cdot y'_r \cdot (\text{wgt}(q) - \text{wgt}(r))^2. \quad (6)$$

► **Theorem 4.3.** *The maximum in (6) under constraints (1) – (3), (1') – (3') is  $\mathbb{V}_{\mathcal{M}}^{\text{dem}}(\text{WR})$ .*

The quadratic objective function (6) is not concave. However, it is *bilinear* and *separable*. This means that the variables can be split into two sets, the primed and the unprimed variables, such that the quadratic terms only contain products of variables from different sets and each constraint contains only variables from the same set. In general, checking whether the solution to a separable bilinear program exceeds a given threshold is NP-hard [23]. Nevertheless, solution methods tailored for bilinear programs that perform well in practice have been developed (see, e.g., [19]). Further, bilinearity allows us to conclude:

► **Corollary 4.4.** *There is a pair of memoryless deterministic schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  for  $\mathcal{M}$  such that  $\mathbb{V}_{\mathcal{M}}^{\text{dem}}(\text{WR}) = \mathbb{V}_{\mathcal{M}}^{\mathfrak{S},\mathfrak{T}}(\text{WR})$ .*

For the complexity of the threshold problem, we can conclude an NP upper bound. Whether the computation of the demonic variance is possible in polynomial time is left open.

► **Corollary 4.5.** *Given  $\mathcal{M}$ ,  $\text{wgt}$  and  $\vartheta \in \mathbb{Q}$ , deciding whether  $\mathbb{V}_{\mathcal{M}}^{\text{dem}}(\text{WR}) \geq \vartheta$  is in NP.*

## 5 Accumulated rewards

One of the most important random variables studied on MDPs are accumulated rewards: Let  $\mathcal{M} = (S, \text{Act}, P, s_{\text{init}})$  be an MDP and let  $\text{rew}: S \rightarrow \mathbb{N}$  be a reward function. We extend the reward function to paths  $\pi = s_0 \alpha_0 s_1 \dots$  by  $\text{rew}(\pi) = \sum_{i=0}^{\infty} \text{rew}(s_i)$ . For this random variable, we prove the following result:

**Main result.** *The maximal variance  $\mathbb{V}_{\mathcal{M}}^{\max}(rew)$  and an optimal randomized finite-memory scheduler can be computed in exponential time.*

*The demonic variance  $\mathbb{V}_{\mathcal{M}}^{\text{dem}}(rew)$  can be computed as the solution to a bilinear program that can be constructed in exponential time. Furthermore, there is a pair of deterministic finite-memory schedulers realizing the demonic variance.*

We provide a sketch outlining the proof strategy. For a detailed exposition, see [26].

**Proof sketch for the main result.** It can be checked in polynomial time whether  $\mathbb{E}_{\mathcal{M}}^{\max}(rew) < \infty$  [14]. If this is the case, this allows us to perform the same preprocessing as in Section 4 that removes all end components without changing the possible distributions of  $rew$  [14].

**Bounding expected values and expectation maximizing actions.** After the pre-processing, a terminal state is reached almost surely. As shown in [28], this allows to obtain a bound  $Q$  on  $\mathbb{E}_{\mathcal{M}}^{\max}(rew^2)$  in polynomial time. Further, the maximal expectation  $\mathbb{E}_{\mathcal{M},s}^{\max}(rew)$  from each state  $s$  can be computed in polynomial time [7, 14]. From these values, a set of *maximizing actions*  $Act^{\max}(s)$  for each state  $s$  can be computed. After the preprocessing, a scheduler is expectation optimal iff it only chooses actions from these sets. If a scheduler  $\mathfrak{S}$  initially chooses a non-maximizing action in a state  $s$ , the expected value  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(rew)$  is strictly smaller than  $\mathbb{E}_{\mathcal{M},s}^{\max}(rew)$ . We define  $\delta$  to be the minimal difference between these values ranging over all starting states and non-maximizing actions. So,  $\delta$  is the “minimal loss” in expected value of  $rew$  received by choosing a non-maximizing action.

**Switching to expectation maximization.** Using the values  $Q$  and  $\delta$ , we provide a bound  $B$  such that any scheduler choosing a non-maximizing action with positive probability after a path  $\pi$  with  $rew(\pi) \geq B$  cannot realize the maximal variance. Intuitively, the reason is that the influence of accumulating future rewards on the variance grows with the amount of rewards already accumulated due to the quadratic nature of variance. The bound  $B$  can be computed in polynomial time and its numerical value is exponential in the size of the input.

It follows that variance maximizing schedulers have to maximize the future expected rewards after a reward of at least  $B$  has been accumulated. Furthermore, we can show that among all expectation maximizing schedulers, a variance maximizing scheduler has to be used above the reward bound  $B$ . In [28], it is shown that a memoryless deterministic expectation maximizing scheduler  $\mathfrak{U}$  that maximizes the variance among all expectation maximizing schedulers can be computed in polynomial time. So, schedulers maximizing the variance of  $rew$  can be chosen to behave like  $\mathfrak{U}$  once a reward of at least  $B$  has been accumulated.

**Quadratic program.** Now, we can unfold the MDP  $\mathcal{M}$  by storing in the state space how much reward has been accumulated up to the bound  $B$ . This results on an exponentially larger MDP  $\mathcal{M}'$ . Using the expected values  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{U}}(rew)$  and the variances  $\mathbb{V}_{\mathcal{M},s}^{\mathfrak{U}}(rew)$  under  $\mathfrak{U}$  from each state  $s$ , we can formulate a quadratic program similar to the one for weighted reachability in Section 4 for this unfolded MDP  $\mathcal{M}'$ . From the solution to this quadratic program, the maximal variance and an optimal memoryless scheduler  $\mathfrak{S}$  for  $\mathcal{M}'$  can be extracted. Transferred back to  $\mathcal{M}$ , the scheduler  $\mathfrak{S}$  corresponds to a reward-based finite-memory scheduler that keeps track of the accumulated reward up to bound  $B$ . As the quadratic program is convex, these computations can be carried out in exponential time.

**Demonic variance.** For the demonic variance, the overall proof follows the same steps. Similar to the bound  $B$  above, a bound  $B'$  can be provided such that in any pair of scheduler  $\mathfrak{S}$  and  $\mathfrak{T}$  realizing the demonic variance, both schedulers can be assumed to switch to the behavior of the memoryless deterministic scheduler  $\mathfrak{U}$  above the reward bound  $B'$ .

Again by unfolding the state space up to this reward bound, the demonic variance can be computed via a bilinear program of exponential size similar to the one used in Section 4 for weighted reachability. Furthermore, the pair of optimal memoryless deterministic schedulers in the unfolded MDP, which can be extracted from the solution, corresponds to a pair of deterministic reward-based finite-memory schedulers in the original MDP  $\mathcal{M}$ . ◀

## 6 Conclusion

We introduced the notion of demonic variance that quantifies the uncertainty under probabilism *and* non-determinism of a random variable  $X$  in an MDP  $\mathcal{M}$ . As this demonic variance is at most twice as big as the maximal variance of  $X$ , we used it to define the NDS for MDPs.

The demonic variance can be used to provide new types of guarantees on the behavior of systems. A first step in this direction is the variant of Chebyshev’s Inequality using the demonic variance proved in this paper. Furthermore, the demonic variance and the NDS can serve as the basis for notions of responsibility. On the one hand, such notions could ascribe responsibility for the uncertainty to non-determinism and probabilism. On the other hand, comparing the NDS from different starting states can be used to identify regions of the state space in which the non-deterministic choices are of high importance.

For weighted reachability and accumulated rewards, we proved that randomized finite-memory schedulers are sufficient to maximize the variance. For the demonic variance, even pairs of deterministic finite-memory schedulers are sufficient. While we obtained upper bounds via the formulation of the computation problems as quadratic programs, determining the precise complexities is left as future work. In the case of accumulated rewards, we restricted to non-negative rewards. When dropping this restriction, severe difficulties have to be expected as several related problems on MDPs exhibit inherent number-theoretic difficulties rendering the decidability status of the corresponding decision problems open [27].

Of course the investigation of the demonic variance and NDS for further random variables constitutes an interesting direction for future work. For practical purposes, studying also the approximability of the maximal and demonic variance is important.

Finally, in the spirit of the demonic variance, further notions can be defined to quantify the uncertainty in  $X$  if the non-determinism in two executions of  $\mathcal{M}$  is not resolved independently, but information can be passed between the two executions. This could be useful, e.g., to analyze the potential power of coordinated attacks on a network. Formally, such a notion could be defined as  $\sup_{\mathfrak{S}} \mathbb{E}_{\mathcal{M} \otimes \mathcal{M}}^{\mathfrak{S}}((X_1 - X_2)^2)$  where  $\mathfrak{S}$  ranges over all schedulers for  $\mathcal{M} \otimes \mathcal{M}$ . In this context, also using an asynchronous product of  $\mathcal{M}$  with  $\mathcal{M}$  could be reasonable.

---

## References

- 1 Erika Ábrahám and Borzoo Bonakdarpour. Hyperpctl: A temporal logic for probabilistic hyperproperties. In *International Conference on Quantitative Evaluation of Systems*, pages 20–35. Springer, 2018. doi:10.1007/978-3-319-99154-2\_2.
- 2 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.
- 3 Christel Baier, Marcus Daum, Clemens Dubslaff, 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.

- 4 Christel Baier, Clemens Dubsclaff, Florian Funke, Simon Jantsch, Rupak Majumdar, Jakob Piribauer, and Robin Ziemek. From verification to causality-based explications (invited talk). In Nikhil Bansal, Emanuela Merelli, and James Worrell, editors, *48th International Colloquium on Automata, Languages, and Programming, ICALP 2021, July 12-16, 2021, Glasgow, Scotland (Virtual Conference)*, volume 198 of *LIPICs*, pages 1:1–1:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICs.ICALP.2021.1.
- 5 Christel Baier, Florian Funke, and Rupak Majumdar. A game-theoretic account of responsibility allocation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pages 1773–1779. ijcai.org, 2021. doi:10.24963/IJCAI.2021/244.
- 6 Christel Baier, Florian Funke, and Rupak Majumdar. Responsibility attribution in parameterized markovian models. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11734–11743. AAAI Press, 2021. doi:10.1609/aaai.v35i13.17395.
- 7 Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991. doi:10.1287/moor.16.3.580.
- 8 Tomáš Brázdil, Václav Brožek, Krishnendu Chatterjee, Vojtěch Forejt, and Antonín Kučera. Markov decision processes with multiple long-run average objectives. *Logical Methods in Computer Science*, 10, 2014. doi:10.2168/LMCS-10(1:13)2014.
- 9 Tomáš Brázdil, Krishnendu Chatterjee, Vojtěch Forejt, and Antonín Kučera. Trading performance for stability in Markov decision processes. *Journal of Computer and System Sciences*, 84:144–170, 2017. doi:10.1016/j.jcss.2016.09.009.
- 10 Hana Chockler and Joseph Y. Halpern. Responsibility and Blame: A Structural-Model Approach. *J. Artif. Int. Res.*, 22(1):93–115, October 2004. doi:10.1613/jair.1391.
- 11 Michael R Clarkson, Bernd Finkbeiner, Masoud Koleini, Kristopher K Micinski, Markus N Rabe, and César Sánchez. Temporal logics for hyperproperties. In *Principles of Security and Trust: Third International Conference, POST*, pages 265–284. Springer, 2014. doi:10.1007/978-3-642-54792-8\_15.
- 12 Michael R Clarkson and Fred B Schneider. Hyperproperties. *Journal of Computer Security*, 18(6):1157–1210, 2010. doi:10.3233/JCS-2009-0393.
- 13 EJ Collins. Finite-horizon variance penalised Markov decision processes. *Operations-Research-Spektrum*, 19(1):35–39, 1997.
- 14 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.
- 15 Rayna Dimitrova, Bernd Finkbeiner, and Hazem Torfah. Probabilistic hyperproperties of Markov decision processes. In *International Symposium on Automated Technology for Verification and Analysis, ATVA*, pages 484–500. Springer, 2020. doi:10.1007/978-3-030-59152-6\_27.
- 16 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.
- 17 Christoph Haase and Stefan Kiefer. The odds of staying on budget. In *42nd International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 9135 of *Lecture Notes in Computer Science*, pages 234–246. Springer, 2015. doi:10.1007/978-3-662-47666-6\_19.
- 18 Lodewijk Kallenberg. *Markov Decision Processes*. Lecture Notes. University of Leiden, 2016.
- 19 Scott Kolodziej, Pedro M Castro, and Ignacio E Grossmann. Global optimization of bilinear programs with a multiparametric disaggregation technique. *Journal of Global Optimization*, 57:1039–1063, 2013. doi:10.1007/s10898-012-0022-1.

- 20 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.
- 21 Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Inf. Fusion*, 85(C):1–22, September 2022. doi:10.1016/j.inffus.2022.03.003.
- 22 Petr Mandl. On the variance in controlled Markov chains. *Kybernetika*, 7(1):1–12, 1971. URL: <http://www.kybernetika.cz/content/1971/1/1>.
- 23 Olvi L Mangasarian. The linear complementarity problem as a separable bilinear program. *Journal of Global Optimization*, 6(2):153–161, 1995. doi:10.1007/BF01096765.
- 24 Shie Mannor and John N. Tsitsiklis. Mean-variance optimization in Markov decision processes. In *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).
- 25 Corto Mascle, Christel Baier, Florian Funke, Simon Jantsch, and Stefan Kiefer. Responsibility and verification: Importance value in temporal logics. In *36th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS*, pages 1–14. IEEE, 2021. doi:10.1109/LICS52264.2021.9470597.
- 26 Jakob Piribauer. Demonic variance and a non-determinism score for Markov decision processes, 2024. doi:10.48550/arXiv.2406.18727.
- 27 Jakob Piribauer and Christel Baier. On Skolem-hardness and saturation points in Markov decision processes. In Artur Czumaj, Anuj Dawar, and Emanuela Merelli, editors, *47th International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 168 of *LIPICs*, pages 138:1–138:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICs.ICALP.2020.138.
- 28 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.
- 29 Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994. doi:10.1002/9780470316887.
- 30 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.
- 31 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.
- 32 Tom Verhoeff. Reward variance in Markov chains: A calculational approach. In *Proceedings of Eindhoven FASTAR Days*. Citeseer, 2004.
- 33 Vahid Yazdanpanah, Mehdi Dastani, Wojciech Jamroga, Natasha Alechina, and Brian Logan. Strategic Responsibility Under Imperfect Information. In *Proc. of the 18th Intern. Conf. on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 592–600. AAMAS Foundation, 2019. URL: <http://dl.acm.org/citation.cfm?id=3331745>.