The variance-penalized stochastic shortest path problem

Jakob Piribauer, Ocan Sankur, Christel Baier

The variance-penalized stochastic shortest path problem

Jakob Piribauer
Technische Universität Dresden, Germany

Ocan Sankur
Univ Rennes, Inria, CNRS, IRISA, France

Christel Baier
Technische Universität Dresden, Germany

Abstract

The stochastic shortest path problem (SSPP) asks to resolve the non-deterministic choices in a Markov decision process (MDP) such that the expected accumulated weight before reaching a target state is maximized. This paper addresses the optimization of the variance-penalized expectation (VPE) of the accumulated weight, which is a variant of the SSPP in which a multiple of the variance of accumulated weights is incurred as a penalty. It is shown that the optimal VPE in MDPs with non-negative weights as well as an optimal deterministic finite-memory scheduler can be computed in exponential space. The threshold problem whether the maximal VPE exceeds a given rational is shown to be EXPTIME-hard and to lie in NEXPTIME. Furthermore, a result of interest in its own right obtained on the way is that a variance-minimal scheduler among all expectation-optimal schedulers can be computed in polynomial time.

2012 ACM Subject Classification Theory of computation → Verification by model checking

Keywords and phrases Markov decision process, variance, stochastic shortest path problem

1 Introduction

Markov decision processes (MDPs) are a standard operational model comprising randomization and non-determinism and are widely used in verification, articifical intelligence, robotics, and operations research. In each state of an MDP, there is a non-deterministic choice from a set of actions. Each action is equipped with a weight and a probability distribution according to which the successor state is chosen randomly. In the analysis of systems modelled as MDPs, one typically is interested in the worst- or best-case behavior, where worst and best case range over all resolutions of the non-determinism. So, the resulting algorithmic problems on MDPs usually ask to resolve non-deterministic choices by specifying a *scheduler* such that the resulting probabilistic behavior is optimized with respect to an objective function. If the weights are used to model one of various quantitative aspects of a system such as costs, resource consumption, rewards, or utility, a frequently encountered such optimization problem is the *stochastic shortest path problem* (SSPP) [**?**, **?**]. It asks to optimize the expected value of the accumulated weight before reaching a target state. Example applications include the analysis of worst-case expected termination times of probabilistic programs or finding the optimal controls in a motion planning scenario with random external influences.

While a solution to the SSPP provides guarantees on the behavior of a system in all environments or indicates the optimal control to maximize expected rewards, it completely disregards all other aspects of the resulting probability distribution of the accumulated weight besides the expected value. In almost all practical applications, however, the uncertainty coming with the probabilistic behavior cannot be neglected. In traffic control systems or energy grids, for example, large variability in the throughput comes at a high cost due to the risk of traffic jams or the difficulty of storing surplus energy. Also a probabilistic program

employed in a complex environment might be of more use with a higher expected termination time in exchange for a lower chance of extreme termination times.

To overcome these shortcomings of the SSPP, various additional optimization problems have been studied in the literature: Optimizing conditional expected accumulated weights under the condition that certain system states are reached allows for a more fine-grained system analysis by making it possible to determine the worst- or best-case expectation in different scenarios [**?**, **?**]. Given a probability p, quantiles on the accumulated weight in MDPs, also called *values-at-risk* in the context of risk analysis, are the best bound B such that the accumulated weight exceeds B with probability at most p in the worst or best case [**?**, **?**]. The *conditional value-at-risk* and the *entropic value-at-risk* are more involved measures that have been studied in this context [**?**, **?**]. They quantify how far the probability mass of the tail of the probability distribution lies above the value-at-risk. The arguably most prominent measure for the deviation of a random variable from its expected value is the *variance*. The computation of the variance of accumulated weights has been studied in Markov chains [**?**] and in MDPs [**?**, **?**]. The investigations of variance in MDPs in the literature is discussed in more detail in the 'Related Work' section below.

Variance-penalized expectation (VPE). In this paper, we investigate a variant of the SSPP in which the costs caused by probabilistic uncertainty are priced in to the objective function: We study the optimization of the *variance-penalized expectation* (VPE), a well-known measure that combines the expected value μ and the variance σ^2 into the single objective function $\mu - \lambda \cdot \sigma^2$ where λ is a parameter that can be varied to aim for different tradeoffs between expectation and variance. In the context of optimization problems on MDPs, the VPE has been studied, e.g., in [**?**, **?**].

Furthermore, the VPE finds use in an area of research primarily concerned with the tradeoffs between expected performance and risks, namely, the theory of financial markets and investment decision-making: In 1952, Harry Markowitz introduced *modern portfolio theory* that evaluates portfolios in terms of expected returns and variance of the returns [**?**], for which he was later awarded the Nobel Prize in economics. A portfolio lies on the *Markowitz efficient frontier* if the expected return cannot be increased without increasing the variance and, vice versa, the variance cannot be decreased without decreasing the expectation. The final choice of a portfolio on the efficient frontier depends on the investors preferences. In this context, the VPE $\mu - \lambda \cdot \sigma^2$ is a simple, frequently used way to express the preference of an investor using the single parameter λ capturing the risk-aversion of the investor (see, e.g., [**?**]). In more involved accounts, the investor's preference is described in terms of a utility function mapping returns to utilities. For the commonly used exponential utility function $u(x) = -e^{-\alpha x}$ and normally distributed returns, the objective of an investor trying to maximize expected utility turns out to be equivalent to the maximization of the VPE with parameter $\lambda = \alpha/2$ [**?**, **?**].

For an illustration of the VPE, consider the following example:

▶ **Example 1.** Consider the MDP $\mathcal{M}$ depicted in Figure 1 where non-trivial probability values as well as the weights accumulated are denoted next to the transitions. We want to analyze the possible trade-offs between the variance and the expected value of the accumulated weight that we can achieve in this MDP.

The only non-deterministic choice is in the state s_{init}. Choosing action α leads to *goal* with expected weight and variance 0. For the remaining actions, the accumulated weight follows a geometric distribution where in each step some weight k is accumulated and *goal* is reached with some probability p after the step. For such a distribution, it is well-known that the expected accumulated weight is k/p and the variance is $(k/p)^2 \cdot (1-p)$. Plugging in the

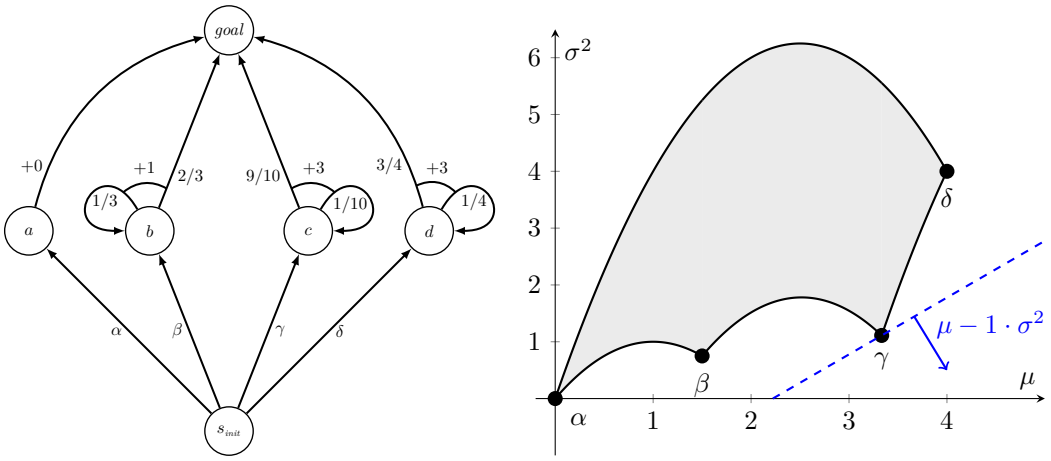

Figure 1 The left hand side shows the MDP $\mathcal{M}$ for Example 1. On the right hand side, all possible combinations of expected accumulated weight and variance for schedulers for $\mathcal{M}$ are depicted. The points corresponding to the four deterministic schedulers are marked by the corresponding action. Furthermore, the blue line indicates all points at which $\mu - 1 \cdot \sigma^2 = 20/9$ and the arrow indicates the direction in which the value of this objective function increases.

respective values for the distributions reached after actions β, γ, and δ, we obtain the pairs of expectations and variances as depicted on the right-hand side of Figure 1. In particular, choosing γ leads to an expectation of $10/3$ and a variance of $10/9$.

Making use of randomization over two different actions τ and σ with probability p and $1-p$, respectively, for some $p \in (0,1)$, we will see in Remark 12 in Section 4 that the expected values and variances under the resulting schedulers lie on a parabolic line segment depicted in black that is uniquely determined by the expected values and variances under τ and σ. By further randomization over multiple actions, combinations of expectation and variance in the gray region in Figure 1 can be realized.

Consider now the VPE with parameter $\lambda = 1$. The dashed blue line in Figure 1 marks all points at which $\mu - 1 \cdot \sigma^2 = 20/9$. The arrow indicates in which direction the value of the VPE increases. So, it turns out that choosing action γ maximizes the VPE in this case; the slightly lower expectation compared to δ is compensated by a significantly lower variance. Geometrically, we can observe that the optimal point for the VPE for any parameter will always lie on the border of the convex hull of the region of feasible points in the μ-σ^2-plane as the VPE is a linear function of expectation and variance. For varying values of λ, also α (for $\lambda \geq 3$) and δ (for $\lambda \leq 1/13$) can constitute the optimal choice in s_{init} for the maximization of the VPE, while β is not optimal for any choice of λ as it lies in the interior of the convex hull of the feasible region. The results of Section 4 will show that in general, the optimal point for the VPE can be achieved by a deterministic finite-memory scheduler. ■

Contribution. The main results of this paper are the following:

1. Among all schedulers that optimize the expected accumulated weight before reaching a target, a variance-minimal scheduler can be computed in polynomial time and chosen to be memoryless and deterministic (Section 3).
2. The maximal VPE in MDPs with non-negative weights can be computed in exponential space. The maximum is obtained by a deterministic scheduler that can be computed in exponential space as well (Section 4). As memory, an optimal scheduler only needs to keep track of the accumulated weight up to a bound computable in polynomial time. As soon as

the bound is reached, optimal schedulers can switch to the behavior of a variance-minimal scheduler among the expectation-minimal schedulers that can be computed by result 1.

3. The threshold problem whether the maximal VPE is greater or equal to a rational ϑ is in NEXPTIME and EXPTIME-hard (Section 4).

Related work. *Accumulated rewards.* In [**?**], a characterization of variance-minimal scheulders among the schedulers maximizing the expected accumulated weight in MDPs is given. Here, we provide a simpler proof based on the calculations of [**?**]; we moreover show how to compute such schedulers in polynomial time. [**?**] also contains hints for a similar characterization of discounted reward, and developments for mean payoff. Another closely related work is [**?**] which study the following multi-objective problem for the accumulated weight in finite-horizon MDPs: given η, ν is there a scheduler achieving an expectation of at least η, and a variance of at most ν? This problem is shown to be NP-hard, and exact pseudo-polynomial time algorithm is given for the existence of a scheduler with expectation η and variance $\leq \nu$. Furthermore, pseudo-polynomial approximation algorithms are given for optimizing the expectation under a constraint on the variance, and optimizing the variance under a constraint on the expectation.

Discounted rewards. In [**?**], the author proves that memoryless *moment-optimal* schedulers exist for the discounted reward, that is, schedulers that maximize the expectation, minimize the variance, maximize the third moment, and so on. Moreover, an algorithm is described to compute such schedulers. In [**?**], a formula for the variance of the discounted reward is given for memoryless schedulers and for the finite-horizon case, in MDPs and semi-MDPs. Variance-minimal schedulers among those maximizing the expected discounted reward until a target set is reached are studied in [**?**] for MDPs with varying discount factors. [**?**] presents a policy iteration algorithm to minimize variance of the discounted weight among schedulers achieving an expectation equal to a given constant.

Mean payoff. For mean payoff objectives, variance was studied in [**?**] for memoryless strategies, and algorithms were given to compute schedulers that achieve given bounds on the expectation and the variance [**?**]. The latter paper also considers the minimization of the variability, which is the average of the squared differences between the expected mean-payoff and each observed one-step reward. In [**?**], the author considers optimizing the expected mean payoff and the average variance. Average variance is defined as the limsup of the variances of the partial sums. They show how to minimize average variance among ϵ-optimal strategies for the expected mean payoff. Policy iteration algorithms were given in [**?**, **?**] to minimize variance or variability of the mean payoff (without constraints on the expectation).

Variance-penalized expectation. The VPE was studied for finite-horizon MDPs with terminal rewards in [**?**]. In [**?**], this notion was studied for the expectation and the variability of both mean payoff and discounted rewards. [**?**] presents a policy iteration algorithm converging against *local* optima for a similar measure.

2 Preliminaries

We give basic definitions and present our notation (for details, see, e.g., [**?**]). Afterwards, we provide auxiliary results on expected frequencies used in the subsequent sections.

2.1 Notation and definitions

Notations for Markov decision processes. A *Markov decision process* (MDP) is a tuple $\mathcal{M} = (S, Act, P, s_{init}, goal, wgt)$ where S is a finite set of states, Act a finite set of

actions, $P\colon S \times Act \times S \to [0,1] \cap \mathbb{Q}$ the transition probability function, $s_{init} \in S$ the initial state, $goal \in S$ a designated target state, and $wgt\colon S \times Act \to \mathbb{Z}$ the weight function. 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 α 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 $Act(s)$. In this paper, for all MDPs, we assume that *goal* is the only *trap* state in which no actions are enabled, that *goal* is reachable from all other states s, and that all states are reachable from s_{init}. The paths of $\mathcal{M}$ are finite or infinite sequences $s_0\,\alpha_0\,s_1\,\alpha_1\ldots$ where states and actions alternate such that $P(s_i,\alpha_i,s_{i+1}) > 0$ for all $i \geq 0$. For $\pi = s_0\,\alpha_0\,s_1\,\alpha_1\,\ldots\alpha_{k-1}\,s_k$, $wgt(\pi) = wgt(s_0,\alpha_0) + \ldots + wgt(s_{k-1},\alpha_{k-1})$ denotes the accumulated weight of π, $P(\pi) = P(s_0,\alpha_0,s_1)\cdot\ldots\cdot P(s_{k-1},\alpha_{k-1},s_k)$ 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 $wgt(s,\alpha)$.

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 cycles whose accumulated weight is 0 (so-called 0-cycles) so that the accumulated weight is bounded on all (infinite) paths in the end component. We will further use the *mean payoff* measure as tool to classify end-components. For an infinite path ζ, the mean payoff is defined as $\mathbb{MP}(\zeta) = \liminf_{n\to\infty} \frac{1}{n} wgt(pref(\zeta,n))$ where $pref(\zeta,n)$ is the prefix of length n of ζ.

Scheduler. A *scheduler* for $\mathcal{M}$ is a function $\mathfrak{S}$ that assigns to each non-maximal path π 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 π and π' 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 $\mathfrak{S}$ that satisfies $\mathfrak{S}(\pi) = \mathfrak{S}(\pi')$ for all pairs of finite paths π and π' with $last(\pi) = last(\pi')$ and $wgt(\pi) = wgt(\pi')$ is called *weight-based* and can be viewed as a function from state-weight pairs $S \times \mathbb{Z}$ to probability distributions over actions. If there is a finite set X of memory modes and a memory update function $U : S \times Act \times S \times X \to X$ 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 π in which case we also view the scheduler as a mapping to actions in $Act(last(\pi))$. Given a scheduler $\mathfrak{S}$, $\zeta = s_0\,\alpha_0\,s_1\,\alpha_1\ldots$ is a $\mathfrak{S}$-path iff ζ is a path and $\mathfrak{S}(s_0\,\alpha_0\ldots\alpha_{k-1}\,s_k)(\alpha_k) > 0$ for all $k \geq 0$. Given a scheduler $\mathfrak{S}$ and a finite $\mathfrak{S}$-path π, we define the residual scheduler $\mathfrak{S}{\uparrow}\pi$ by $\mathfrak{S}{\uparrow}\pi(\rho) = \mathfrak{S}(\pi \circ \rho)$ for each finite path ρ starting in $last(\pi)$.

Probability measure. We write $\Pr^{\mathfrak{S}}_{\mathcal{M},s}$ 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 σ-algebra generated by the cylinder sets $Cyl(\pi)$ of all maximal extensions of a finite path $\pi = s_0\,\alpha_0\,s_1\,\alpha_1\,\ldots\alpha_{k-1}\,s_k$ starting in state s, i.e., $s_0 = s$, by assigning to $Cyl(\pi)$ the probability that π is realized under $\mathfrak{S}$, which is $\mathfrak{S}(s_0)(\alpha_0)\cdot P(s_0,\alpha_0,s_1)\cdot\mathfrak{S}(s_0\alpha_0 s_1)(\alpha_1)\cdot\ldots\cdot\mathfrak{S}(s_0\alpha_0\ldots s_{k-1})(\alpha_{k-1})\cdot P(s_{k-1},\alpha_{k-1},s_k)$. For details, see [**?**].

For a random variable X that is defined on (some) 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 $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}(X)$. We define $\mathbb{E}^{\min}_{\mathcal{M},s}(X) = \inf_{\mathfrak{S}} \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}(X)$ and $\mathbb{E}^{\max}_{\mathcal{M},s}(X) = \sup_{\mathfrak{S}} \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}(X)$ 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 $\mathbb{V}^{\mathfrak{S}}_{\mathcal{M},s}(X)$ and defined by

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

Furthermore, for a measurable set of paths ψ with positive probability, $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}(X|\psi)$ denotes the conditional expectation of X under ψ. If $s = s_{init}$, we sometimes drop the subscript s.

These notations are extended to end-components of a given MDP, which are themselves seen as MDPs. We may, for instance, write $\mathbb{E}^{\min}_{\mathcal{E},s}(X)$ where $\mathcal{E}$ is an end-component of $\mathcal{M}$, and s is a state in $\mathcal{E}$, and the minimization ranges over schedulers of $\mathcal{M}$ that do not leave $\mathcal{E}$.

Accumulated weight. For maximal paths ζ of $\mathcal{M}$, we define the following random variable $\Diamond\!\!\!\!+ goal$:

$$\Diamond\!\!\!\!+ goal(\zeta) = \begin{cases} wgt(\zeta) & \text{if } \zeta \vDash \Diamond goal, \\ undefined & \text{otherwise.} \end{cases}$$

Recall that we only take schedulers under which a random variable is defined almost surely into account when addressing minimal or maximal expected values. For the expected value of $\Diamond\!\!\!\!+ goal$ to be defined, it is necessary that *goal* is reached almost surely. We call a scheduler $\mathfrak{S}$ with $\Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) = 1$ *proper*. So, in the definition of the maximal (or minimal) expected accumulated weight $\mathbb{E}^{\max}_{\mathcal{M}}(\Diamond\!\!\!\!+ goal) = \sup_{\mathfrak{S}} \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond\!\!\!\!+ goal)$, $\mathfrak{S}$ ranges over all proper schedulers.

2.2 Auxiliary conclusions from results on expected frequencies

In this section, we present conclusions from well-known results on the expected frequencies of state-weight pairs in MDPs in the formulation in which we use them in the paper. Let $\mathcal{M} = (S, Act, P, s_{init}, goal, wgt)$ be an MDP with weights in $\mathbb{Z}$ and let $\mathfrak{S}$ be a scheduler. For each state-weight pair $(s, w) \in S \times \mathbb{Z}$, we define the *expected frequency* $\vartheta^{\mathfrak{S}}_{s,w}$ under $\mathfrak{S}$ by

$$\vartheta^{\mathfrak{S}}_{s,w} \stackrel{\text{def}}{=} \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\text{number of visits to } s \text{ with accumulated weight } w)$$

where the random variable "number of visits to s with accumulated weight w" counts the number of prefixes π of a maximal paths ζ with $last(\pi) = s$ and $wgt(\pi) = w$. Note also that in MDPs $\mathcal{M}$ in which all end components have negative maximal expected mean-payoff, the expected frequencies of all state-weight pairs are finite under any scheduler.

▶ **Lemma 2.** *Let $\mathcal{M}$ be an MDP and let $\mathfrak{S}$ be a scheduler such that the expected frequency $\vartheta^{\mathfrak{S}}_{s,w}$ are finite for all state-weight pairs $(s, w) \in S \times \mathbb{Z}$. Then, there is a weight-based (randomized) scheduler $\mathfrak{T}$ with $\vartheta^{\mathfrak{S}}_{s,w} = \vartheta^{\mathfrak{T}}_{s,w}$ for all $(s, w) \in S \times \mathbb{Z}$.*

Proof sketch. Analogous to [**?**, Theorem 5.5.1]: For each state-weight pair (s, w) and each action $\alpha \in Act(s)$, let $\vartheta^{\mathfrak{S}}_{s,w,\alpha}$ be the expected number of times that α is chosen under $\mathfrak{S}$ after finite path ending in state s with weight w. Define the scheduler $\mathfrak{T}$ as a function from $S \times \mathbb{Z} \to \text{Distr}(Act)$ by letting

$$\mathfrak{T}(s, w)(\alpha) \stackrel{\text{def}}{=} \frac{\vartheta^{\mathfrak{S}}_{s,w,\alpha}}{\vartheta^{\mathfrak{S}}_{s,w}}.$$ ◀

▶ **Corollary 3.** *Let $\mathcal{M}$ be an MDP. Let $\mathfrak{S}$ be a scheduler for which $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond\!\!\!\!+ goal)$ and $\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond\!\!\!\!+ goal)$ are defined and for which the expected frequency $\vartheta^{\mathfrak{S}}_{s,w}$ are finite for all state-weight pairs $(s, w) \in S \times \mathbb{Z}$. Then, there is a weight-based scheduler $\mathfrak{T}$ with*

$$\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond\!\!\!\!+ goal) = \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond\!\!\!\!+ goal) \qquad \text{and} \qquad \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond\!\!\!\!+ goal) = \mathbb{V}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond\!\!\!\!+ goal).$$

Proof. The expected value and the variance of $\diamondsuit goal$ under a scheduler $\mathfrak{S}$ depend only on the expected frequencies $\vartheta^{\mathfrak{S}}_{goal,w}$ with $w \in \mathbb{Z}$. ◀

In this paper, we address questions concerning the possible combinations of expected value and variance of the random variable $\diamondsuit goal$. Due to this corollary, we can restrict our attention to weight-based schedulers for all investigations in the sequel.

Given two scheduler $\mathfrak{S}$ and $\mathfrak{T}$, our definition of schedulers does not directly allow us to define a new scheduler $\mathfrak{R}$ that behaves according to $\mathfrak{S}$ with probability $p \in (0,1)$ and according to $\mathfrak{T}$ with probability $1-p$. For each state-weight pair (s,w) the expected frequency under the hypothetical scheduler $\mathfrak{R}$ would be $p \cdot \vartheta^{\mathfrak{S}}_{s,w} + (1-p) \cdot \vartheta^{\mathfrak{T}}_{s,w}$. The following lemma states that a scheduler achieving these frequencies exists:

▶ **Lemma 4.** *Let $\mathcal{M}$ be an MDP as above and let $\mathfrak{S}$ and $\mathfrak{T}$ be schedulers such that the expected frequency $\vartheta^{\mathfrak{S}}_{s,w}$ and $\vartheta^{\mathfrak{T}}_{s,w}$ are finite for all state-weight pairs $(s,w) \in S \times \mathbb{Z}$. Further, let $p \in (0,1)$. Then, there exists a scheduler $\mathfrak{R}$ such that $\vartheta^{\mathfrak{R}}_{s,w} = p \cdot \vartheta^{\mathfrak{S}}_{s,w} + (1-p) \cdot \vartheta^{\mathfrak{T}}_{s,w}$ for all state-weight pairs (s,w).*

Proof sketch. Let $\vartheta^{\mathfrak{S}}_{s,w,\alpha}$ be defined as in the proof above. We define the weight-based scheduler $\mathfrak{R}$ as follows: For all state-weight pairs (s,w) and all $\alpha \in Act(s)$, let

$$\mathfrak{R}(s,w)(\alpha) = \frac{p \cdot \vartheta^{\mathfrak{S}}_{s,w,\alpha} + (1-p) \cdot \vartheta^{\mathfrak{T}}_{s,w,\alpha}}{p \cdot \vartheta^{\mathfrak{S}}_{s,w} + (1-p) \cdot \vartheta^{\mathfrak{T}}_{s,w}}.$$

The proof of the correctness is analogous to [**?**, Theorem 9.12]. ◀

This lemma allows us to introduce the following notation:

▶ **Definition 5.** *Given $\mathcal{M}$, $\mathfrak{S}$ and $\mathfrak{T}$ as in the previous lemma, we denote the scheduler $\mathfrak{R}$ whose existence is stated in the lemma by $p \cdot \mathfrak{S} \oplus (1-p) \cdot \mathfrak{T}$.*

3 Minimal variance among expectation-optimal schedulers

Let us call a scheduler *expectation-optimal* if it maximizies the expectation of $\diamondsuit goal$ from a given state s. In this section, we prove a result that is of interest in its own right and that will play a crucial role in our investigation of the optimization of the VPE in the following section. Namely, we show how to compute a scheduler that minimizes the variance among expectation-optimal schedulers in polynomial time. Note that in MDPs with weights in $\mathbb{Z}$, the mimimization of the expectation of $\diamondsuit goal$ can be reduced to the maximization by multiplying all weights with -1. This change of weights does not affect the variance and hence all results of this section also apply to expectation-minimal schedulers.

We assume that in a given MDP $\mathcal{M} = (S, Act, P, s_{init}, wgt, goal)$, the maximal achievable expectation of $\diamondsuit goal$ is finite. This can be checked in polynomial time [**?**] and, when this value is finite, it is achievable by memoryless deterministic strategies. By [**?**], all end components E of $\mathcal{M}$ are then either 0-end components or satisfy $\mathbb{E}^{\max}_E(\mathbb{MP}) < 0$.

The algorithm proceeds as follows. First, a transformation is applied so as to ensure that the only end-components in $\mathcal{M}$ are such that the maximal achievable expected mean payoff is negative; while preserving the expectation and the variance of $\diamondsuit goal$ (Lemma 6). We then prune the MDP so that all actions are optimal for maximizing the expected $\diamondsuit goal$ (Lemma 7). It follows that all schedulers then achieve the same expected $\diamondsuit goal$. We then derive an equation system in which the variances at each state are unknowns, while the expectations are known constants (Lemma 8). We conclude by showing that this equation

system admits a unique solution and is solvable in polynomial time. Omitted proofs can be found in Appendix A.

▶ **Lemma 6 ([?]).** *Let $\mathcal{M} = (S, Act, P, s_{init}, wgt, goal)$ be an MDP with $\mathbb{E}^{\max}_{\mathcal{M}}(\Diamond goal) < \infty$. There is a polynomial transformation which outputs an MDP $\mathcal{M}'$ with the following properties:*

1. *$\mathcal{M}'$ has no 0-end-components,*
2. *there is a mapping f from schedulers of $\mathcal{M}$ to those of $\mathcal{M}'$ such that for all* proper *schedulers $\mathfrak{S}$ for $\mathcal{M}$, $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) = \mathbb{E}^{f(\mathfrak{S})}_{\mathcal{M}'}(\Diamond goal)$, and $\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) = \mathbb{V}^{f(\mathfrak{S})}_{\mathcal{M}'}(\Diamond goal)$.*
3. *there is a mapping g from schedulers of $\mathcal{M}'$ to those of $\mathcal{M}$ such that for all* proper *schedulers $\mathfrak{S}$ for $\mathcal{M}'$, $\mathbb{E}^{g(\mathfrak{S})}_{\mathcal{M}}(\Diamond goal) = \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal)$, and $\mathbb{V}^{g(\mathfrak{S})}_{\mathcal{M}}(\Diamond goal) = \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal)$.*

From now on, by the previous lemma, we assume that $\mathcal{M}$ only has end-components E with $\mathbb{E}^{\max}_{\mathcal{M}}(\mathbb{MP}) < 0$. We start by computing $\mathbb{E}^{\max}_{\mathcal{M}}(\Diamond goal)$ with the following equation:

$$\mu_s = \begin{cases} 0 & \text{if } s = goal, \\ \max_{a \in Act(s)} \sum_{s' \in S} P(s,a,s')(wgt(s,a) + \mu_{s'}) & \text{otherwise.} \end{cases} \qquad (*)$$

By [**?**], $(*)$ has the unique solution $\mu_s = \mathbb{E}^{\max}_{\mathcal{M}}(\Diamond goal)$ and this solution is computable in polynomial time via linear programming. Let us define $Act^{\max}(s)$ as the set of actions from s which satisfy $(*)$ with equality, i.e. $Act^{\max}(s) \stackrel{\text{def}}{=} \{a \in Act(s) \mid \mu_s = wgt(s,a) + \sum_{s' \in S} P(s,a,s')\mu_{s'}\}$, and let $\mathcal{M}'$ be obtained by restricting $\mathcal{M}$ to actions from $Act^{\max}$. By standard arguments (see Appendix A), we can show the following lemma:

▶ **Lemma 7.** *Let $(\mu_s)_{s \in S}$ be the solution of $(*)$ for an MDP $\mathcal{M}$. Let $\mathcal{M}'$ obtained from $\mathcal{M}$ as above. Then, $\mathcal{M}'$ has no end-components. Moreover, for all $s \in S$, all schedulers $\mathfrak{S}$ of $\mathcal{M}'$ achieve $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}[\Diamond goal] = \mu_s$.*

So, in order to find the variance-minimal scheduler among expectation optimal schedulers for $\mathcal{M}$, it is sufficient to find a variance-minimal scheduler for $\mathcal{M}'$. We derive the following lemma by adapting [**?**] to MDPs.

▶ **Lemma 8.** *Consider an MDP $\mathcal{M}$, and assume that there is a vector $(\mu_s)_{s \in S}$ of values such that all schedulers $\mathfrak{S}$ satisfy $\forall s \in S, \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}(\Diamond goal) = \mu_s$. Then, $(\mathbb{V}^{\inf}_{\mathcal{M},s}(\Diamond goal))_{s \in S}$ is the unique solution of the following equation:*

$$V_s = \begin{cases} 0 & \textit{if } s = goal, \\ \min_{a \in Act(s)} \sum_{t \in S} P(s,a,t)\left((wgt(s,a) + \mu_t - \mu_s)^2 + V_t\right) & \textit{otherwise.} \end{cases} \qquad (**)$$

Note that the equation system $(**)$ is the same as the equation system used to minimize the expected accumulated weight before reaching *goal* under the weight function wgt' that assigns the non-negative weight $(wgt(s,a) + \mu_t - \mu_s)^2$ to the transition (s, α, t). So, this equation system is solvable in polynomial time [**?**]. Using that all schedulers in $\mathcal{M}'$ achieve an expected accumulated weight of μ_s when starting in state s, the results of this section can be combined to the following theorem.

▶ **Theorem 9.** *Given an MDP $\mathcal{M}$ such that $\mathbb{E}^{\max}_{\mathcal{M}}[\Diamond goal] < \infty$, a memoryless deterministic, expectation-optimal scheduler $\mathfrak{S}$ such that $\mathbb{V}^{\mathfrak{S}}_{\mathcal{M},s}[\Diamond goal]$ is minimal among all expectation-optimal schedulers for any state s is computable in polynomial time.*

4 Variance-penalized expectation

The goal of this section is to develop an algorithm to compute the optimal *variance-penalized expectation* (VPE). Given a rational $\lambda > 0$, we define the VPE with parameter λ under a

scheduler $\mathfrak{S}$ as

$$\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} \stackrel{\text{def}}{=} \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) - \lambda \cdot \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) = \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) - \lambda \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal^2) + \lambda \cdot (\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal))^2.$$

Task: Compute the maximal variance-penalized expectation

$$\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}} \stackrel{\text{def}}{=} \sup_{\mathfrak{S}} \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$$

where the supremum ranges over all proper schedulers. Furthermore, compute an optimal scheduler $\mathfrak{S}$ with $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$.

Throughout this section, we will restrict ourselves to MDPs $\mathcal{M} = (S, Act, P, s_{init}, wgt, goal)$ with a weight function $wgt\colon S \times Act \to \mathbb{N}$, i.e., we only consider MDPs with non-negative weights. Key results established in this section do not hold in the general setting with arbitrary weights and further complications arise. In the conclusions we will briefly discuss these complications.

As before, we are only interested in schedulers that reach the goal with probability 1. If the maximal expectation $\mathbb{E}^{\max}_{\mathcal{M}}(\Diamond goal) < \infty$, it is well-known that in this case of non-negative weights, all end components of $\mathcal{M}$ are 0-end components [**?**,**?**]. Hence, w.l.o.g., we can assume that $\mathcal{M}$ has no end components throughout this section by Lemma 6. In this case, $\Diamond goal$ is defined on almost all paths under any scheduler. So, in particular the values $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)$ and $\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)$ are defined for all schedulers $\mathfrak{S}$. Furthermore, as we have seen in Corollary 3, it is sufficient to consider weight-based schedulers for the optimization of VPEs. The main result of this section is the following:

▶ **Main result.** Given an MDP $\mathcal{M}$ and λ as above, the optimal value $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$ and an optimal scheduler $\mathfrak{S}$ can be computed in exponential space. Given a rational ϑ, the threshold problem whether $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}} \geq \vartheta$ is in NEXPTIME and EXPTIME-hard.

To obtain the main result, we will first prove that the maximal VPE is obtained by a deterministic scheduler (Section 4.1). This result can then be used for the EXPTIME-hardness proof for the threshold problem (Section 4.2). The key step to obtain the upper bounds of the main result is to show that optimal schedulers have to *minimize* the weight that is expected to still be accumulated after a computable bound of accumulated weight has been exceeded. We call such a bound a *saturation point* (Section 4.3). Finally, we show how to utilize the saturation point result to solve the threshold problem and to compute the optimal VPE (Section 4.4). Proofs omitted in this section can be found in Appendix B.

▶ Remark 10. In the formulation presented here, the goal is to maximize the expected accumulated weight with a penalty for the variance. All results and proofs in this section, however, hold analogously for the variant $\sup_{\mathfrak{S}} -\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) - \lambda \cdot \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)$ of the maximal VPE in which the goal is to minimize the expected accumulated weight while receiving a penalty for the variance. In particular, the same saturation point works and optimal schedulers still have to minimize the expected accumulated weight as soon as the accumulated weight exceeds the saturation point. ■

4.1 Existence of optimal deterministic schedulers

We begin this section with a lemma describing how the variance of accumulated weight behaves under convex combinations of schedulers. This will allow us to show that the maximal VPE can be approximated by deterministic schedulers with the help of Lemma 14 describing

a connection between randomization and convex combinations. This first lemma follows via basic arithmetic form the fact that the expected values of $\Diamond goal$ and $\Diamond goal^2$ depend linearly on the expected frequencies of the state-weight pairs $(goal, w)$ with $w \in \mathbb{N}$.

▶ **Lemma 11.** *Let $\mathcal{M} = (S, Act, P, s_{init}, wgt, goal)$ be an MDP with non-negative weights and no end components. Let $\mathfrak{S}$ and $\mathfrak{T}$ be two schedulers for $\mathcal{M}$. Let $p \in (0,1)$. The scheduler $\mathfrak{R} \stackrel{def}{=} p \cdot \mathfrak{S} \oplus (1-p) \cdot \mathfrak{T}$ satisfies*

$$\mathbb{V}^{\mathfrak{R}}_{\mathcal{M}}(\Diamond goal) = p \cdot \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + (1-p) \cdot \mathbb{V}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal) + p \cdot (1-p) \cdot (\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal))^2.$$

Proof sketch. The claim follows from straight-forward calculations using that the expected values of $\Diamond goal$ and $\Diamond goal^2$ depend linearly on the expected frequencies of the state-weight pairs $(goal, w)$ with $w \in \mathbb{N}$. ◀

▶ Remark 12. Given two schedulers $\mathfrak{S}$ and $\mathfrak{S}'$ under which the expectation and variance are (η, ν) and (η', ν'), respectively, such that $\eta < \eta'$, there is a unique convex combination $\mathfrak{T}$ of the two schedulers with expectation x for all $x \in [\eta, \eta']$. Viewing the variance of these convex combinations as a function $V : [\eta, \eta'] \to \mathbb{R}$, we can observe the following using the previous Lemma 11:

$$V(x) = \nu + \frac{x-\eta}{\eta'-\eta} \cdot (\nu'-\nu) + \frac{x-\eta}{\eta'-\eta} \cdot \frac{\eta'-x}{\eta'-\eta} \cdot (\eta'-\eta)^2 = \nu + \frac{x-\eta}{\eta'-\eta} \cdot (\nu'-\nu) + (x-\eta) \cdot (\eta'-x).$$

The coefficient before x^2 in this quadratic polynomial hence is always -1. ■

The following lemma stating the continuity of the VPE will be useful in several ways: If we manipulate schedulers at one state-weight pair at a time, we can reason about the scheduler we obtain in the limit after manipulating the scheduler at all state-weight pairs, e.g., in the proof of Theorem 15 below. Further, it will allow us to prove that there is an optimal scheduler, i.e., that the supremum in the definition of $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$ is in fact a maximum.

▶ **Lemma 13** (Continuity of VPE)**.** *Let $\mathcal{M}$ and $\lambda > 0$ be as above. The variance-penalized expectation as a function from weight-based schedulers to $\mathbb{R}$ is (uniformly) continuous in the following sense: Given $\varepsilon > 0$, there is a natural number N_ε such that for all weight-based schedulers $\mathfrak{S}$ and $\mathfrak{T}$ that agree on all state-weight pairs (s, w) with $w \leq N_\varepsilon$, we have*

$$\left| \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} - \mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}} \right| < \varepsilon.$$

Proof sketch. The claim follows from the fact that the probability that a high amount of weight w is accumulated under any scheduler decreases exponentially as w tends to ∞. ◀

For the final ingredient to show that deterministic schedulers approximate the optimal VPE, we take a closer look at the relation of randomization to convex combinations of schedulers. For the following lemma, let $\mathfrak{S}$ be a weight-based scheduler for an MDP $\mathcal{M}$ as before. Assume that there is a state-weight pair $(s, w) \in S \times \mathbb{N}$ reachable under $\mathfrak{S}$ such that $\mathfrak{S}$ chooses two different actions α and β with probabilities q and $1-q$, respectively, for some $q \in (0,1)$. Let $\mathfrak{S}_\alpha$ be the scheduler that agrees with $\mathfrak{S}$ on all state-weight pairs except for (s, w) and that chooses α with probability 1 at (s, w). Define $\mathfrak{S}_\beta$ analogously. The technical proof of the following lemma can be found in Appendix B.

▶ **Lemma 14.** *Let $\mathcal{M}$, $\mathfrak{S}$, $\mathfrak{S}_\alpha$, $\mathfrak{S}_\beta$, and q be as above. There is a value $p \in (0,1)$ such that the expected frequencies of all state-weight pairs are the same under $\mathfrak{S}$ and $p \cdot \mathfrak{S}_\alpha \oplus (1-p) \cdot \mathfrak{S}_\beta$.*

▶ **Theorem 15** (Deterministic schedulers approximate optimal VPE). *Let $\mathcal{M}$ be an MDP with non-negative weights and without end components and let $\lambda > 0$. For each scheduler $\mathfrak{S}$, there is a deterministic weight-based scheduler $\mathfrak{T}$ with*

$$\mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}.$$

Proof sketch. W.l.o.g., we can assume that $\mathfrak{S}$ is weight-based by Corollary 3. At a single state-weight pair (s, w) at which $\mathfrak{S}$ makes use of randomization between, we can (potentially repeatedly) apply Lemma 14 and Lemma 11 to find a scheduler $\mathfrak{S}'$ that does not make use of this randomization but satisfies $\mathbb{VPE}[\lambda]^{\mathfrak{S}'}_{\mathcal{M}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$. Going through all state-weight pairs in this fashion, we can construct an infinite sequence of schedulers with non-decreasing VPE in which randomization is successively removed at all state-weight pairs. In the limit, we obtain a well defined deterministic weight-based scheduler $\mathfrak{T}$. Lemma 13 allows us to conclude that $\mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$. ◀

In the definition of the maximal variance-penalized expectation $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}} = \sup_{\mathfrak{S}} \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$, it is sufficient to let the supremum range over deterministic weight-based schedulers $\mathfrak{S}$ in the light of this theorem. For the proof of the existence of optimal schedulers, we make use of an analytic argument: Continuous functions on compact space obtain their maximum. The continuity shown in Lemma 13 applied to the space of deterministic weight-based schedulers can be reformulated as continuity with respect to a metric on this space. Namely, we define the metric $d_{\mathcal{M}}$ on the set of deterministic weight-based schedulers as follows: Given two deterministic weight-based schedulers $\mathfrak{S}$ and $\mathfrak{T}$ for $\mathcal{M}$, first let

$$m(\mathfrak{S}, \mathfrak{T}) \stackrel{\text{def}}{=} \min\{w \mid \text{there is a state } s \in S \text{ with } \mathfrak{S}(s, w) \neq \mathfrak{T}(s, w)\}.$$

We then define $d_{\mathcal{M}}(\mathfrak{S}, \mathfrak{T}) \stackrel{\text{def}}{=} 2^{-m(\mathfrak{S},\mathfrak{T})}$. This metric indeed turns the set of deterministic weight-based schedulers into a compact space as shown in [**?**]:

▶ **Lemma 16** (Compactness of the space of deterministic weight-based schedulers [**?**]). *Let $\mathcal{M}$ be as above. The space of all deterministic weight-based schedulers with the topology induced by the metric $d_{\mathcal{M}}$ is compact.*

▶ **Theorem 17** (Existence of an optimal deterministic weight-based scheduler). *Let $\mathcal{M}$ and $\lambda > 0$ be as above. There is a deterministic weight-based scheduler $\mathfrak{S}$ with*

$$\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}.$$

Proof. The claim follows from Lemma 13, Theorem 15, and Lemma 16, as continuous functions on compact spaces obtain their maximum. ◀

4.2 Hardness of the threshold problem

The result that the maximal variance-penalized expectation can be achieved by a deterministic scheduler can be used for the following hardness result:

▶ **Theorem 18.** *Given an MDP $\mathcal{M}$ with non-negative weights and two rationals $\lambda, \vartheta > 0$, deciding whether $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}} \geq \vartheta$ is EXPTIME-hard. Furthermore, for acyclic MDPs $\mathcal{M}$, the problem is PSPACE-hard.*

Proof sketch. We reduce from the following problem which is shown to be EXPTIME-hard in general and PSPACE-hard for acyclic MDPs in [**?**]: Given an MDP $\mathcal{M}$ and a natural

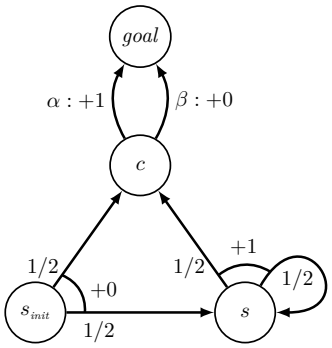

Figure 2 The MDP $\mathcal{M}$ used in Example 19.

number $T > 0$ such that *goal* is reached in $\mathcal{M}$ almost surely under all schedulers, decide whether there is a scheduler $\mathfrak{S}$ such that $\Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = T) = 1$.

The idea is to construct an MDP $\mathcal{M}'$ that reaches *goal* with weight T with probability $1/2$ directly and otherwise behaves like $\mathcal{M}$. By choosing λ sufficiently large, we can show that $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}'} \geq T$ is only possible if and only if there is a scheduler with $\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) = 0$. This scheduler then has to reach *goal* with weight T on all paths. The necessary technical calculations can be found in Appendix B.2. ◀

4.3 Saturation Point

In the sequel, we will provide a series of results that allow us to further restrict the class of deterministic schedulers that we have to consider when maximizing the variance-penalized expectation. In the end, we obtain a finite set of deterministic finite-memory schedulers among which there is a scheduler achieving the optimal variance-penalized expectation. In particular, this means that the optimum is computable.

The key step is the insight that we can provide a natural number K computable in polynomial time such that an optimal scheduler $\mathfrak{S}$ for the variance-penalized expectation has to minimize the expected accumulated weight before reaching *goal* once a weight of at least K has already been accumulated on a run. Furthermore, the behavior of $\mathfrak{S}$ after a weight of at least K has been accumulated must minimize the variance of the weight that will still be accumulated among all expectation-minimal schedulers. We call this value K a *saturation point*.

▶ **Example 19.** The MDP $\mathcal{M}$ in Figure 2 aims to provide some intuition on the results of this section. The state c in this MDP is reached with accumulated weight n with probability $(1/2)^{n+1}$ for all $n \in \mathbb{N}$. Then, the choice has to be made whether to collect weight $+1$ or 0 before moving to *goal*. We want to take a closer look at a family of special weight-based deterministic finite-memory schedulers for $\mathcal{M}$: Let $\mathfrak{S}_k$ be the scheduler that chooses action α in c if the accumulated weight is less than k and otherwise chooses action β.

For these schedulers, we can explicitly provide expectation and variance: The probability that the scheduler $\mathfrak{S}_k$ chooses α in c is $1 - (1/2)^k$. As the expected accumulated weight before reaching c is 1, we obtain a total expected accumulated weight of

$$\mathbb{E}^{\mathfrak{S}_k}_{\mathcal{M}}(\Diamond goal) = 2 - \frac{1}{2^k}.$$

To obtain the variance, we compute $\mathbb{E}^{\mathfrak{S}_k}_{\mathcal{M}}(\Diamond goal^2)$:

$$\mathbb{E}^{\mathfrak{S}_k}_{\mathcal{M}}(\Diamond goal^2) = \sum_{n=0}^{k-1} \frac{1}{2^{k+1}} \cdot (n+1)^2 + \sum_{n=k}^{\infty} \frac{1}{2^{k+1}} \cdot n^2$$
$$= \sum_{n=0}^{\infty} \frac{1}{2^{k+1}} \cdot (n+1)^2 - \sum_{n=k}^{\infty} \frac{1}{2^{k+1}} \cdot (2n+1) = 6 - \frac{2k+3}{2^k}.$$

We can then easily compute the variance

$$\mathbb{V}^{\mathfrak{S}_k}_{\mathcal{M}}(\Diamond goal) = \mathbb{E}^{\mathfrak{S}_k}_{\mathcal{M}}(\Diamond goal^2) - (\mathbb{E}^{\mathfrak{S}_k}_{\mathcal{M}}(\Diamond goal))^2 = 2 + \frac{1-2k}{2^k} - \frac{1}{4^k}.$$

For $\lambda = 1$, we obtain the following VPE:

$$\mathbb{VPE}[\lambda]^{\mathfrak{S}_k}_{\mathcal{M}} = \frac{k-1}{2^{k-1}} + \frac{1}{4^k}.$$

Comparing scheduler $\mathfrak{S}_k$ to $\mathfrak{S}_{k+1}$, we obtain: $\mathbb{VPE}[\lambda]^{\mathfrak{S}_{k+1}}_{\mathcal{M}} - \mathbb{VPE}[\lambda]^{\mathfrak{S}_k}_{\mathcal{M}} = 2 - k/2^k - 3/4^k$. This difference is negative for $k \geq 2$ and positive for $k = 1$. We conclude that among the schedulers $\mathfrak{S}_k$, the scheduler $\mathfrak{S}_2$ is VPE-optimal. Interestingly, this means choosing not to accumulate the additional weight $+1$ by choosing α is better already for small amounts of accumulated weight. Intuitively, the reason is that choosing α for an accumulated weight ≥ 2 has a larger effect on the variance than on the expectation. Increasing the expectation in particular also increases the squared deviation of the path that reach *goal* with weight 1 which has probability $1/2$ under $\mathfrak{S}_k$ for $k \geq 2$. The saturation point result of this section will tell us that an optimal scheduler always has to minimize the weight that is expected to still be accumulated weight once sufficiently much weight has already been accumulated. ∎

Let $\mathcal{M} = (S, Act, P, s_{init}, wgt, goal)$ be an MDP without end components and with non-negative weights as above and let $\lambda > 0$ be a rational. Before we define K and show that it can be computed in polynomial time, we need some additional notation.

For each state $s \in S$, define $e_s \stackrel{\text{def}}{=} \mathbb{E}^{\min}_{\mathcal{M},s}(\Diamond goal)$. For each state $s \in S \setminus \{goal\}$, we define the subset $Act^{\min}(s) \subseteq Act(s)$ of actions allowing to minimize the expectation analogously to $Act^{\max}$ before: $Act^{\min}(s) \stackrel{\text{def}}{=} \{\alpha \in Act(s) \mid e_s = wgt(s,\alpha) + \sum_{t \in S} P(s,\alpha,t) \cdot e_t\}$. Choosing an action not belonging to $Act^{\min}(s)$ in state s ensures that the expected accumulated weight before reaching *goal* is higher than the minimal possible value. Further, we can define the minimal amount by which choosing a non-minimizing action increases the expectation:

$$\delta \stackrel{\text{def}}{=} \min\{(wgt(s,\alpha) + \sum_{t \in S} P(s,\alpha,t) \cdot e_t) - e_s \mid s \in S \setminus \{goal\} \text{ and } \alpha \in Act(s) \setminus Act^{\min}(s)\}.$$

If the set on the right hand side is empty, all schedulers minimize the expected accumulated weight before reaching *goal* and the claims of this section hold trivially. So, we can assume that this set is non-empty. By the definition of $Act^{\min}$, we observe that $\delta > 0$.

Next, we can compute an upper bound U_1 for $\mathbb{E}^{\max}_{\mathcal{M},s}(\Diamond goal)$ for all states s by computing the maximal value $U_1 \stackrel{\text{def}}{=} \max_{s \in S} \mathbb{E}^{\max}_{\mathcal{M},s}(\Diamond goal)$.

Finally, let ε be the minimal transition probability present in $\mathcal{M}$. As $\mathcal{M}$ has no end components, the only trap state *goal* is reached within $n \stackrel{\text{def}}{=} |S|$ steps under each scheduler with probability at least ε^n. Let W be the largest weight in $\mathcal{M}$. Within n steps at most a weight of $n \cdot W$ is accumulated. We use these observations for the following two values:

- First, we can provide a value $B_{1/2}$ such that the probability that a weight above $B_{1/2}$ is accumulated under any scheduler is at most $1/2$: For this, let $b_{1/2}$ be such that $((1-\varepsilon^n))^{b_{1/2}} \leq 1/2$. This is the case if and only if $b_{1/2}$ is at least

$$\frac{\log(1/2)}{\log(1-\varepsilon^n)} = -\frac{1}{\log(1-\varepsilon^n)} < \frac{1}{\varepsilon^n}.$$

  So, we can choose $b_{1/2}$ to be $\frac{1}{\varepsilon^n}$. Then, with probability at most $1/2$, a path has length at least $n \cdot b_{1/2}$. This allows us to defined $B_{1/2} \stackrel{\text{def}}{=} b_{1/2} \cdot n \cdot W$.
- Second, we compute an upper bound U_2 for $\max_{s \in S} \mathbb{E}^{\max}_{\mathcal{M},s}(\Diamond goal^2)$: With probability ε^n a path has weight at most $n \cdot W$; with probability $(1-\varepsilon^n) \cdot \varepsilon^n$ it has weight at most $2 \cdot n \cdot W$; with probability $(1-\varepsilon^n)^2 \cdot \varepsilon^n$ it has weight at most $3 \cdot n \cdot W$; and so on. So, we get that $\max_{s \in S} \mathbb{E}^{\max}_{\mathcal{M},s}(\Diamond goal^2) \leq \sum_{i=0}^{\infty} (1-\varepsilon^n)^i \cdot \varepsilon^n \cdot ((i+1) \cdot n \cdot W)^2$. This allows us to define

$$U_2 \stackrel{\text{def}}{=} \frac{2 \cdot n^2 \cdot W^2}{\varepsilon^{2n}} \geq \frac{(2-\varepsilon^n) \cdot n^2 \cdot W^2}{\varepsilon^{2n}} = \sum_{i=0}^{\infty} (1-\varepsilon^n)^i \cdot \varepsilon^n \cdot ((i+1) \cdot n \cdot W)^2.$$

We are now in the position to define the saturation point K: Let K be the least natural number with

$$K \geq B_{1/2} = \frac{n \cdot W}{\varepsilon^n} \qquad \text{and} \qquad K \geq \frac{U_1/\lambda + U_2 + 2U_1 + U_1^2/2}{\delta} + 1.$$

The definition of K is arguably a bit cumbersome, but the choices will become clear in the proof of Theorem 20. All values involved except for δ and U_1 can be computed directly from n, W, and ε in polynomial time. The values δ and U_1 require to maximize or minimize the expected value of $\Diamond goal$ from all states, i.e., to solve an SSPP which can be done in polynomial time by linear programming [**?**, **?**].

▶ **Theorem 20** (Saturation point). *Let $\mathcal{M}$, $\lambda > 0$ and K be as above. Let $\mathfrak{S}$ be a scheduler with $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$. Then, for each finite $\mathfrak{S}$-path π with $wgt(\pi) \geq K$, the residual scheduler $\mathfrak{S}{\uparrow}\pi$ satisfies*

$$\mathbb{E}^{\mathfrak{S}\uparrow\pi}_{\mathcal{M},last(\pi)}(\Diamond goal) = \mathbb{E}^{\min}_{\mathcal{M},last(\pi)}(\Diamond goal).$$

Proof sketch. Let $\mathfrak{S}$ be a scheduler with $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$. Suppose there is a $\mathfrak{S}$-path π' with $wgt(\pi') \geq K$ such that $\mathbb{E}^{\mathfrak{S}\uparrow\pi'}_{\mathcal{M},last(\pi')}(\Diamond goal) > \mathbb{E}^{\min}_{\mathcal{M},last(\pi')}(\Diamond goal)$. Then, there must be an $\mathfrak{S}$-path π that extends π' such that $\mathfrak{S}$ chooses an action $\alpha \notin Act^{\min}(last(\pi))$ with positive probability.

The residual scheduler $\mathfrak{T}$ of $\mathfrak{S}$ after π in case $\mathfrak{S}$ chooses α then satisfies $\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},last(\pi)}(\Diamond goal) \geq \mathbb{E}^{\min}_{\mathcal{M},last(\pi)}(\Diamond goal) + \delta$. We let $\mathfrak{S}'$ be a scheduler that behaves like $\mathfrak{S}$ unless $\mathfrak{S}$ chooses α after π. In this case, $\mathfrak{S}'$ minimizes the expected value of $\Diamond goal$ from then on. We consider the difference $\mathbb{VPE}[\lambda]^{\mathfrak{S}'}_{\mathcal{M}} - \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$. Using the bounds U_1 and U_2 and that the probability of π is at most $1/2$ as $K \geq B_{1/2}$, we obtain a lower bound for this difference that consists of an expression in terms of U_1, U_2, and λ plus the term

$$\lambda \cdot wgt(\pi) \cdot (\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},last(\pi)}(\Diamond goal) - \mathbb{E}^{\min}_{\mathcal{M},last(\pi)}(\Diamond goal)).$$

Observing that this term is greater or equal to $\lambda \cdot K \cdot \delta$, the definition of K was chosen exactly so that we can conclude that $\mathbb{VPE}[\lambda]^{\mathfrak{S}'}_{\mathcal{M}} - \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} > 0$. So, $\mathfrak{S}$ was not VPE-maximal yielding a contradiction. ◀

By the results of Section 3, there is a memoryless deterministic scheduler $\mathfrak{V}$ that minimizes the variance among all schedulers minimizing the expected accumulated weight before reaching *goal*. More precisely, for all states s, the scheduler $\mathfrak{V}$ satisfies $\mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal) = \mathbb{E}^{\min}_{\mathcal{M},s}(\Diamond goal)$ and $\mathbb{V}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal) = \inf_{\mathfrak{M}} \mathbb{V}^{\mathfrak{M}}_{\mathcal{M},s}(\Diamond goal)$ where the infimum ranges over all schedulers $\mathfrak{M}$ with $\mathbb{E}^{\mathfrak{M}}_{\mathcal{M},s}(\Diamond goal) = \mathbb{E}^{\min}_{\mathcal{M},s}(\Diamond goal)$. We use the existence of this scheduler in the following theorem.

▶ **Theorem 21.** *Let $\mathcal{M}$, $\lambda > 0$, K, and $\mathfrak{V}$ be as above. Let $\mathfrak{S}$ be a deterministic scheduler with $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$. Let $\mathfrak{T}$ be the scheduler that agrees with $\mathfrak{S}$ on all paths π with weight less than K and that chooses actions according to the memoryless deterministic scheduler $\mathfrak{V}$ after paths π' with $wgt(\pi') \geq K$. This scheduler $\mathfrak{T}$ satisfies $\mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$, too.*

Proof sketch. Given a scheduler $\mathfrak{S}$ with $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$, and a path π with $wgt(\pi) \geq K$, we compare the scheduler $\mathfrak{S}$ to the scheduler $\mathfrak{S}'$ that behaves like $\mathfrak{S}$, but switches to the behavior of $\mathfrak{V}$ after π. We obtain that $\mathbb{VPE}[\lambda]^{\mathfrak{S}'}_{\mathcal{M}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$ is equivalent to $\mathbb{E}^{\mathfrak{V}}_{\mathcal{M}}(\Diamond goal^2) \leq \mathbb{E}^{\mathfrak{S}\upharpoonright\pi}_{\mathcal{M}}(\Diamond goal^2)$. This holds because $\mathbb{V}^{\mathfrak{V}}_{\mathcal{M}}(\Diamond goal) \leq \mathbb{V}^{\mathfrak{S}\upharpoonright\pi}_{\mathcal{M}}(\Diamond goal)$ as $\mathfrak{V}$ and $\mathfrak{S}\upharpoonright\pi$ achieve the same expectation. Using a continuity argument as before, we show that changing the behavior of $\mathfrak{S}$ to $\mathfrak{V}$ after all paths with weight at least K does not decrease the VPE. ◀

Put together, we have shown that the maximal VPE is obtained by a weight-based deterministic scheduler that switches to the memoryless behavior of $\mathfrak{V}$ as soon as a weight of at least K has been accumulated, which also means that it uses only finite memory.

4.4 Computation of the optimal VPE

Given an MDP $\mathcal{M} = (S, Act, P, s_{init}, wgt, goal)$ with non-negative weights and without end components and $\lambda > 0$ as before, let K be the saturation point given above. Note that K is computable in polynomial time and that hence its numerical value is at most exponential in the size of $\mathcal{M}$. We construct the following MDP $\mathcal{M}'$ that encodes the weights that are accumulated until the saturation point is exceeded into the state space: Let W be the maximal weight occurring in $\mathcal{M}$. The set of states is $S' = S \times \{0, 1, \ldots, K + W - 1\}$. The set of actions remains unchanged. The new probability transition function is given by $P'((s, w), (t, w + wgt(s, \alpha))) \stackrel{\text{def}}{=} P(s, \alpha, t)$ for all $s, t \in S$, all $w < K$, and all $\alpha \in Act(s)$. All remaining transition probabilities are 0. Note that this means that all states of the form $(goal, w)$ with $w \in \{0, 1, \ldots, K + W - 1\}$ and of the form (s, w) with $s \in S$ and $w \in \{K, K + 1, \ldots, K + W - 1\}$ are trap states in $\mathcal{M}'$. The initial state is $s'_{init} \stackrel{\text{def}}{=} (s_{init}, 0)$. The weight function is not relevant in $\mathcal{M}'$.

Let $\mathfrak{V}$ be the memoryless deterministic scheduler for $\mathcal{M}$ as in Theorem 21 that specifies the optimal behavior in order to maximize the variance-penalized expectation as soon as a weight of at least K has been accumulated. Let us call the set of weight-based deterministic schedulers for $\mathcal{M}$ that behave like $\mathfrak{V}$ after a weight of at least K has been accumulated by $\mathrm{WD}_K(\mathcal{M})$. Clearly, there is a natural one-to-one-correspondence between memoryless deterministic schedulers for $\mathcal{M}'$ and schedulers in $\mathrm{WD}_K(\mathcal{M})$.

By the results of Section 3, for each state $s \in S$, we can compute the values

$$e_s \stackrel{\text{def}}{=} \mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal) \qquad \text{and} \qquad q_s \stackrel{\text{def}}{=} \mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal^2) = \mathbb{V}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal) + e_s^2$$

in polynomial time. The following lemma now allows us to express the VPE in $\mathcal{M}$ in terms of reachability probabilities in $\mathcal{M}'$.

▶ **Lemma 22.** *Let $\mathcal{M}$, $\mathcal{M}'$, K and λ be as above. Given a scheduler $\mathfrak{S} \in \mathrm{WD}_K(\mathcal{M})$ also viewed as a memoryless deterministic scheduler for $\mathcal{M}'$, let*

$$\mu \stackrel{def}{=} \sum_{w<K} \Pr^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond(goal, w)) \cdot w + \sum_{s\in S, w\geq K} \Pr^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond(s,w)) \cdot (w + e_s). \tag{†}$$

Then,

$$\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mu - \lambda \cdot \Big(\sum_{w<K} \Pr^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond(goal, w)) \cdot (w-\mu)^2) + \sum_{s\in S, w\geq K} \Pr^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond(s,w)) \cdot ((w-\mu)^2 + 2(w-\mu)e_s + q_s)) \Big). \tag{‡}$$

Proof. It is clear that $\mu = \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)$. So, we have to show that

$$\begin{aligned} &\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) = \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}((\Diamond goal - \mu)^2) \\ =& \sum_{w<K} \Pr^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond(goal, w)) \cdot (w-\mu)^2) \\ &+ \sum_{s\in S, w\geq K} \Pr^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond(s,w)) \cdot ((w-\mu)^2 + 2(w-\mu)e_s + q_s)). \end{aligned} \tag{∘}$$

The event $\Diamond(s,w)$ that (s,w) is reached in $\mathcal{M}'$ corresponds to the event that a path in $\mathcal{M}$ has a prefix of weight w ending in s. We denote this event in $\mathcal{M}$ also by $\Diamond(s,w)$. If $\Pr^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond(s,w)) > 0$ for $w \geq K$, then

$$\begin{aligned} &\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}((\Diamond goal - \mu)^2 | \Diamond(s,w)) \\ =&\mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}((\Diamond goal + w - \mu)^2) \\ =&(w-\mu)^2 + 2(w-\mu)\cdot \mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal) + \mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal^2). \end{aligned}$$

So, the sums in equation (∘) sum up the conditional expectation of $(\Diamond goal - \mu)^2$ in $\mathcal{M}$ under the conditions that $(goal, w)$ is reached for $w < K$ or that the state s is the first one reached when the accumulated weight exceeds K with weight $w \geq K$, multiplied by the respective probabilities of the conditions. ◀

Putting everything together, we arrive at the main result.

▶ **Theorem 23.** *Let $\mathcal{M}$ and λ be as above. Given a rational ϑ, the threshold problem whether $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}} \geq \vartheta$ is in NEXPTIME. The optimal value $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$ and an optimal scheduler can be computed in exponential space.*

Proof. The threshold problem can be decided in non-deterministic exponential time as follows: Given $\mathcal{M}$ and λ, compute K in polynomial time and construct $\mathcal{M}'$ as above (of exponential size) in exponential time. Guess a memoryless deterministic scheduler $\mathfrak{S}$ for $\mathcal{M}'$ also viewed as a scheduler in $\mathrm{WD}_K(\mathcal{M})$. The reachability probabilities for all trap states in $\mathcal{M}'$ under $\mathfrak{S}$ can then be computed in time polynomial in the size of $\mathcal{M}'$. With the help of equations (†) and (‡) from Lemma 22, $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$ can be computed from these reachability probabilities in time polynomial in the size of $\mathcal{M}'$. If $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} \geq \vartheta$, accept. By Theorem 21, $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}} \geq \vartheta$ iff there is a scheduler $\mathfrak{S}$ in $\mathrm{WD}_K(\mathcal{M})$ with $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} \geq \vartheta$. Due to the one-to-one correspondence between schedulers in $\mathrm{WD}_K(\mathcal{M})$ and memoryless deterministic schedulers for $\mathcal{M}'$, this establishes the correctness of the algorithm.

To compute the optimal value $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$, we compute $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$ for all schedulers $\mathfrak{S}$ in $\mathrm{WD}_K(\mathcal{M})$ in the same fashion and always store the highest value found so far. As the memoryless schedulers for $\mathcal{M}'$ have an exponentially large representation, this can be done in exponential space and the optimal scheduler can be returned as well. ◀

5 Conclusion

In our results, there remains a complexity gap between the EXPTIME-lower bounds and the exponential-space and NEXPTIME-upper bounds for the optimization of the VPE in MDPs with non-negative weights and the corresponding threshold problem, respectively. Here, we want to shed some light on this complexity gap: It is well-known that the possible vectors of expected frequencies of all states in an MDP can be characterized by a linear equation system (see, e.g., [**?**]). Using this linear equation system for the exponentially large MDP constructed in Section 4.4 and equations (†) and (‡) from that section, the threshold problem for the maximal VPE can be reformulated as the satisfiability problem of an exponentially sized system of quadratic inequalities. The optimization problem can likewise be formulated as an exponentially large *quadratically constrained quadratic program* (QCQP). This satisfiability problem and QCQPs are NP-hard in general. The question whether the inequality system of exponential size we obtain here has a special structure which allows it to be solved in exponential time remains open here.

This observation stands in contrast to conceptually similar saturation point results straight-forwardly leading to exponential time algorithms (see, e.g., [**?**]). For example, the threshold problem for conditional expectations: "Given a set $T \subseteq S$, is there a scheduler $\mathfrak{S}$ with $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal \mid \Diamond T) \geq \vartheta$?" admits a saturation point result in MDPs with non-negative weights as well [**?**]. Deriving a system of inequalities as above, however, leads to a system of linear inequalities after straight-forward transformations. Hence, this approach directly leads to an exponential time algorithm for the threshold problem for conditional expectations. For the VPE, the system of inequalities seems to be inherently of a polynomial nature which can be seen as an indication that the situation here is fundamentally more difficult.

Further, we restricted our attention to MDPs with non-negative weights. When allowing positive and negative weights, the key result, the existence of a saturation point, does not hold anymore. For conditional expectations and other problems relying on the existence of a saturation point, the switch to integer weights makes the problems even at least as hard as the Positivity problem for linear recurrence sequences, a number theoretic problem whose decidability has been open for many decades (see [**?**, **?**]). The question whether such a hardness result for the threshold problem of the VPE, rendering decidability impossible without a breakthrough in number theory, can be established remains as future work.

Further possible directions of research include the investigation of the following multi-objective threshold problem: Given η and ν, is there a scheduler with expectation at least η and variance at most ν? As the variance treats good and bad outcomes symmetrically, replacing the variance in the VPE by a one-sided deviation measure, such as the lower semi-variance that only takes the outcomes worse than the expected value into account, constitutes another natural extension of this work.

A Omitted proofs of Section 3

▶ **Lemma 6 ([?]).** *Let $\mathcal{M} = (S, Act, P, s_{init}, wgt, goal)$ be an MDP with $\mathbb{E}^{\max}_{\mathcal{M}}(\Diamond goal) < \infty$. There is a polynomial transformation which outputs an MDP $\mathcal{M}'$ with the following properties:*

1. *$\mathcal{M}'$ has no 0-end-components,*
2. *there is a mapping f from schedulers of $\mathcal{M}$ to those of $\mathcal{M}'$ such that for all* proper *schedulers $\mathfrak{S}$ for $\mathcal{M}$, $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) = \mathbb{E}^{f(\mathfrak{S})}_{\mathcal{M}'}(\Diamond goal)$, and $\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) = \mathbb{V}^{f(\mathfrak{S})}_{\mathcal{M}'}(\Diamond goal)$.*
3. *there is a mapping g from schedulers of $\mathcal{M}'$ to those of $\mathcal{M}$ such that for all* proper *schedulers $\mathfrak{S}$ for $\mathcal{M}'$, $\mathbb{E}^{g(\mathfrak{S})}_{\mathcal{M}}(\Diamond goal) = \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal)$, and $\mathbb{V}^{g(\mathfrak{S})}_{\mathcal{M}}(\Diamond goal) = \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal)$.*

Proof sketch. The so-called spider construction of [**?**, Lemma 3.7] was shown to produce an MDP $\mathcal{M}'$ without 0-ECs, and which provides the mappings f and g, preserving, in particular, the probabilities of all properties ϕ_k defined by the set of paths that reach *goal* with cost at least k. The construction actually preserves all $\mathcal{E}$*-invariant* properties where $\mathcal{E}$ is the eliminated 0-EC; and it was shown that ϕ_k are such properties [1]. The equalities of all moments, and in particular, the expectation and variance immediately follow. ◀

▶ **Lemma 7.** *Let $(\mu_s)_{s\in S}$ be the solution of $(*)$ for an MDP $\mathcal{M}$. Let $\mathcal{M}'$ obtained from $\mathcal{M}$ as above. Then, $\mathcal{M}'$ has no end-components. Moreover, for all $s \in S$, all schedulers $\mathfrak{S}$ of $\mathcal{M}'$ achieve $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}[\Diamond goal] = \mu_s$.*

Proof. Towards a contradiction, assume that $\mathcal{M}'$ has an end-component $\mathcal{E}$. Let s be a state in $\mathcal{E}$, $\mathfrak{S}$ a memoryless deterministic scheduler that stays forever in $\mathcal{E}$. Consider $k > 0$, let B_k denote the set of $\mathfrak{S}$-paths of $\mathcal{E}$ of length k that start at s. Using that all actions in $\mathcal{E}$ belong to $Act^{\max}$, we can write

$$\begin{aligned}\mu_s &= \sum_{\rho\in B_k} \Pr^{\mathfrak{S}}_{\mathcal{E}}(\rho)(wgt(\rho) + \mu_{last(\rho)}) \\ &= \sum_{\rho\in B_k} \Pr^{\mathfrak{S}}_{\mathcal{E}}(\rho)wgt(\rho) + \sum_{\rho\in B_k} \Pr^{\mathfrak{S}}_{\mathcal{E}}(\rho)\mu_{last(\rho)} \\ &\leq \sum_{\rho\in B_k} \Pr^{\mathfrak{S}}_{\mathcal{E}}(\rho)wgt(\rho) + \max_{t\in S} \mu_t.\end{aligned}$$

More rigorously, these equations follow easily by induction on k using the definition of $Act^{\max}$. Notice that $\sum_{\rho\in B_k} P(\rho)wgt(\rho)$ is the expected accumulated weight over k steps. However, we know that $\mathbb{E}^{\max}_{\mathcal{E}}(\mathbb{MP}) < 0$, so the expected mean payoff inside this end component is also negative. This implies that $\lim_{k\to\infty} \sum_{\rho\in B_k} \Pr^{\mathfrak{S}}_{\mathcal{E}}(\rho)wgt(\rho) = -\infty$ which contradicts the above inequality. Thus, $\mathcal{M}'$ has no end-components. It also follows that all schedulers are proper.

We now show that all schedulers are expectation-optimal by first observing that all memoryless deterministic schedulers are expectation-optimal. In fact, if $\mathfrak{S}_{\mathsf{MD}}$ denotes such a scheduler in $\mathcal{M}'$, then $(\mathbb{E}^{\mathfrak{S}_{\mathsf{MD}}}_{\mathcal{M}',s}(\Diamond goal))_{s\in S}$ is a solution of $(*)$ and hence agrees with $(\mathbb{E}^{\max}_{\mathcal{M}',s}(\Diamond goal))_{s\in S}$. As maximal and minimal expectation of $\Diamond goal$ are obtained by memoryless deterministic schedulers [**?**], all schedulers have the same expected values from each state. ◀

[1] Note that in item S3.2 of [**?**, Lemma 3.7], $p^{\mathfrak{S}}_s = 0$ if $\mathfrak{S}$ is proper.

▶ **Lemma 8.** *Consider an MDP $\mathcal{M}$, and assume that there is a vector $(\mu_s)_{s\in S}$ of values such that all schedulers $\mathfrak{S}$ satisfy $\forall s \in S, \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}(\Diamond goal) = \mu_s$. Then, $(\mathbb{V}^{\inf}_{\mathcal{M},s}(\Diamond goal))_{s\in S}$ is the unique solution of the following equation:*

$$V_s = \begin{cases} 0 & \text{if } s = goal, \\ \min_{a\in Act(s)} \sum_{t\in S} P(s,a,t)\left((wgt(s,a)+\mu_t-\mu_s)^2 + V_t\right) & \text{otherwise.} \end{cases} \qquad (**)$$

Proof. We consider the random variables $X_0, X_1, \ldots$ which are the i-th visited state; R which is the sum of all weights until *goal* is reached (i.e. this stands for $\Diamond goal$), and R' the sum of all weights but the first one until *goal* is reached.

$$\begin{aligned}
\mathbb{V}^{\inf}_{\mathcal{M},s} &= \inf_{\mathfrak{S}} \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}[(R-\mu_s)^2] \\
&= \min_{a\in Act(s)} \inf_{\mathfrak{S}:\mathfrak{S}(s)=a} \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}[(R-\mu_s)^2] \\
&= \min_{a\in Act(s)} \inf_{\mathfrak{S}:\mathfrak{S}(s)=a} \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}[(wgt(s,a)+R'-\mu_s)^2] \\
&= \min_{a\in Act(s)} \inf_{\mathfrak{S}:\mathfrak{S}(s)=a} \sum_{t\in S} P(s,a,t)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s}[(wgt(s,a)+R'-\mu_s)^2 \mid X_1 = t] \\
&= \min_{a\in Act(s)} \inf_{\mathfrak{S}:\mathfrak{S}(s)=a} \sum_{t\in S} P(s,a,t)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},t}[(wgt(s,a)+R-\mu_s)^2] \\
&= \min_{a\in Act(s)} \sum_{t\in S} P(s,a,t)\left((wgt(s,a)-\mu_s+\mu_t)^2 + \inf_{\mathfrak{S}} \mathbb{V}^{\mathfrak{S}}_{\mathcal{M},t}[R]\right).
\end{aligned}$$

Here, we used $\mathbb{E}[(c+X)^2]) = (c+\mathbb{E}[X])^2 + \mathbb{V}[X]$ and the fact that $\mu_t = \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},t}[R]$ is constant on the last line. This shows that $\mathbb{V}^{\inf}_{\mathcal{M},s}[R]$ satisfies the given equation.

Notice that this has a form similar to $(*)$, where the weight function not only depends on the state s and chosen action a but also on the next state t, that is, $wgt'(s,a,t) = (wgt(s,a)-\mu_s+\mu_t)^2$. Alternatively, the dependence of the weight on the state t can also be modeled using intermediary states. Since $\mathcal{M}$ has no end-components, this has a unique solution by [**?**]. ◀

B Omitted proofs of Section 4

B.1 Existence of optimal deterministic schedulers

▶ **Lemma 11.** *Let $\mathcal{M} = (S, Act, P, s_{init}, wgt, goal)$ be an MDP with non-negative weights and no end components. Let $\mathfrak{S}$ and $\mathfrak{T}$ be two schedulers for $\mathcal{M}$. Let $p \in (0,1)$. The scheduler $\mathfrak{R} \overset{def}{=} p\cdot\mathfrak{S} \oplus (1-p)\cdot\mathfrak{T}$ satisfies*

$$\mathbb{V}^{\mathfrak{R}}_{\mathcal{M}}(\Diamond goal) = p\cdot\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + (1-p)\cdot\mathbb{V}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal) + p\cdot(1-p)\cdot(\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal))^2.$$

Proof. As the expected values of $\Diamond goal$ and $\Diamond goal^2$ depend linearly on the expected frequencies of the state-weight pairs $(goal, w)$ with $w \in \mathbb{N}$, we know that

$$\mathbb{E}^{\mathfrak{R}}_{\mathcal{M}}(\Diamond goal) = p\cdot\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + (1-p)\cdot\mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)$$

and

$$\mathbb{E}^{\mathfrak{R}}_{\mathcal{M}}(\Diamond goal^2) = p\cdot\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal^2) + (1-p)\cdot\mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal^2).$$

The claim follows by basic arithmetic using $\mathbb{V}^{\mathfrak{R}}_{\mathcal{M}}(\Diamond goal) = \mathbb{E}^{\mathfrak{R}}_{\mathcal{M}}(\Diamond goal^2) - \mathbb{E}^{\mathfrak{R}}_{\mathcal{M}}(\Diamond goal)^2$:

$$
\begin{aligned}
&\mathbb{V}^{\mathfrak{R}}_{\mathcal{M}}(\Diamond goal)\\
&= p \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal^2) + (1-p) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal^2) - (p \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + (1-p) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal))^2\\
&= p \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal^2) + (1-p) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal^2)\\
&\quad - \Big(p^2 \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)^2 + 2p(1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal) + (1-p)^2 \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)^2\Big)\\
&\quad \underbrace{+(p^2-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)^2 - (p^2-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)^2}_{=0}\\
&\quad \underbrace{+((1-p)^2-(1-p))\mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)^2 - ((1-p)^2-(1-p))\mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)^2}_{=0}\\
&= p \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal^2) - p \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)^2 + (1-p) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal^2) - (1-p) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)^2\\
&\quad - 2p(1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)\\
&\quad - (p^2-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)^2 - ((1-p)^2-(1-p))\mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)^2\\
&= p \cdot \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + (1-p) \cdot \mathbb{V}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)\\
&\quad - 2p(1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)\\
&\quad - p \cdot (p-1) \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)^2 - ((1-p)-1) \cdot (1-p) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)^2\\
&= p \cdot \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + (1-p) \cdot \mathbb{V}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)\\
&\quad + p \cdot (1-p) \cdot \Big(\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)^2 - 2\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal) + \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)^2\Big)\\
&= p \cdot \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + (1-p) \cdot \mathbb{V}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal) + p \cdot (1-p) \cdot (\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal))^2. \qquad \blacktriangleleft
\end{aligned}
$$

▶ **Lemma 13** (Continuity of VPE). *Let $\mathcal{M}$ and $\lambda > 0$ be as above. The variance-penalized expectation as a function from weight-based schedulers to $\mathbb{R}$ is (uniformly) continuous in the following sense: Given $\varepsilon > 0$, there is a natural number N_ε such that for all weight-based schedulers $\mathfrak{S}$ and $\mathfrak{T}$ that agree on all state-weight pairs (s, w) with $w \leq N_\varepsilon$, we have*

$$\left|\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} - \mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}}\right| < \varepsilon.$$

Proof. As $\mathcal{M}$ has no end components, the values $E_1 \stackrel{\text{def}}{=} \max_{s \in S} \mathbb{E}^{\max}_{\mathcal{M},s}(\Diamond goal)$ and $E_2 \stackrel{\text{def}}{=} \max_{s \in S} \mathbb{E}^{\max}_{\mathcal{M},s}(\Diamond goal^2)$ are finite (an explicit bound on the latter is also given in Section 4.3). Furthermore, let W be the maximal weight occurring in $\mathcal{M}$, let n be the number of states of $\mathcal{M}$, and let δ be the minimal non-zero probability occurring in $\mathcal{M}$. From any state, *goal* is reached within n steps with probability at least δ^n. Let $p \stackrel{\text{def}}{=} (1 - \delta^n)$. So, for any $k \in \mathbb{N}$, the probability that a weight of more than $n \cdot W \cdot k$ is accumulated under any scheduler is bounded by p^n.

We now show that $\mathbb{E}^{\cdot}_{\mathcal{M}}(\Diamond goal)$ and $\mathbb{E}^{\cdot}_{\mathcal{M}}(\Diamond goal^2)$ as functions from schedulers to real numbers are continuous in the sense of the lemma. The claim then follows immediately.

For the former, let $\varepsilon > 0$ be given. Let k_ε be such that $p^{k_\varepsilon} \cdot E_1 < \epsilon$. Now, let $\mathfrak{S}$ and $\mathfrak{T}$ be two weight-based schedulers that agree on all state-weight pairs with weight at most $N_\varepsilon \stackrel{\text{def}}{=} k_\varepsilon \cdot n \cdot W$. If a weight of more than N_ε cannot be accumulated under $\mathfrak{S}$ and $\mathfrak{T}$, there is nothing to show. Otherwise,

$$
\begin{aligned}
&|\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal)|\\
&\leq \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal > N_\varepsilon) \cdot |\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal | \Diamond goal > N_\varepsilon) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M}}(\Diamond goal | \Diamond goal > N_\varepsilon)|\\
&\leq p^{k_\varepsilon} \cdot E_1 < \varepsilon.
\end{aligned}
$$

For the function $\mathbb{E}^{\cdot}_{\mathcal{M}}(\Diamond goal^2)$, the claim follows analogously using the bound E_2. $\blacktriangleleft$

▶ **Lemma 14.** *Let $\mathcal{M}$, $\mathfrak{S}$, $\mathfrak{S}_\alpha$, $\mathfrak{S}_\beta$, and q be as above. There is a value $p \in (0,1)$ such that the expected frequencies of all state-weight pairs are the same under $\mathfrak{S}$ and $p \cdot \mathfrak{S}_\alpha \oplus (1-p) \cdot \mathfrak{S}_\beta$.*

In the proof, we use the following notation: Let $\mathfrak{S}$ be a weight-based scheduler, (s,w) a state-weight pair, and π a path with $last(\pi) = s$ and $wgt(\pi) = w$. Then, $\mathfrak{S}{\uparrow}(s,w)$ denotes the scheduler $\mathfrak{S}{\uparrow}\pi$. As $\mathfrak{S}$ is weight-based, this definition does not depend on the choice of the path π.

Proof. To obtain a candidate, for the value p, we first consider the expected frequency $\vartheta^{\mathfrak{S}}_{s,w}$, $\vartheta^{\mathfrak{S}_\alpha}_{s,w}$, and $\vartheta^{\mathfrak{S}_\beta}_{s,w}$, of the state-weight pair (s,w) under $\mathfrak{S}$, $\mathfrak{S}_\alpha$, and $\mathfrak{S}_\beta$, respectively. Under all three schedulers, the probability $t > 0$ that the state-weight pair (s,w) is reached is the same. The expected frequencies now depend on the probability that a run returns to (s,w) after leaving (s,w). Let r_α be the probability that $\mathfrak{S}_\alpha{\uparrow}(s,w)$ returns to state s without accumulating additional weight, i.e., the probability that a run under $\mathfrak{S}_\alpha{\uparrow}(s,w)$ starting in s has a prefix π of length > 1 with $last(\pi) = s$ and $wgt(\pi) = 0$. Let r_β be defined analogously. The probability that $\mathfrak{S}$ returns from (s,w) to this same state-weight pair is then $q \cdot r_\alpha + (1-q) \cdot r_\beta$. Note that r_α and r_β are less than 1 as all end components have negative maximal expected mean-payoff in $\mathcal{M}$. We obtain the following expected frequencies of (s,w):

$$\vartheta^{\mathfrak{S}_\alpha}_{s,w} = t\frac{1}{1-r_\alpha}, \qquad \vartheta^{\mathfrak{S}_\beta}_{s,w} = t\frac{1}{1-r_\beta},$$
$$\vartheta^{\mathfrak{S}}_{s,w} = t\frac{1}{1-qr_\alpha-(1-q)r_\beta}.$$

The value p has to satisfy $p \cdot \vartheta^{\mathfrak{S}_\alpha}_{s,w} + (1-p) \cdot \vartheta^{\mathfrak{S}_\beta}_{s,w} = \vartheta^{\mathfrak{S}}_{s,w}$ which is equivalent to

$$\frac{1}{1-qr_\alpha-(1-q)r_\beta} = p \cdot \frac{1}{1-r_\alpha} + (1-p) \cdot \frac{1}{1-r_\beta}. \tag{$\diamond$}$$

The value $\frac{1}{1-qr_\alpha-(1-q)r_\beta}$ lies between $\frac{1}{1-r_\alpha}$ and $\frac{1}{1-r_\beta}$. If $r_\alpha = r_\beta$, any value $p \in [0,1]$ satisfies the equation. In this case, we also have $\vartheta^{\mathfrak{S}}_{s,w} = \vartheta^{\mathfrak{S}_\alpha}_{s,w} = \vartheta^{\mathfrak{S}_\beta}_{s,w}$. Otherwise, there is a unique value $p \in (0,1)$ satsifying ($\diamond$).

Next, we have to check sure that also actions α and β are chosen with the correct expected frequency in (s,w) under $p \cdot \mathfrak{S}_\alpha \oplus (1-p) \cdot \mathfrak{S}_\beta$. Note that $\vartheta^{\mathfrak{S}}_{s,w,\alpha} = q \cdot \vartheta^{\mathfrak{S}}_{s,w}$ and that $\vartheta^{\mathfrak{S}_\alpha}_{s,w,\alpha} = \vartheta^{\mathfrak{S}_\alpha}_{s,w}$ and $\vartheta^{\mathfrak{S}_\alpha}_{s,w,\beta} = 0$, and analogously for $\mathfrak{S}_\beta$.

If $r_\alpha = r_\beta$, we can simply choose $p = q$. If $r_\alpha \neq r_\beta$, we have to check that

$$q \cdot t \cdot \frac{1}{1-qr_\alpha-(1-q)r_\beta} = p \cdot t \cdot \frac{1}{1-r_\alpha}.$$

This equation follows from ($\diamond$) by basic arithmetic for $r_\alpha \neq r_\beta$ by multiplying out equation ($\diamond$) and dividing by $(r_\alpha - r_\beta)$.

So, indeed there exists a unique value p such that the expected frequencies of (s,w) and of the expected number of times α and β, respectively, are chosen at (s,w) agree under $\mathfrak{S}$ and $p \cdot \mathfrak{S}_\alpha \oplus (1-p) \cdot \mathfrak{S}_\beta$. As schedulers $\mathfrak{S}$, $\mathfrak{S}_\alpha$, and $\mathfrak{S}_\beta$ behave identically at all state-weight pairs different to (s,w), the expected frequencies of all other state weight pairs under $\mathfrak{S}$ and $p \cdot \mathfrak{S}_\alpha \oplus (1-p) \cdot \mathfrak{S}_\beta$ are the same as well. ◀

▶ **Theorem 15** (Deterministic schedulers approximate optimal VPE)**.** *Let $\mathcal{M}$ be an MDP with non-negative weights and without end components and let $\lambda > 0$. For each scheduler $\mathfrak{S}$, there is a deterministic weight-based scheduler $\mathfrak{T}$ with*

$$\mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}.$$

Proof. Let $\mathfrak{S}$ be a scheduler that is not deterministic. By Corollary 3, we can assume w.l.o.g. that $\mathfrak{S}$ is weight-based. So, we view $\mathfrak{S}$ as a function from $S \times \mathbb{N}$ to probability distributions over actions. There is now a reachable state-weight pair (s, w) at which $\mathfrak{S}$ schedules an action α with probability $0 < q < 1$. In a first step, we show that we can modify the scheduler such that it chooses α with probability 0 or 1 without decreasing the variance-penalized expectation.

First, we suppose that the scheduler only chooses one other action β at (s, w) with positive probability to keep notation simpler. We explain how to treat the case with multiple actions afterwards. So, the scheduler $\mathfrak{S}$ chooses α at (s, w) with probability q and β with probability $1 - q$. Let $\mathfrak{S}_\alpha$ be the scheduler that agrees with $\mathfrak{S}$ on all state-weight pairs except for (s, w) where it chooses α with probability 1. Let $\mathfrak{S}_\beta$ be defined analogously. By Lemma 14, there is a $p \in (0, 1)$ such that $\mathfrak{S}$ and $p \cdot \mathfrak{S}_\alpha + (1 - p) \cdot \mathfrak{S}_\beta$ lead to the same expected frequency of all state-weight pairs. Using Lemma 11, we obtain:

$$\begin{aligned}
\mathbb{VPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}} &= \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}}(\Diamond\!\!\!\!-\, goal) - \lambda \mathbb{V}_{\mathcal{M}}^{\mathfrak{S}}(\Diamond\!\!\!\!-\, goal) \\
&= p \cdot (\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_\alpha}(\Diamond\!\!\!\!-\, goal) - \lambda \mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_\alpha}(\Diamond\!\!\!\!-\, goal)) + (1 - p) \cdot (\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_\beta}(\Diamond\!\!\!\!-\, goal) - \lambda \mathbb{V}_{\mathcal{M}}^{\mathfrak{S}_\beta}(\Diamond\!\!\!\!-\, goal)) \\
&\quad - \lambda \cdot p \cdot (1 - p)(\mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_\alpha}(\Diamond\!\!\!\!-\, goal) - \mathbb{E}_{\mathcal{M}}^{\mathfrak{S}_\beta}(\Diamond\!\!\!\!-\, goal))^2 \\
&\leq p \cdot \mathbb{VPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}_\alpha} + (1 - p) \cdot \mathbb{VPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}_\beta}.
\end{aligned}$$

We conclude that $\mathbb{VPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}_\alpha} \geq \mathbb{VPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}}$ or $\mathbb{VPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}_\beta} \geq \mathbb{VPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}}$. So, indeed a scheduler that behaves like $\mathfrak{S}$ except at (s, w) where it chooses α with probability 0 or 1 achieves at least the same variance-penalized expectation.

We assumed for simplicity that the scheduler $\mathfrak{S}$ chooses only two actions with positive probability. In general, the scheduler $\mathfrak{S}$ might choose action α with probability $0 < q < 1$ and further actions $\beta_1, \ldots, \beta_\ell \in \mathit{Act}(\mathit{last}(\pi)) \setminus \{\alpha\}$ for some ℓ with positive probabilities $(1-q) \cdot \Delta(\beta_1), \ldots, (1-q) \cdot \Delta(\beta_\ell)$ where Δ is a probability distribution over $\mathit{Act}(\mathit{last}(\pi)) \setminus \{\alpha\}$. The argument above still works if we use β as an abbreviation for choosing actions $\beta_1, \ldots, \beta_\ell \in \mathit{Act}(\mathit{last}(\pi)) \setminus \{\alpha\}$ according to the probability distribution Δ. Our proof then shows that at least one of the schedulers choosing α with probability 1 or the remaining actions according to Δ does not decrease the variance-penalized expectation. In the latter case if $\ell > 1$, the argument can successively be repeated by letting action β_1 take the role of α. After at most ℓ changes to the scheduler, we find a scheduler that chooses one of the actions α and $\beta_1, \ldots, \beta_\ell$ with probability 1 at (s, w) and otherwise behaves like $\mathfrak{S}$.

To obtain a deterministic scheduler, we enumerate all (countably many) state-weight pairs $(s_0, w_0), (s_1, w_1), \ldots$ Let $\mathfrak{S}_0 \stackrel{\text{def}}{=} \mathfrak{S}$. Once scheduler $\mathfrak{S}_i$ is defined, we let $\mathfrak{S}_{i+1}$ be a scheduler that makes a deterministic choice at the state-weight pair (s_i, w_i) and otherwise behaves like $\mathfrak{S}_i$ and that satisfies

$$\mathbb{VPE}[\lambda]_{\mathcal{M}}^{\mathfrak{S}_{i+1}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}_i}.$$

In the limit, we obtain a well-defined deterministic weight-based scheduler $\mathfrak{T}$: The choice of $\mathfrak{T}$ at any state-weight pair (s_i, w_i) is given by $\mathfrak{S}_{i+1}(s_i, w_i)$ which is equal to $\mathfrak{S}_j(s_i, w_i)$ for all $j > i$.

We claim that

$$\mathbb{VPE}[\lambda]_{\mathcal{M}}^{\mathfrak{T}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}.$$

To see this, let $\varepsilon > 0$ be arbitrary and let N_ε be the natural number as in Lemma 13. There are only finitely many state-weight pairs (s, w) with $w \leq N_\varepsilon$. Let k_ε be a natural number

such that all state-weight pairs (s, w) with $w \leq N_\varepsilon$ occur before $(s_{k_\varepsilon}, w_{k_\varepsilon})$ in our enumeration. Then $\mathfrak{S}_{k_\varepsilon}$ and $\mathfrak{T}$ agree on all all state-weight pairs (s, w) with $w \leq N_\varepsilon$. By Lemma 13, we conclude that

$$\mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}_{k_\varepsilon}}_{\mathcal{M}} - \varepsilon \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} - \varepsilon.$$

As ε was arbitrary, this implies that $\mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}$. ◀

B.2 Hardness of the threshold problem

▶ **Theorem 18.** *Given an MDP $\mathcal{M}$ with non-negative weights and two rationals $\lambda, \vartheta > 0$, deciding whether $\mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}} \geq \vartheta$ is EXPTIME-hard. Furthermore, for acyclic MDPs $\mathcal{M}$, the problem is PSPACE-hard.*

Proof. We reduce from the following problem which is shown to be EXPTIME-hard in [**?**]: Given an MDP $\mathcal{M}$ and a natural number $T > 0$ such that *goal* is reached in $\mathcal{M}$ almost surely under all schedulers, decide whether there is a scheduler $\mathfrak{S}$ such that $\Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = T) = 1$. Observe that $\mathcal{M}$ does not contain any end-components.

Given such an MDP $\mathcal{M}$ and a value $T > 0$, we define the MDP $\mathcal{M}'$ by adding a fresh initial state ι from which a unique action (with weight 0) leads, with probability $1/2$, to the initial state of $\mathcal{M}$; while with probability $1/2$ it leads to a fresh state ι'. From ι', a unique action with weight T leads to *goal*.

Further, we let $\vartheta \stackrel{\text{def}}{=} T$. Let W be the largest weight in the MDP, and n the number of states, and $p_{\min}$ the smallest probability in $\mathcal{M}$. Let $\varepsilon < p^n_{\min}$.

We establish an upper bound on the expectation of $\Diamond goal$ as follows. Since all schedulers are proper, under all schedulers, every n steps, the process reaches *goal* with probability at least ε. We have, for all $\mathfrak{S}$,

$$\begin{aligned}
\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal) &\leq W \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}(\text{steps until } goal) \\
&= W \sum_{k \geq 0} k \Pr^{\mathfrak{S}}_{\mathcal{M}'}(\text{steps until } goal = k) \\
&\leq W \sum_{k \geq 0} k(1-\varepsilon)^{\lfloor k/n \rfloor} \\
&\leq W \sum_{l \geq 0} n(l+n)(1-\varepsilon)^{l} \\
&\leq nW\left(\frac{n}{\varepsilon} + \frac{1-\varepsilon}{\varepsilon^2}\right) \qquad (1)
\end{aligned}$$

where we used the fact that under $\mathfrak{S}$, from any state, there is a path of length n to *goal*, with probability $\geq \varepsilon$. Let us denote $f(n, W, \varepsilon) = nW(\frac{n}{\varepsilon} + \frac{1-\varepsilon}{\varepsilon^2})$ which we can assume to be larger than 1. Define

$$\lambda \stackrel{\text{def}}{=} 18 f(n, W, \varepsilon)$$

Notice that this number can be computed in polynomial time.

We claim that there exists a scheduler $\mathfrak{S}$ with $\Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = T) = 1$ iff there exists $\mathfrak{S}'$ with $\mathbb{VPE}[\lambda]^{\mathfrak{S}'}_{\mathcal{M}'} \geq \vartheta$.

A scheduler $\mathfrak{S}$ with $\Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = T) = 1$ viewed as a scheduler for $\mathcal{M}'$ satisfies $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal) = T$ and $\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal) = 0$. Hence, $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}'} = T = \vartheta$.

Let $\mathfrak{S}$ be a deterministic scheduler for $\mathcal{M}'$ that maximizes $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}'}$, and assume that $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}'} \geq \vartheta$. Such a scheduler exists by the previous Theorem 15 and can be viewed as a scheduler for $\mathcal{M}$ as well.

Let us write $p = \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = T)$, and $\mu^{\mathfrak{S}} = \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)$.

By assuming $p < 1$, we will reach a contradiction.

We first show that $|T - \mu^{\mathfrak{S}}| \leq \frac{1}{3}$. In fact, observe that

$$\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal) = \frac{1}{2}(T - \mu^{\mathfrak{S}})^2 + \frac{1}{2}\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}((\Diamond goal - \mu^{\mathfrak{S}})^2), \tag{2}$$

where the first term corresponds to the path $\iota, \iota', goal$, and the second term to all other paths.

As $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal) \leq f(n, W, \varepsilon)$, we obtain $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}'} \leq f(n, W, \varepsilon) - \lambda \cdot \frac{1}{2}(T - \mu^{\mathfrak{S}})^2$. So if $|T - \mu^{\mathfrak{S}}| > \frac{1}{3}$, then $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}'} \leq f(n, W, \varepsilon) - \frac{1}{18}\lambda < 0$, which is a contradiction.

Now, let us write

$$\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal) = \frac{1}{2}T + \frac{1}{2}pT + \frac{1}{2}\sum_{k \neq T} k \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = k),$$

where the first term corresponds to the path via $\iota, \iota', goal$; the second term is the set of paths that enter $\mathcal{M}$ and achieve T, and the last term is the contribution of all other paths.

For the variance, let us focus on the right term of (2).

$$\begin{aligned}
\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}((\Diamond goal - \mu^{\mathfrak{S}})^2) &= \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = T)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}((\Diamond goal - \mu^{\mathfrak{S}})^2 \mid \Diamond goal = T) \\
&\quad + \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal \neq T)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}((\Diamond goal - \mu^{\mathfrak{S}})^2 \mid \Diamond goal \neq T) \\
&= p(T - \mu^{\mathfrak{S}})^2 + (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}((\Diamond - \mu^{\mathfrak{S}})^2 \mid \Diamond goal \neq T) \\
&= p(T - \mu^{\mathfrak{S}})^2 + \sum_{l \neq T}(l - \mu^{\mathfrak{S}})^2 \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = l).
\end{aligned} \tag{3}$$

Here, the right hand side of the last line is obtained as follows.

$$\begin{aligned}
\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}((\Diamond goal - \mu^{\mathfrak{S}})^2 \mid \Diamond goal \neq T) &= \textstyle\sum_{k \geq 0} k \Pr^{\mathfrak{S}}_{\mathcal{M}}((\Diamond goal - \mu^{\mathfrak{S}})^2 = k \mid \Diamond goal \neq T) \\
&= \textstyle\sum_{l \geq 0}(l - \mu^{\mathfrak{S}})^2 \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = l \mid \Diamond goal \neq T) \\
&= \textstyle\sum_{l \neq T}(l - \mu^{\mathfrak{S}})^2 \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = l \mid \Diamond goal \neq T) \\
&= \frac{1}{1-p}\textstyle\sum_{l \neq T}(l - \mu^{\mathfrak{S}})^2 \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = l).
\end{aligned}$$

So, we rewrite (2) as follows.

$$\mathbb{V}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal) = \frac{1+p}{2}(T - \mu^{\mathfrak{S}})^2 + \frac{1}{2}\sum_{k \neq T}(k - \mu^{\mathfrak{S}})^2 \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = k).$$

Let us compute the VPE:

$$\begin{aligned}
\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}'} &= \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal) - \lambda \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}'}(\Diamond goal) \\
&= \frac{1+p}{2}T - \frac{\lambda(1+p)}{2}(T - \mu^{\mathfrak{S}})^2 + \frac{1}{2}\sum_{k \neq T}(k - \lambda(k - \mu^{\mathfrak{S}})^2) \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = k)
\end{aligned}$$

To reach a contradiction, it suffices to show that the above is less than T, which is equivalent to

$$-\lambda(1+p)(T - \mu^{\mathfrak{S}})^2 + \sum_{k \neq T}(k - \lambda(k - \mu^{\mathfrak{S}})^2) \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = k) < (1-p)T. \tag{4}$$

The function $k \mapsto (k - \lambda(k-\mu^{\mathfrak{S}})^2)$ is increasing until $k = \mu^{\mathfrak{S}} + \frac{1}{2\lambda}$ and decreasing afterwards; so its maximum at integers $k \neq T$ is reached either at $T-1$ or $T+1$ since $|T - \mu^{\mathfrak{S}}| \leq \frac{1}{3}$. For $k = T-1$, we have

$$\begin{aligned}
&-\lambda(1+p)(T-\mu^{\mathfrak{S}})^2 + (T-1-\lambda(T-1-\mu^{\mathfrak{S}})^2) \sum_{k \neq T} \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = k) < (1-p)T \\
&\Leftrightarrow -\lambda(1+p)(T-\mu^{\mathfrak{S}})^2 + (1-p)T - (1+\lambda(T-1-\mu^{\mathfrak{S}})^2)(1-p) < (1-p)T \\
&\Leftrightarrow -\lambda(1+p)(T-\mu^{\mathfrak{S}})^2 - (1+\lambda(T-1-\mu^{\mathfrak{S}})^2)(1-p) < 0.
\end{aligned}$$

For $k = T+1$, we have

$$\begin{aligned}
&-\lambda(1+p)(T-\mu^{\mathfrak{S}})^2 + (T+1-\lambda(T+1-\mu^{\mathfrak{S}})^2) \sum_{k \neq T} \Pr^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal = k) < (1-p)T \\
&\Leftrightarrow -\lambda(1+p)(T-\mu^{\mathfrak{S}})^2 + (1-p)(T+1-\lambda(T+1-\mu^{\mathfrak{S}})^2) < (1-p)T \\
&\Leftrightarrow -\lambda(1+p)(T-\mu^{\mathfrak{S}})^2 + (1-p)(1-\lambda(T+1-\mu^{\mathfrak{S}})^2) < 0.
\end{aligned}$$

Here, $(T+1-\mu^{\mathfrak{S}})^2 \geq \frac{4}{9}$ since $|T-\mu^{\mathfrak{S}}| \leq \frac{1}{3}$, so $1-\lambda(T+1-\mu^{\mathfrak{S}})^2 < 0$ since $\lambda > \frac{9}{4}$. This establishes (4), yielding a contradiction.

Addressing the second claim: For acyclic MDPs, the problem we reduce from is shown to be PSPACE-hard in [**?**]. As our construction preserves acyclicity, PSPACE-hardness for the threshold problem for VPE in acyclic MDPs follows as above. ◀

B.3 Saturation point

▶ **Theorem 20** (Saturation point). *Let $\mathcal{M}$, $\lambda > 0$ and K be as above. Let $\mathfrak{S}$ be a scheduler with $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$. Then, for each finite $\mathfrak{S}$-path π with $wgt(\pi) \geq K$, the residual scheduler $\mathfrak{S}{\uparrow}\pi$ satisfies*

$$\mathbb{E}^{\mathfrak{S}\uparrow\pi}_{\mathcal{M},last(\pi)}(\Diamond goal) = \mathbb{E}^{\min}_{\mathcal{M},last(\pi)}(\Diamond goal).$$

Proof. Let $\mathfrak{S}$ be a scheduler with $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$. Suppose there is a $\mathfrak{S}$-path π' with $wgt(\pi') \geq K$ such that

$$\mathbb{E}^{\mathfrak{S}\uparrow\pi'}_{\mathcal{M},last(\pi')}(\Diamond goal) > \mathbb{E}^{\min}_{\mathcal{M},last(\pi')}(\Diamond goal).$$

Then, there must be an $\mathfrak{S}$-path π that extends π' such that $\mathfrak{S}$ chooses an action $\alpha \notin Act^{\min}(last(\pi))$ with positive probability. Let us write $s \stackrel{\text{def}}{=} last(\pi)$ and define

$$p \stackrel{\text{def}}{=} P(\pi) \cdot \mathfrak{S}(\pi)(\alpha).$$

So, p is the probability that π is seen under $\mathfrak{S}$ and that $\mathfrak{S}$ chooses α afterwards. As $wgt(\pi) \geq K \geq B_{1/2}$, we conclude that $p \leq 1/2$.

We claim that we can construct a scheduler $\mathfrak{S}'$ with $\mathbb{VPE}[\lambda]^{\mathfrak{S}'}_{\mathcal{M}} > \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$. Let $\mathfrak{Min}$ be a memoryless deterministic scheduler with

$$\mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) = \mathbb{E}^{\min}_{\mathcal{M},s}(\Diamond goal).$$

We define the scheduler $\mathfrak{S}'$ to behave like $\mathfrak{S}$ unless π is seen and $\mathfrak{S}$ chooses α after π. In this case, $\mathfrak{S}'$ switches to the behavior of $\mathfrak{Min}$ instead.

To compare the schedulers $\mathfrak{S}$ and $\mathfrak{S}'$, let us define $\mathfrak{T}$ to be the residual scheduler of $\mathfrak{S}$ after π when $\mathfrak{S}$ chooses α. I.e., extending the notation for residual schedulers, we define $\mathfrak{T} \stackrel{\text{def}}{=} \mathfrak{S}{\uparrow}(\pi\alpha)$ where

$$\mathfrak{S}{\uparrow}(\pi\alpha)(s)(\alpha) = 1,$$

i.e., on the path only consisting of state s, the scheduler chooses α with probability 1, and for all finite paths ρ starting with s followed by α,

$$\mathfrak{S}{\uparrow}(\pi\alpha)(\rho) = \mathfrak{S}(\pi \circ \rho).$$

So, the schedulers $\mathfrak{S}$ and $\mathfrak{S}'$ agree on all paths except for the extensions of π in case $\mathfrak{S}$ chooses α after π. In this case, $\mathfrak{S}$ behaves like $\mathfrak{T}$ and $\mathfrak{S}'$ behaves like $\mathfrak{Min}$.

Let now X be the event that $\mathfrak{S}$ does not choose α after π. In particular, X contains all paths that do not have π as a prefix. Conditioned on the event X, $\mathfrak{S}$ and $\mathfrak{S}'$ behave identically. Furthermore, the probability of X is $1-p$ under both schedulers. This allows us to split the expected value of $\Diamond\!\!\!\!-goal$ under both schedulers conditioning on X and its complement $\bar{X}$. We get

$$\begin{aligned}\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal) &= (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal \mid X) + p \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal \mid \bar{X}) \\ &= (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal \mid X) + p \cdot (wgt(\pi) + \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond\!\!\!\!-goal)) \end{aligned} \tag{5}$$

and

$$\begin{aligned}\mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal) &= (1-p)\mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal \mid X) + p \cdot \mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal \mid \bar{X}) \\ &= (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal \mid X) + p \cdot (wgt(\pi) + \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond\!\!\!\!-goal)) \end{aligned} \tag{6}$$

where we use that $\mathfrak{S}$ and $\mathfrak{S}'$ behave identically on X in the last equality. Let us denote the term that does not depend on $\mathfrak{T}$ or $\mathfrak{Min}$ and occurs in both equations (5) and (6) by

$$C_1 \stackrel{\text{def}}{=} (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal \mid X).$$

Note that $C_1 \leq \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal) \leq U_1$.

As a direct consequence of (5) and (6), we also get

$$\begin{aligned}&(\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal))^2 \\ &= C_1^2 + 2C_1 \cdot p \cdot (wgt(\pi) + \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond\!\!\!\!-goal)) + p^2 \cdot (wgt(\pi) + \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond\!\!\!\!-goal))^2 \end{aligned} \tag{7}$$

and

$$\begin{aligned}&(\mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal))^2 \\ &= C_1^2 + 2C_1 \cdot p \cdot (wgt(\pi) + \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond\!\!\!\!-goal)) + p^2 \cdot (wgt(\pi) + \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond\!\!\!\!-goal))^2. \end{aligned} \tag{8}$$

Applying the same reasoning as above to the random variable $\Diamond\!\!\!\!-goal^2$, we obtain

$$\begin{aligned}\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal^2) &= (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal^2 \mid X) + p \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal^2 \mid \bar{X}) \\ &= (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal^2 \mid X) + p \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}((wgt(\pi) + \Diamond\!\!\!\!-goal)^2) \end{aligned} \tag{9}$$

and

$$\begin{aligned}\mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal^2) &= (1-p)\mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal^2 \mid X) + p \cdot \mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal^2 \mid \bar{X}) \\ &= (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond\!\!\!\!-goal^2 \mid X) + p \cdot \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}((wgt(\pi) + \Diamond\!\!\!\!-goal)^2). \end{aligned} \tag{10}$$

Again, we abbreviate the first term by

$$C_2 \stackrel{\text{def}}{=} (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond goal^2 \mid X)$$

and note that $C_2 \leq \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond goal^2) \leq U_2$. Further extending equations (9) and (10) and using the linearity of the expected value, we obtain

$$\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond goal^2) = C_2 + p \cdot (wgt(\pi)^2 + 2 \cdot wgt(\pi) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal) + \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal^2)) \tag{11}$$

and

$$\mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond goal^2) = C_2 + p \cdot (wgt(\pi)^2 + 2 \cdot wgt(\pi) \cdot \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) + \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal^2)). \tag{12}$$

The key to showing our claim now lies in the fact that $\mathfrak{T}$ starts in state s by choosing actions $\alpha \notin Act^{\min}(s)$. So,

$$\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal) \geq \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) + \delta. \tag{13}$$

Putting everything together, we will now show that indeed $\mathbb{VPE}[\lambda]^{\mathfrak{S}'}_{\mathcal{M}} > \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$:

$$\begin{aligned}
&\mathbb{VPE}[\lambda]^{\mathfrak{S}'}_{\mathcal{M}} - \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} \\
=\ &\mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond goal) - \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond goal) \\
&- \lambda \cdot (\mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond goal^2) - \mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond goal^2)) \\
&+ \lambda \cdot ((\mathbb{E}^{\mathfrak{S}'}_{\mathcal{M},s_{init}}(\Diamond goal))^2 - (\mathbb{E}^{\mathfrak{S}}_{\mathcal{M},s_{init}}(\Diamond goal))^2) \\
=\ &p \cdot (\mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal)) \\
&- \lambda \cdot p \cdot \big(2 \cdot wgt(\pi) \cdot \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) + \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal^2) \\
&\qquad\qquad - 2 \cdot wgt(\pi) \cdot \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal^2)\big) \\
&+ \lambda \cdot \big(2C_1 \cdot p \cdot (\mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal)) \\
&\qquad + p^2 \cdot ((wgt(\pi) + \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal))^2 - (wgt(\pi) + \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal))^2)\big) \\
=\ &p \cdot (\mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal)) \\
&- \lambda \cdot p \cdot \big(\mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal^2) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal^2)\big) \\
&+ 2\lambda \cdot p \cdot wgt(\pi) \cdot \big(\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal)\big) \\
&+ 2\lambda \cdot p \cdot (\mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal)) \\
&+ 2\lambda \cdot p^2 \cdot wgt(\pi) \cdot \big(\mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal)\big) \\
&+ \lambda \cdot p^2 \cdot \big((\mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal))^2 - (\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal))^2\big) \\
\geq\ &- p \cdot U_1 - \lambda \cdot p \cdot U_2 - 2\lambda \cdot p \cdot U_1 - \lambda \cdot p^2 \cdot U_1^2 \\
&+ 2\lambda \cdot p \cdot wgt(\pi) \cdot \big(\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal)\big) \\
&- 2\lambda \cdot p^2 \cdot wgt(\pi) \cdot \big(\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal)\big) \\
\geq\ &- p \cdot U_1 - \lambda \cdot p \cdot U_2 - 2\lambda \cdot p \cdot U_1 - \lambda \cdot p^2 \cdot U_1^2 \\
&+ \lambda \cdot p \cdot wgt(\pi) \cdot \big(\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal)\big).
\end{aligned} \tag{14}$$

In the last inequality, we use that $p \leq 1/2$. To show that the right hand side is greater than 0, we use that $\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal) \geq \delta$ and first obtain that

$$\begin{aligned}
&- p \cdot U_1 - \lambda \cdot p \cdot U_2 - 2\lambda \cdot p \cdot U_1 - \lambda \cdot p^2 \cdot U_1^2 \\
&+ \lambda \cdot p \cdot wgt(\pi) \cdot \big(\mathbb{E}^{\mathfrak{T}}_{\mathcal{M},s}(\Diamond goal) - \mathbb{E}^{\mathfrak{Min}}_{\mathcal{M},s}(\Diamond goal)\big) \\
\geq\ &- p \cdot U_1 - \lambda \cdot p \cdot U_2 - 2\lambda \cdot p \cdot U_1 - \lambda \cdot p^2 \cdot U_1^2 + \lambda \cdot p \cdot wgt(\pi) \cdot \delta.
\end{aligned}$$

Now,

$$-p \cdot U_1 - \lambda \cdot p \cdot U_2 - 2\lambda \cdot p \cdot U_1 - \lambda \cdot p^2 \cdot U_1^2 + \lambda \cdot p \cdot wgt(\pi) \cdot \delta > 0$$

is equivalent to

$$\lambda \cdot wgt(\pi) \cdot \delta > U_1 + \lambda U_2 + 2\lambda U_1 + \lambda \cdot p \cdot U_1^2.$$

As $wgt(\pi) \geq K$ and $p \leq 1/2$, it is sufficient to show that

$$\lambda \cdot K \cdot \delta > U_1 + \lambda U_2 + 2\lambda U_1 + \lambda \cdot 1/2 \cdot U_1^2.$$

This now, however follows directly from the definition of K as

$$\lambda \cdot K \cdot \delta \geq U_1 + \lambda U_2 + 2\lambda U_1 + \lambda \cdot 1/2 \cdot U_1^2 + \lambda \cdot \delta.$$

This shows that the scheduler $\mathfrak{S}$ does not maximize $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$. This finishes the proof that the defined value K, which is computable in polynomial time as argued before, is as desired.

Note that the estimations (14) and hence the whole proof works analogously for the variant $\sup_{\mathfrak{S}} -\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) - \lambda \cdot \mathbb{V}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal)$ as claimed in Remark 10. ◀

▶ **Theorem 21.** *Let $\mathcal{M}$, $\lambda > 0$, K, and $\mathfrak{V}$ be as above. Let $\mathfrak{S}$ be a deterministic scheduler with $\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$. Let $\mathfrak{T}$ be the scheduler that agrees with $\mathfrak{S}$ on all paths π with weight less than K and that chooses actions according to the memoryless deterministic scheduler $\mathfrak{V}$ after paths π' with $wgt(\pi') \geq K$. This scheduler $\mathfrak{T}$ satisfies $\mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}} = \mathbb{VPE}[\lambda]^{\max}_{\mathcal{M}}$, too.*

Proof. We will show that after each $\mathfrak{S}$-path π with $wgt(\pi) \geq K$ behaving according to $\mathfrak{V}$ instead of $\mathfrak{S}{\upharpoonright}\pi$ does not decrease the variance-penalized expectation. Applying this reasoning successively to all $\mathfrak{S}$-paths that reach a weight level of at least K in their last step leads us to the desired scheduler $\mathfrak{T}$.

So, let π be a $\mathfrak{S}$-path with $wgt(\pi) \geq K$ ending in a state s. By Theorem 20, we know that

$$\mathbb{E}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}(\Diamond goal) = \mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal)$$

and by the definition of $\mathfrak{V}$,

$$\mathbb{V}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}(\Diamond goal) \geq \mathbb{V}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal).$$

As in previous proofs of this section, we express the variance-penalized expectation of $\mathfrak{S}$ by conditioning on the event Π that π is a prefix of a run and its complement $\neg\Pi$. Note that by the definition of K and the fact that π is a $\mathfrak{S}$-path, the probability $p \stackrel{\text{def}}{=} \Pr^{\mathfrak{S}}_{\mathcal{M}}(\pi)$ lies strictly between 0 and 1. We obtain

$$\begin{aligned}
\mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}} &= \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + \lambda \cdot (\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal))^2 - \lambda \cdot \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal^2) \\
&= \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + \lambda \cdot (\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal))^2 \\
&\quad - \lambda \cdot \big((1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal^2 \mid \neg\Pi) + p\mathbb{E}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}((wgt(\pi) + \Diamond goal)^2)\big) \\
&= \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) + \lambda \cdot (\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal))^2 \\
&\quad - \lambda \cdot (1-p)\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal^2 \mid \neg\Pi) \\
&\quad - \lambda \cdot p \cdot \big(wgt(\pi)^2 + 2wgt(\pi) \cdot \mathbb{E}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}(\Diamond goal) + \mathbb{E}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}(\Diamond goal^2)\big). \qquad (\times)
\end{aligned}$$

Let now $\mathfrak{S}'$ be the scheduler that behaves like $\mathfrak{S}$ after all paths that do not have π as a prefix and that behaves like $\mathfrak{V}$ as soon as π has been seen. By the observations on $\mathfrak{S}{\upharpoonright}\pi$

and $\mathfrak{V}$ above, we know that $\mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\Diamond goal) = \mathbb{E}^{\mathfrak{S}'}_{\mathcal{M}}(\Diamond goal)$. Furthermore, conditioned on $\neg\Pi$, the two schedulers behave identically. With the same calculations for $\mathfrak{S}'$ as in $(\times)$ using that $\mathfrak{S}'{\upharpoonright}\pi = \mathfrak{V}$, we get that

$$\mathbb{VPE}[\lambda]^{\mathfrak{S}'}_{\mathcal{M}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}$$

if and only if

$$-\lambda \cdot p \cdot \mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal^2) \geq -\lambda \cdot p \cdot \mathbb{E}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}(\Diamond goal^2)$$

This, however, follows directly from

$$\begin{aligned}
\mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal^2) &= \mathbb{V}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal) + (\mathbb{E}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal))^2 \\
&= \mathbb{V}^{\mathfrak{V}}_{\mathcal{M},s}(\Diamond goal) + (\mathbb{E}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}(\Diamond goal))^2 \\
&\leq \mathbb{V}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}(\Diamond goal) + (\mathbb{E}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}(\Diamond goal))^2 \\
&= \mathbb{E}^{\mathfrak{S}{\upharpoonright}\pi}_{\mathcal{M},s}(\Diamond goal^2).
\end{aligned}$$

Enumerating all $\mathfrak{S}$-paths $\pi_1, \pi_2, \dots$ that reach a weight of at least K in their last step, we can now recursively define a sequence of schedulers $\mathfrak{S}_i$ starting from $\mathfrak{S}_1 = \mathfrak{S}$ with non-decreasing variance-penalized expectation such that $\mathfrak{S}_i$ behaves like $\mathfrak{V}$ after all paths π_j with $j < i$. By continuity arguments as before, we obtain a scheduler $\mathfrak{T}$ in the limit that behaves like $\mathfrak{V}$ as soon as a weight of at least K has been accumulated and that satisfies

$$\mathbb{VPE}[\lambda]^{\mathfrak{T}}_{\mathcal{M}} \geq \mathbb{VPE}[\lambda]^{\mathfrak{S}}_{\mathcal{M}}.$$

By the optimality of $\mathfrak{S}$, the new scheduler $\mathfrak{T}$ maximizes the variance-penalized expectation, too. ◀