The variance-penalized stochastic shortest path problem

Jakob Piribauer[1], Ocan Sankur[2], and Christel Baier[1]

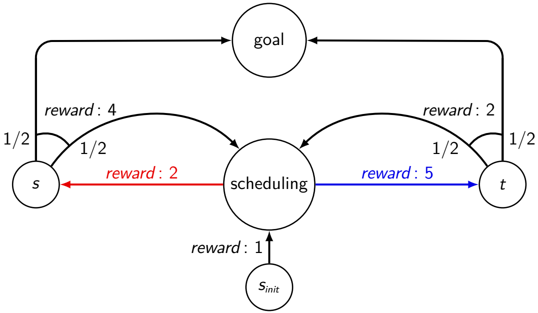
[1] TU Dresden
[2] Univ Rennes, Inria, CNRS, IRISA

October 2022
Reachability Problems

Stochastic shortest path problem

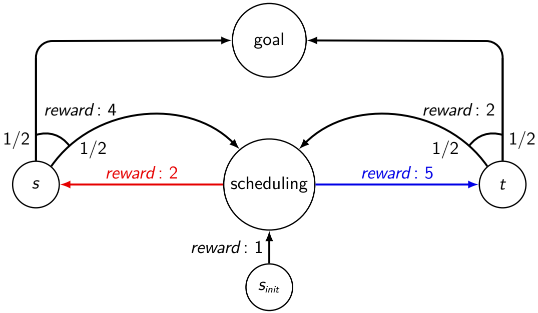
Stochastic shortest path problem

What is the maximal possible reward in expectation?



Stochastic shortest path problem

What is the maximal possible reward in expectation?

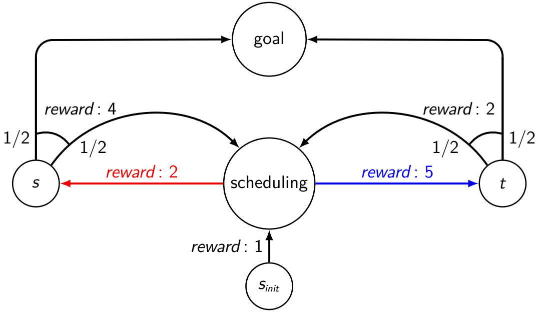


Markov decision process (MDP)

Scheduler:
resolves non-deterministic choices

Stochastic shortest path problem

What is the maximal possible reward in expectation?



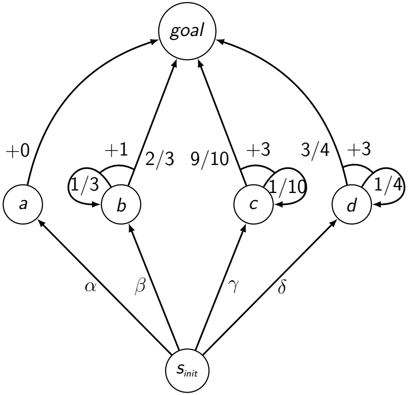
Markov decision process (MDP)

Scheduler:
resolves non-deterministic choices

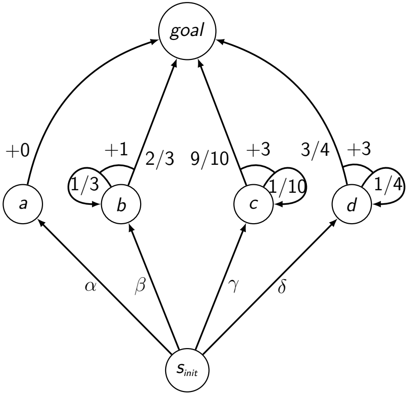
Classical problem:

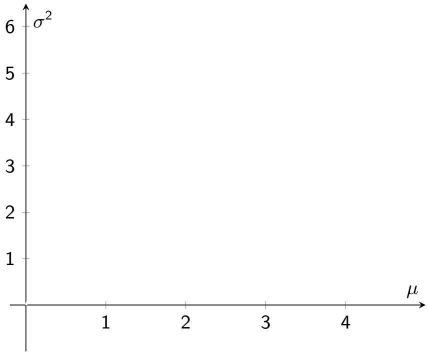
Compute $\mathbb{E}^{\max}_{\mathcal{M}}(\text{acc. reward}) \stackrel{\text{def}}{=} \sup_{\mathfrak{S}} \mathbb{E}^{\mathfrak{S}}_{\mathcal{M}}(\text{acc. reward})$
where $\mathfrak{S}$ ranges over schedulers reaching goal almost surely.

Tradeoff between expectation and variance

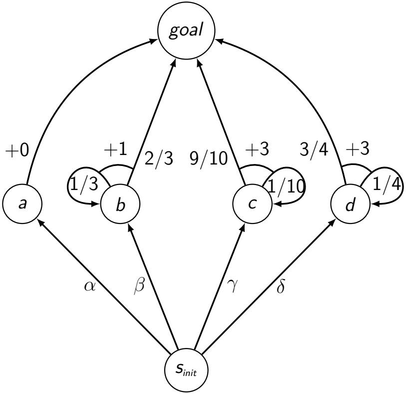


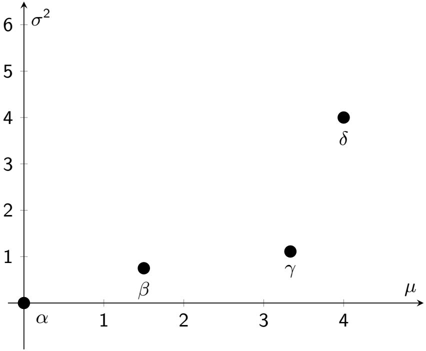
Tradeoff between expectation and variance



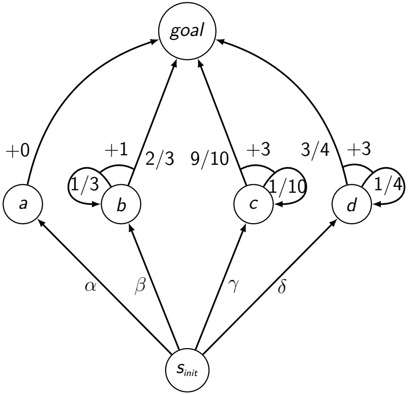


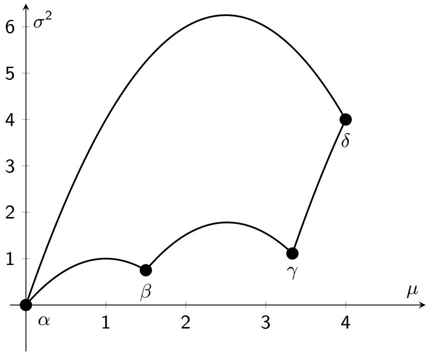
Tradeoff between expectation and variance



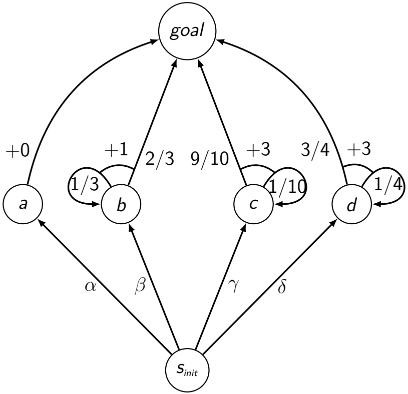


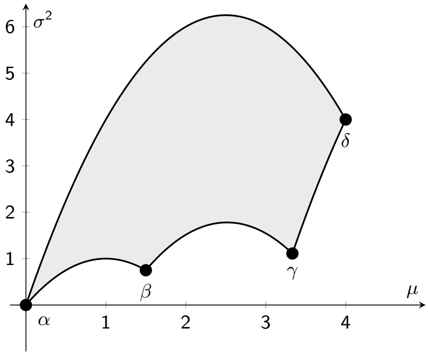
Tradeoff between expectation and variance



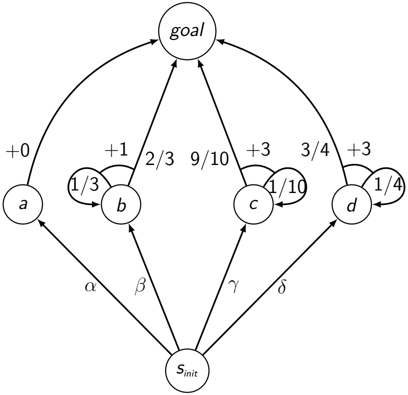


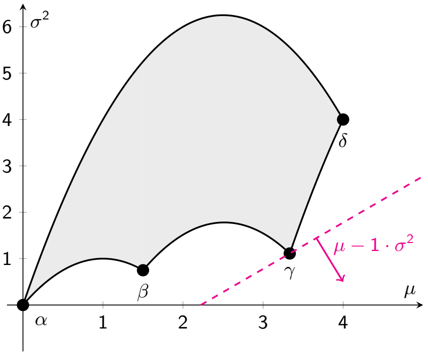
Tradeoff between expectation and variance





Tradeoff between expectation and variance





Variance-penalized expectation (VPE): $\mu - \lambda \cdot \sigma^2$

Motivation to study VPE

- Established objective in MDPs
  (finite horizon[1], discounted expected rewards[2])

[1] See, e.g., Collins (1997)
[2] See, e.g., Filar, Kallenberg, Lee (1989)

Motivation to study VPE

- Established objective in MDPs
  (finite horizon[1], discounted expected rewards[2])

- Markowitz portfolio optimization

[1] See, e.g., Collins (1997)
[2] See, e.g., Filar, Kallenberg, Lee (1989)

Motivation to study VPE

- Established objective in MDPs
  (finite horizon[1], discounted expected rewards[2])
- Markowitz portfolio optimization
- *Weighted factor method* a common approach in multi-objective optimization to obtain a subset of the Pareto-optimal points.[3]

[1] See, e.g., Collins (1997)
[2] See, e.g., Filar, Kallenberg, Lee (1989)
[3] See, e.g., White (1982), Chatterjee, Majumdar, Henzinger (2006).

Motivation to study VPE

- Established objective in MDPs
  (finite horizon[1], discounted expected rewards[2])

- Markowitz portfolio optimization

- *Weighted factor method* a common approach in multi-objective optimization to obtain a subset of the Pareto-optimal points.[3]

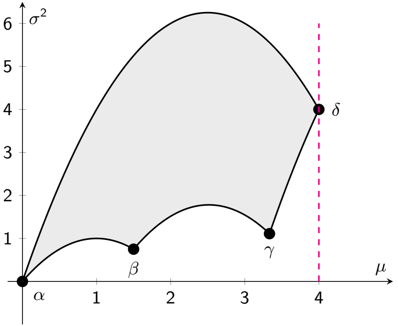
- As a stepping stone towards further problems.

[1] See, e.g., Collins (1997)
[2] See, e.g., Filar, Kallenberg, Lee (1989)
[3] See, e.g., White (1982), Chatterjee, Majumdar, Henzinger (2006).

Our results

Theorem

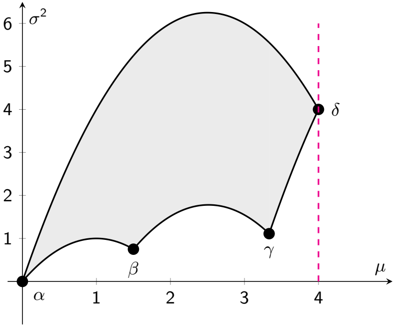
In an MDP with arbitrary (integer) weights, a memoryless, deterministic, and variance-minimal scheduler among all expectation-optimal schedulers can be computed in polynomial time.



Our results

Theorem

In an MDP with arbitrary (integer) weights, a memoryless, deterministic, and variance-minimal scheduler among all expectation-optimal schedulers can be computed in polynomial time.



If expected weight is known (and independent of the history) from each state, minimal variance can be computed via a linear program.

Illustration of the difficulties of maximizing VPE

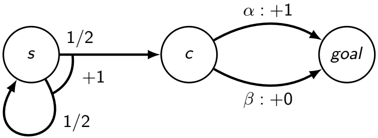
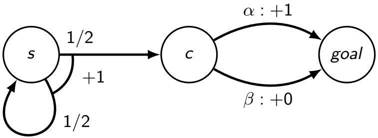


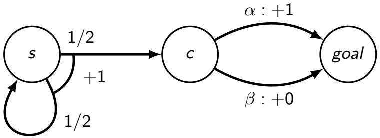
Illustration of the difficulties of maximizing VPE



X: accumulated weight

Maximize $\mathbb{E}^{\mathfrak{S}}(X) - \lambda \mathbb{V}^{\mathfrak{S}}(X) = \mathbb{E}^{\mathfrak{S}}(X) - \lambda\Big(\mathbb{E}^{\mathfrak{S}}(X^2) - (\mathbb{E}^{\mathfrak{S}}(X))^2\Big)$.

Illustration of the difficulties of maximizing VPE

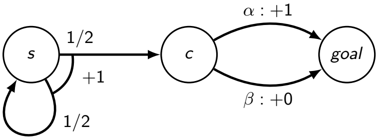


X: accumulated weight

Maximize $\mathbb{E}^{\mathfrak{S}}(X) - \lambda \mathbb{V}^{\mathfrak{S}}(X) = \mathbb{E}^{\mathfrak{S}}(X) - \lambda\big(\mathbb{E}^{\mathfrak{S}}(X^2) - (\mathbb{E}^{\mathfrak{S}}(X))^2\big)$.

Let $\mathfrak{B}$ and $\mathfrak{A}$ be two schedulers that behave identically except:

Illustration of the difficulties of maximizing VPE

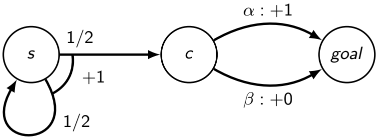


X: accumulated weight

Maximize $\mathbb{E}^{\mathfrak{S}}(X) - \lambda \mathbb{V}^{\mathfrak{S}}(X) = \mathbb{E}^{\mathfrak{S}}(X) - \lambda\Big(\mathbb{E}^{\mathfrak{S}}(X^2) - (\mathbb{E}^{\mathfrak{S}}(X))^2\Big)$.

Let $\mathfrak{B}$ and $\mathfrak{A}$ be two schedulers that behave identically except:
$\mathfrak{B}$ chooses β if weight k has been accumulated; $\mathfrak{A}$ chooses α instead.
(This happens with prob $p = \frac{1}{2^{k-1}}$.)

Illustration of the difficulties of maximizing VPE



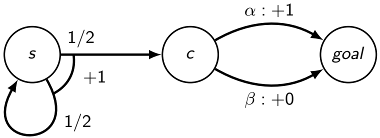
X: accumulated weight

Maximize $\mathbb{E}^{\mathfrak{S}}(X) - \lambda \mathbb{V}^{\mathfrak{S}}(X) = \mathbb{E}^{\mathfrak{S}}(X) - \lambda\big(\mathbb{E}^{\mathfrak{S}}(X^2) - (\mathbb{E}^{\mathfrak{S}}(X))^2\big)$.

Let $\mathfrak{B}$ and $\mathfrak{A}$ be two schedulers that behave identically except:
$\mathfrak{B}$ chooses β if weight k has been accumulated; $\mathfrak{A}$ chooses α instead.
(This happens with prob $p = \frac{1}{2^{k-1}}$.)

$$
\begin{aligned}
\mathbb{E}^{\mathfrak{A}}(X) - \mathbb{E}^{\mathfrak{B}}(X) &= p. \\
(\mathbb{E}^{\mathfrak{A}}(X))^2 - (\mathbb{E}^{\mathfrak{B}}(X))^2 &= 2p\mathbb{E}^{\mathfrak{B}}(X) + p^2. \\
\mathbb{E}^{\mathfrak{A}}(X^2) - \mathbb{E}^{\mathfrak{B}}(X^2) &= p((k+1)^2 - k^2) = p(2k+1).
\end{aligned}
$$

Illustration of the difficulties of maximizing VPE



X: accumulated weight

Maximize $\mathbb{E}^{\mathfrak{S}}(X) - \lambda \mathbb{V}^{\mathfrak{S}}(X) = \mathbb{E}^{\mathfrak{S}}(X) - \lambda\big(\mathbb{E}^{\mathfrak{S}}(X^2) - (\mathbb{E}^{\mathfrak{S}}(X))^2\big)$.

Let $\mathfrak{B}$ and $\mathfrak{A}$ be two schedulers that behave identically except:
$\mathfrak{B}$ chooses β if weight k has been accumulated; $\mathfrak{A}$ chooses α instead.
(This happens with prob $p = \frac{1}{2^{k-1}}$.)

$$\mathbb{E}^{\mathfrak{A}}(X) - \mathbb{E}^{\mathfrak{B}}(X) = p.$$
$$(\mathbb{E}^{\mathfrak{A}}(X))^2 - (\mathbb{E}^{\mathfrak{B}}(X))^2 = 2p\mathbb{E}^{\mathfrak{B}}(X) + p^2.$$
$$\mathbb{E}^{\mathfrak{A}}(X^2) - \mathbb{E}^{\mathfrak{B}}(X^2) = p((k+1)^2 - k^2) = p(2k+1).$$

$$VPE(\mathfrak{A}) - VPE(\mathfrak{B}) = p\big(1 + \lambda(2\mathbb{E}^{\mathfrak{B}}(X) + p) \quad -\lambda(2k+1)\big).$$

Saturation point

Lemma

*Given an MDP $\mathcal{M}$ with non-negative weights and a rational penalty factor λ, we can compute a bound K in polynomial time such that any VPE-optimal scheduler has to **minimize** the expected accumulated weight as soon as a weight of at least K has been accumulated.*

Saturation point

Lemma

Given an MDP $\mathcal{M}$ with non-negative weights and a rational penalty factor λ, we can compute a bound K in polynomial time such that any VPE-optimal scheduler has to ***minimize*** *the expected accumulated weight as soon as a weight of at least K has been accumulated.*

Above the *saturation point* K:

Fix a memoryless deterministic scheduler minimizing the variance among all expectation-minimal scheduler.

Saturation point

Lemma

Given an MDP $\mathcal{M}$ with non-negative weights and a rational penalty factor λ, we can compute a bound K in polynomial time such that any VPE-optimal scheduler has to ***minimize*** *the expected accumulated weight as soon as a weight of at least K has been accumulated.*

Above the *saturation point* K:

Fix a memoryless deterministic scheduler minimizing the variance among all expectation-minimal scheduler.

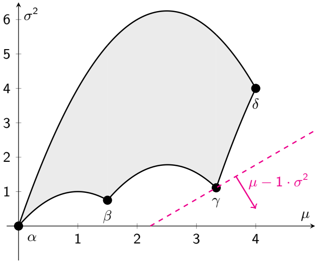
Computable in polynomial time.

Our results

Our results

Theorem

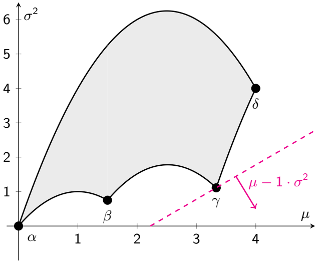
In an MDP with non-negative weights, the maximal VPE (for a given penalty factor λ) can be computed in exponential space.
Optimal schedulers can be chosen to be deterministic finite-memory schedulers and can be computed in exponential space as well.



Our results

Theorem

In an MDP with non-negative weights, the maximal VPE (for a given penalty factor λ) can be computed in exponential space.
Optimal schedulers can be chosen to be deterministic finite-memory schedulers and can be computed in exponential space as well.



Theorem

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

Outlook

- Decidability of the existence of a scheduler with expectation $\geq \eta$ and variance $\leq \nu$?

Outlook

- Decidability of the existence of a scheduler with expectation $\geq \eta$ and variance $\leq \nu$?

- VPE in MDPs with arbitrary weights (Positivity-hard?)

Outlook

- Decidability of the existence of a scheduler with expectation $\geq \eta$ and variance $\leq \nu$?
- VPE in MDPs with arbitrary weights (Positivity-hard?)
- Investigation of further risk and deviation measures

References

- Collins. "Finite-horizon variance penalised Markov decision processes." Operations-Research-Spektrum 19.1 (1997): 35-39.
- Filar, Kallenberg, Lee. "Variance-penalized Markov decision processes." Mathematics of Operations Research 14.1 (1989): 147-161.
- White. "Multi-objective infinite-horizon discounted Markov decision processes." Journal of mathematical analysis and applications 89.2 (1982): 639-647.
- Chatterjee, Majumdar, Henzinger. "Markov decision processes with multiple objectives." Annual symposium on theoretical aspects of computer science. Springer, Berlin, Heidelberg, 2006.