Mean Payoff in Markov Decision Processes

Jan Křetínský

Technical University of Munich, Germany

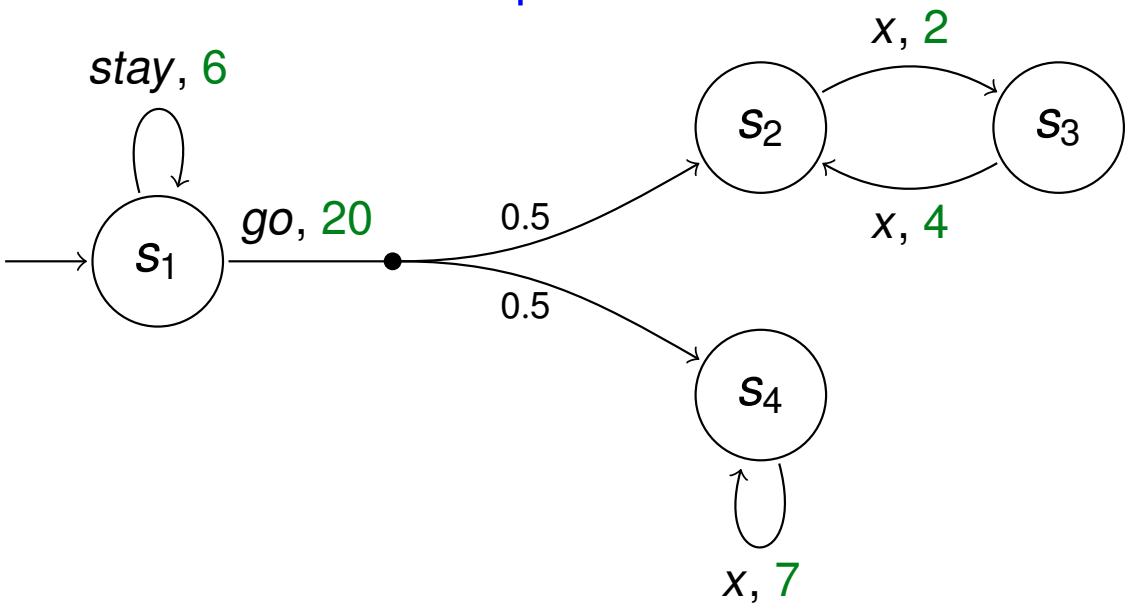
joint work with P. Ashok, T. Meggendorfer (TUM),
K. Chatterjee, P. Daca (IST Austria)

Highlights

London, UK
September 14, 2017

Problem

Model: Markov decision process



Objective: Mean payoff / long-run average reward

- $MP(20\ 2\ 4\ 2\ 4\ \cdots) = 3$
- $\mathbb{E}_{\sigma_{go}}[MP] = \frac{3+7}{2} = 5$
- $\max_\sigma \mathbb{E}_\sigma[MP] = 6$

State of the art and our contributions

Solution techniques

- Linear programming (LP)
  - precise
  - polynomial time, but practically slow
- Strategy iteration (SI)
  - precise
  - monotonically improving
  - can utilize domain knowledge
  - exponential time and **costly evaluation**, but quite ok
- Value iteration (VI)
  - convergent
  - **no general stopping criterion / current error bound**
  - exponential, but fast

State of the art and our contributions

Solution techniques

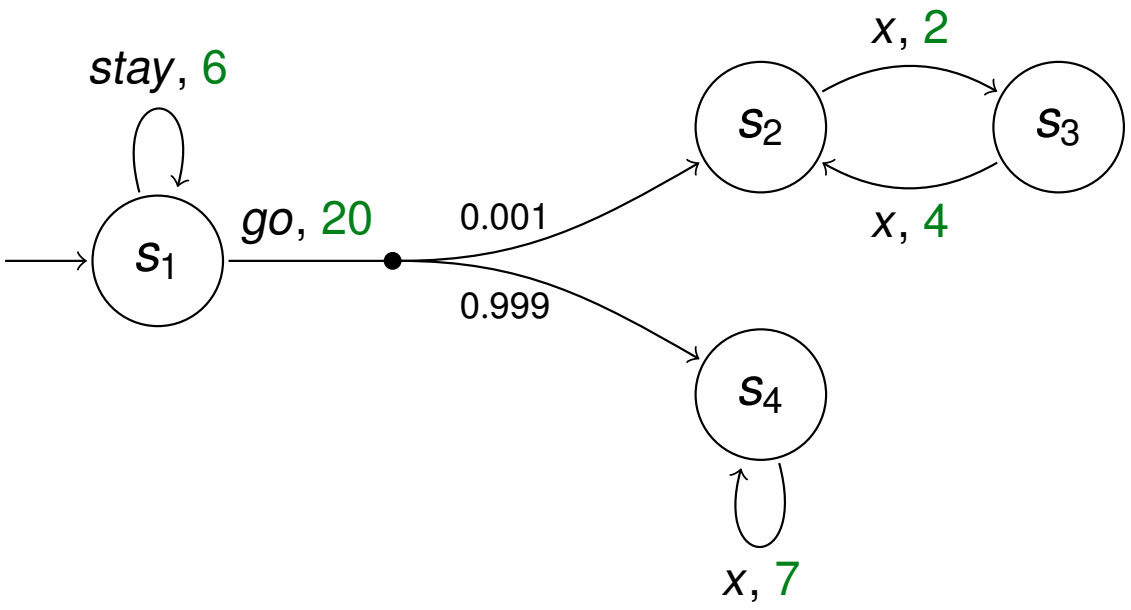
- Linear programming (LP)
  - precise
  - polynomial time, but practically slow
- Strategy iteration (SI)
  - precise
  - monotonically improving
  - can utilize domain knowledge
  - exponential time and **costly evaluation**, but quite ok
- Value iteration (VI)
  - convergent
  - **no general stopping criterion / current error bound**
  - exponential, but fast

Our contribution: improving SI [ATVA'17] and VI [CAV'17]

- stopping criterion / current error bound for VI
  - refuting old conjecture
- speeding up VI and strategy evaluation in SI
  - *simulations and machine learning* help to focus

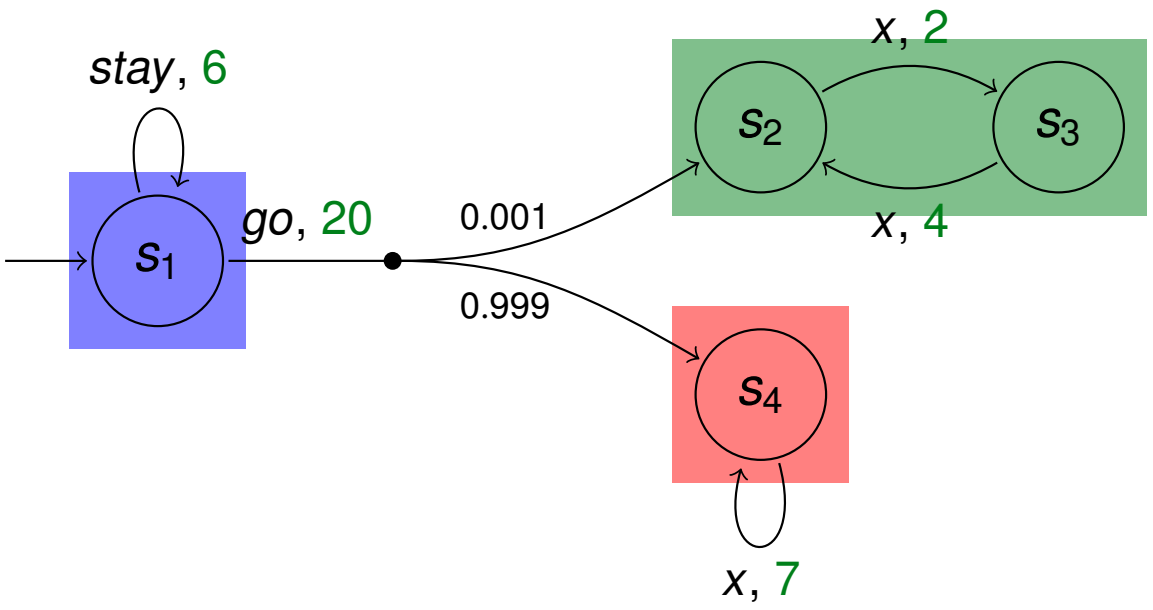
Main idea 1/2

Note that $\max_\sigma \mathbb{E}_\sigma[MP] = \max_\sigma \sum_{MEC\ M} \overbrace{\mathbb{P}_\sigma[\Diamond M]}^{\text{reachability}} \cdot \overbrace{MP(M^\sigma)}^{\text{MP on EC}}$



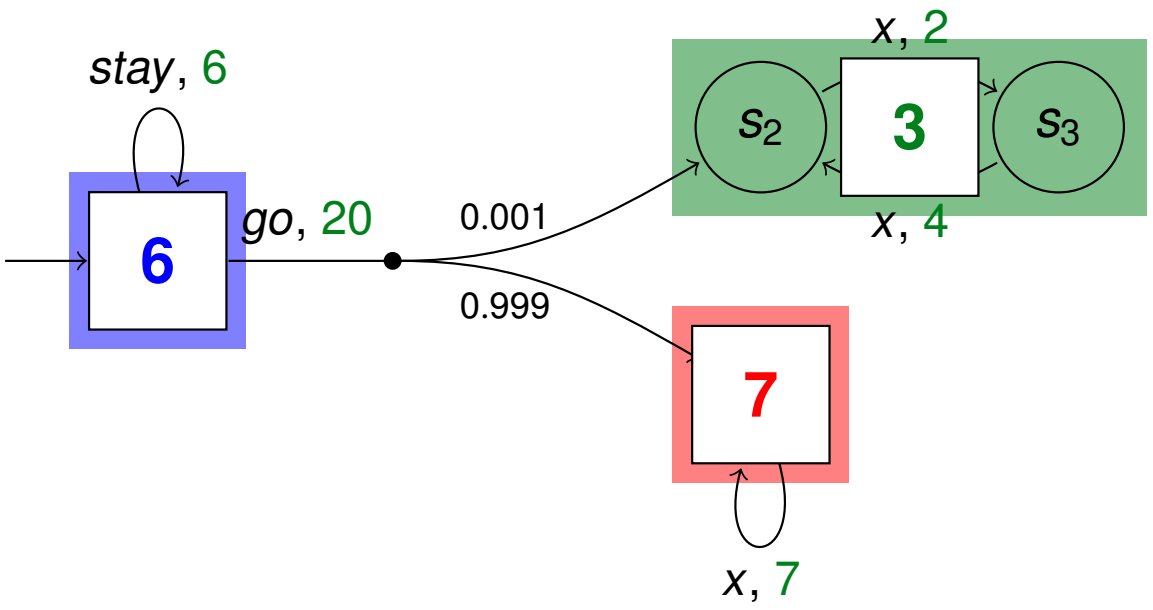
Main idea 1/2

Note that $\max_\sigma \mathbb{E}_\sigma[MP] = \max_\sigma \sum_{MEC\ M} \overbrace{\mathbb{P}_\sigma[\Diamond M]}^{\text{reachability}} \cdot \overbrace{MP(M^\sigma)}^{\text{MP on EC}}$



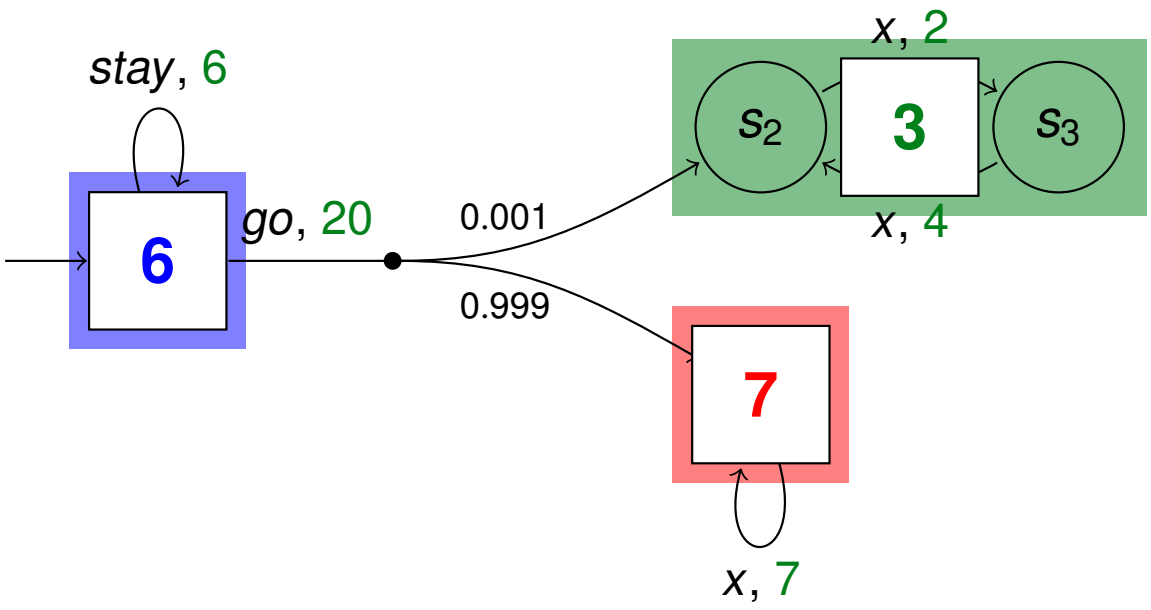
Main idea 1/2

Note that $\max_\sigma \mathbb{E}_\sigma[MP] = \max_\sigma \sum_{MEC\ M} \overbrace{\mathbb{P}_\sigma[\Diamond M]}^{\text{reachability}} \cdot \overbrace{MP(M^\sigma)}^{\text{MP on EC}}$



Main idea 1/2

Note that $\max_\sigma \mathbb{E}_\sigma[MP] = \max_\sigma \sum_{MEC\ M} \overbrace{\mathbb{P}_\sigma[\Diamond M]}^{\text{reachability}} \cdot \overbrace{MP(M^\sigma)}^{\text{MP on EC}}$



Desiderata:

- ignore states with low reachability probability/approx. error/profit
- focus on **highly reachable, uncertain and profitable** states

Main idea 2/2

Solution:

1. keep both lower and upper bounds
   - collapse end components (graph transformation, on the fly)
   - $\Longrightarrow$ error bound, uncertainty
   - $\Longrightarrow$ treat only highly uncertain states

Main idea 2/2

Solution:

1. keep both lower and upper bounds
   - collapse end components (graph transformation, on the fly)
   - $\Longrightarrow$ error bound, uncertainty
   - $\Longrightarrow$ treat only highly uncertain states
2. simulation guided + reinforcement learning
   - transition probabilities $\Longrightarrow$ treat only highly reachable states
   - pick currently best actions $\Longrightarrow$ treat only highly profitable states

Experimental results

Table: Timeout: 10m, Memory: 8GB, VI precision: 10^{-6}

Model	LP	SI	VI	SI*	VI*
cs_nfail3 (184, 38)	2	17	–	4	4
cs_nfail4 (960, 176)	5	1129	–	5	5
sensors1 (462, 132)	3	–	–	4	5
sensors2 (7860, 4001)	101	–	–	13	15
mer3 (15622, 9451)	–	–	–	16	15
mer4 (119305, 71952)	–	–	–	42	64
zeroconf(4730203, ?)	–	–	–	–	10

Conclusion

Summary

- stopping criterion / error bounds for VI
- improving VI by orders of magnitude
- SI made on par with VI
- *tools:*
  - on-the-fly graph transformations (collapsing end components)
  - simulations
  - machine learning
- for details see [CAV'17, ATVA'17]

Conclusion

Summary

- stopping criterion / error bounds for VI
- improving VI by orders of magnitude
- SI made on par with VI
- *tools:*
  - on-the-fly graph transformations (collapsing end components)
  - simulations
  - machine learning
- for details see [CAV'17, ATVA'17]

Open questions and future work

- ignore less important parts of end components
- stochastic games
- initialization for SI using machine learning

Conclusion

Summary

- stopping criterion / error bounds for VI
- improving VI by orders of magnitude
- SI made on par with VI
- *tools:*
  - on-the-fly graph transformations (collapsing end components)
  - simulations
  - machine learning
- for details see [CAV'17, ATVA'17]

Open questions and future work

- ignore less important parts of end components
- stochastic games
- initialization for SI using machine learning

Thank you