Register at: essai.si

ESSAI & ACAI 2023
LJUBLJANA, SLOVENIA

# MODEL UNCERTAINTY IN SEQUENTIAL DECISION MAKING


**DAVID PARKER**
*University of Oxford*



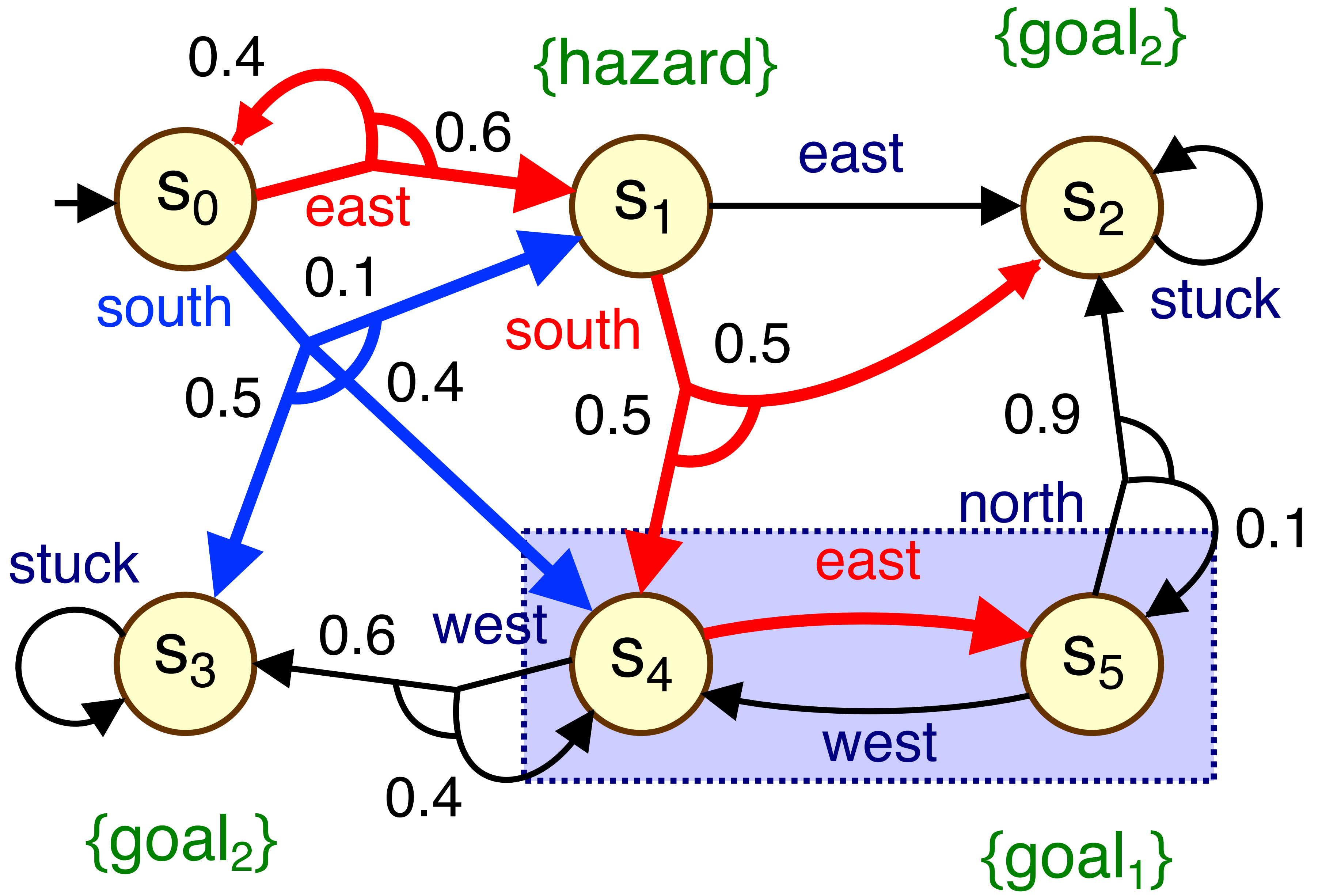
**BRUNO LACERDA**
*University of Oxford*



**NICK HAWES**
*University of Oxford*

# Recap

- Introduction
  - aleatoric vs. epistemic uncertainty
- Markov decision processes (MDPs)
  - sequential decision making under uncertainty
  - policies and objectives
    - MaxProb, SSP, finite-horizon, temporal logic
  - solving MDPs (optimal policy generation)
    - linear programming (PTIME)
    - or dynamic programming (value iteration)



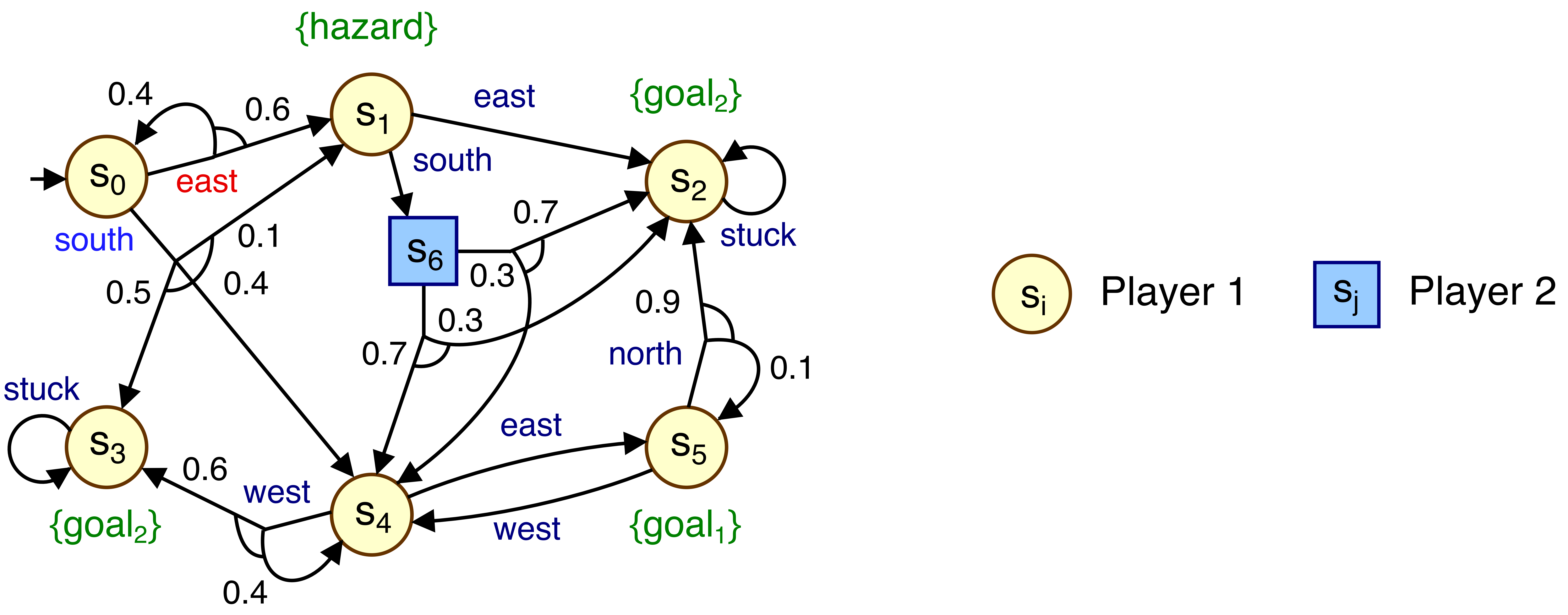
# Course contents

- ~~Markov decision processes (MDPs) and stochastic games~~
    - ~~MDPs: key concepts and algorithms~~
    - stochastic games: adding adversarial aspects
- Uncertain MDPs
    - MDPs + epistemic uncertainty, robust control, robust dynamic programming, interval MDPs, uncertainty set representation, challenges, tools
- Sampling-based uncertain MDPs
    - removing the transition independence assumption
- Bayes-adaptive MDPs
    - maintaining a distribution over the possible models

# Stochastic games

# Running example

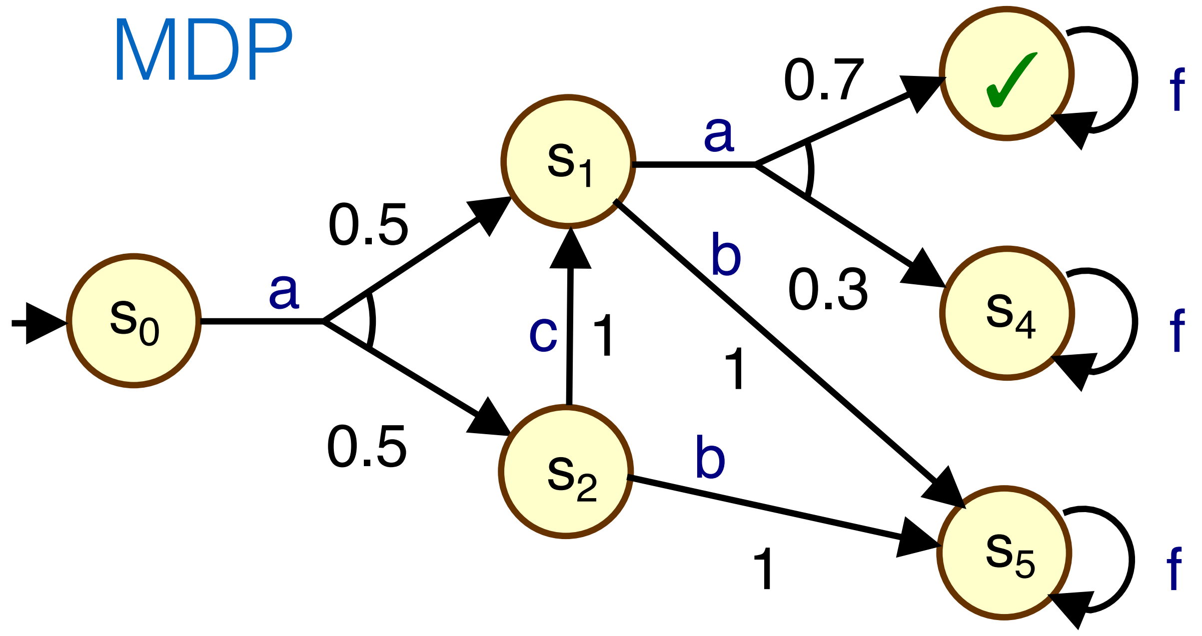
- Interaction with a second robot



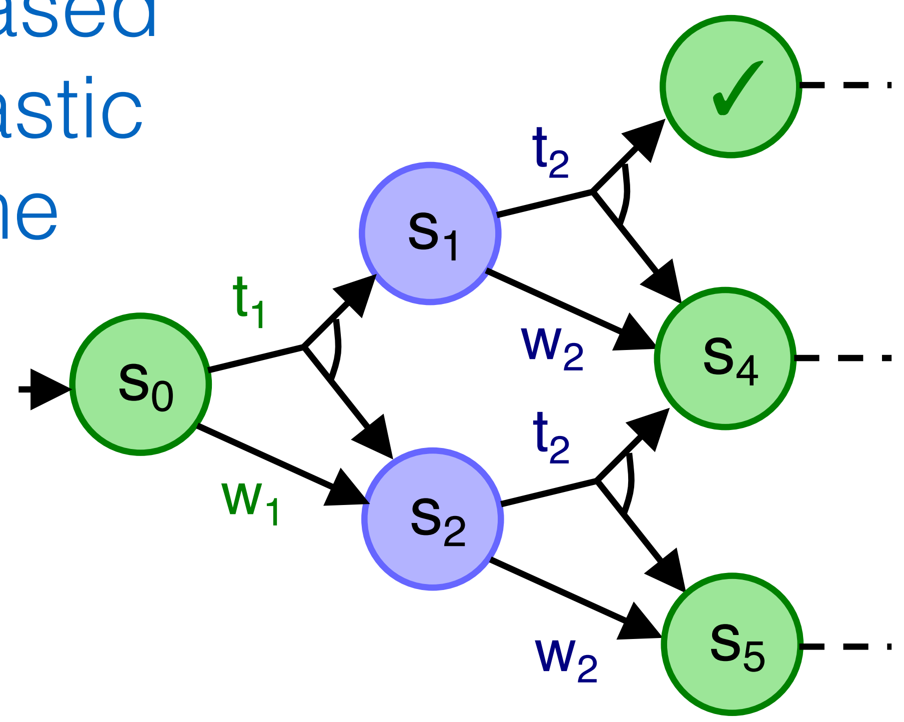
# Stochastic games

- MDPs model sequential decision making
  - for a single agent, under stochastic uncertainty
  - we may need adversarial (uncontrollable) decisions
  - or collaborative decision making for multiple agents
- A (turn-based, two-player) stochastic game
  - takes the form $\mathcal{G} = (\{1,2\}, S, \langle S_1, S_2 \rangle, s_0, A, P)$ where:
  - states $S$, initial state $s_0$ and actions $A$ are as for MDPs
  - $S_1, S_2 \subseteq S$ are the (disjoint) states controlled by players 1 and 2
  - transition function $P : S \times A \times S \to [0,1]$ is also as for MDPs
- Another possibility: concurrent stochastic games
  - with $P : S \times (A_1 \times A_2) \times S \to [0,1]$

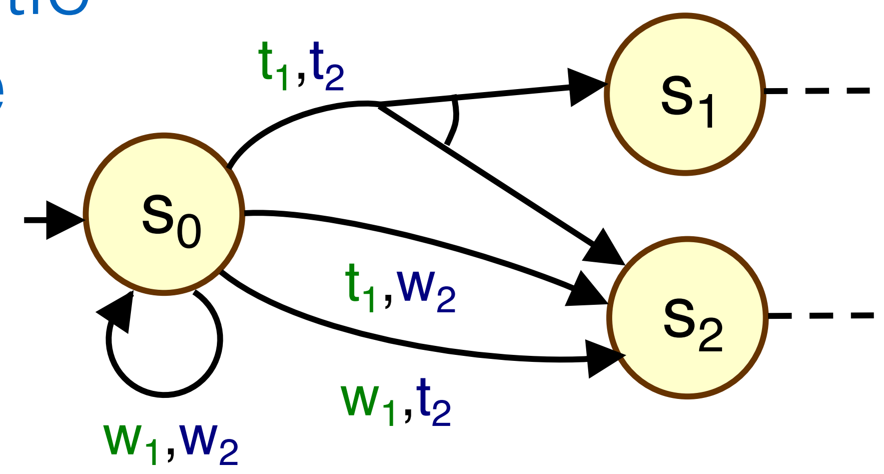
MDP




turn-based stochastic game

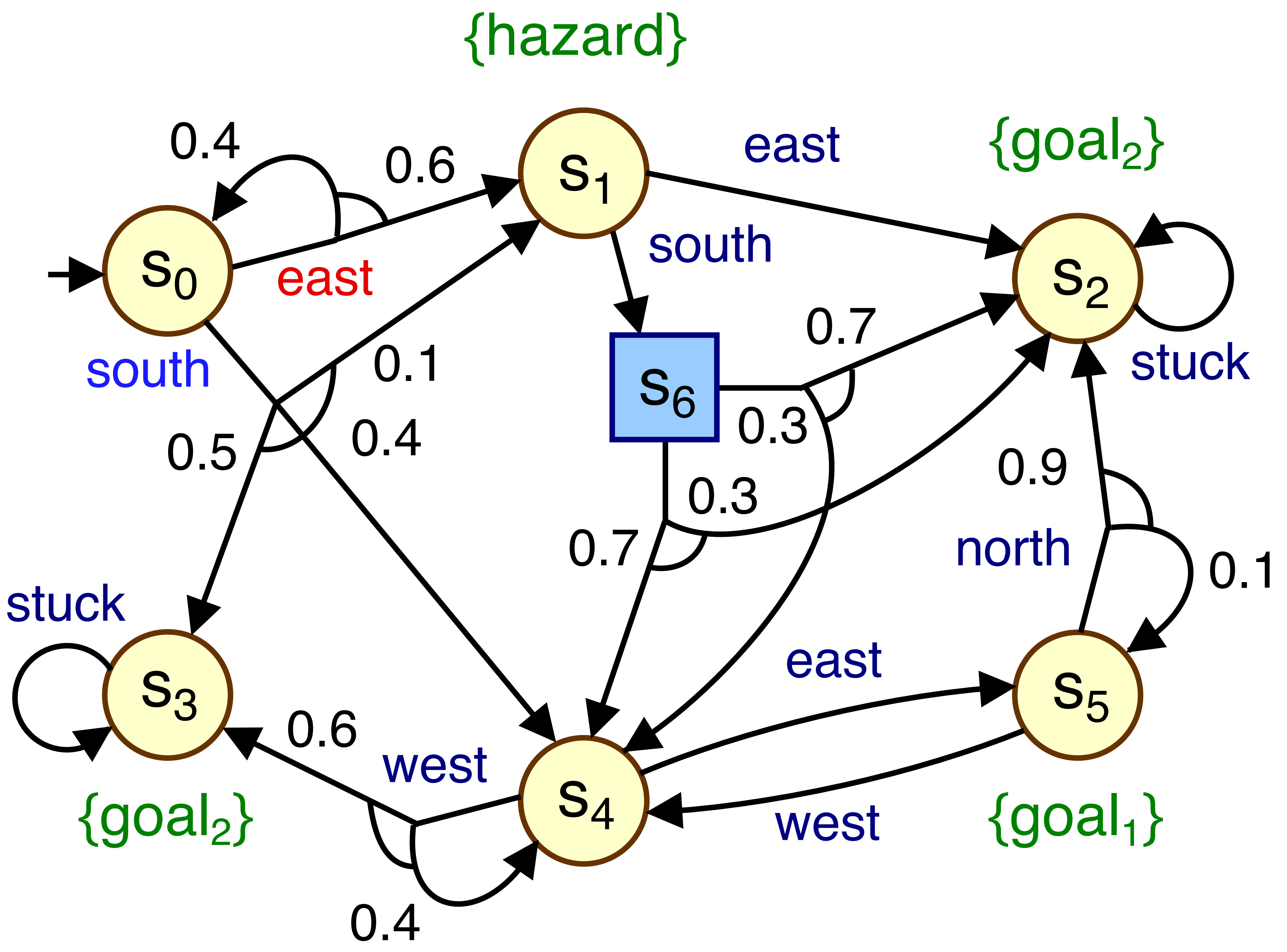



concurrent stochastic game

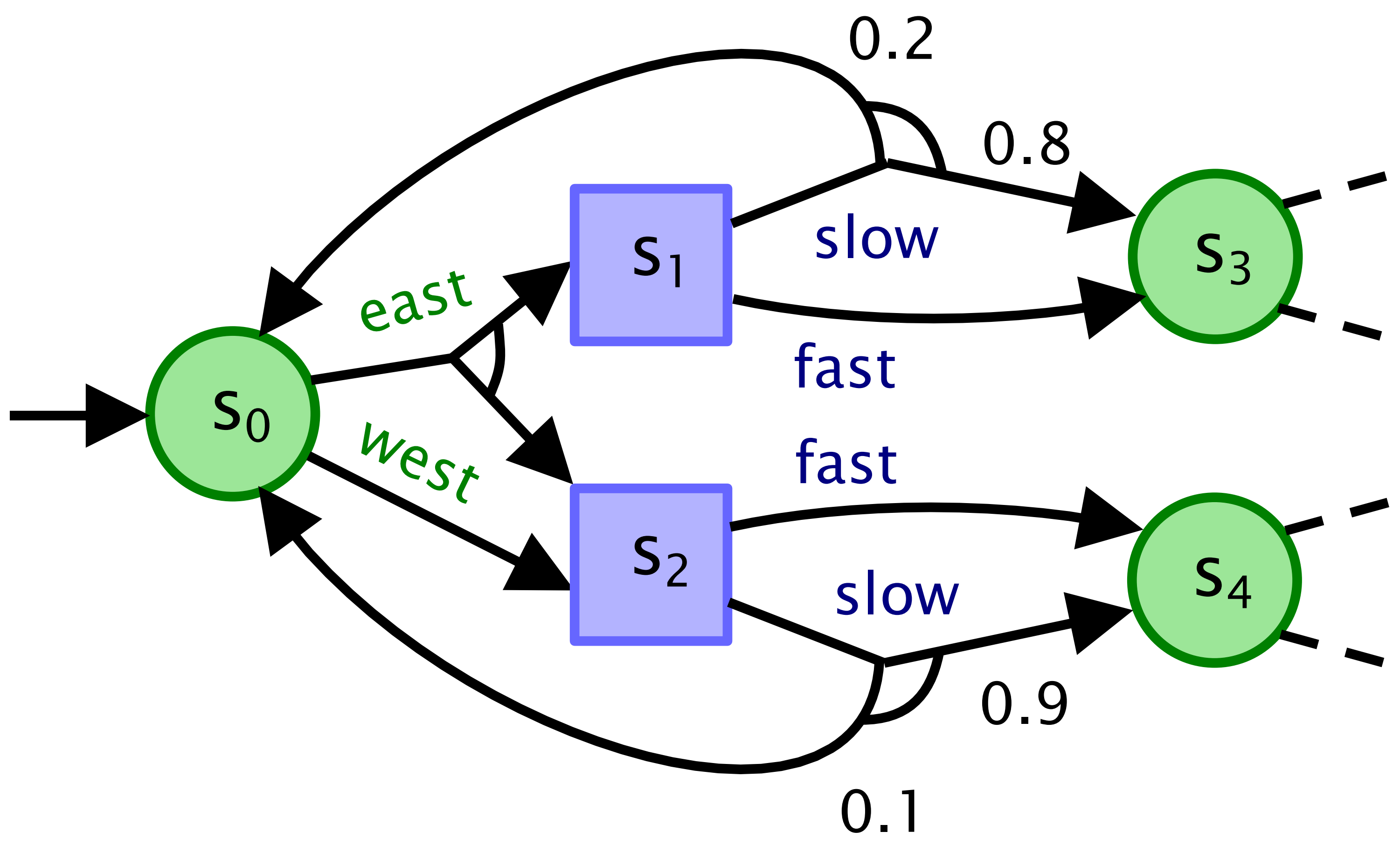


# Turn-based stochastic games

uncontrollable/unknown interference

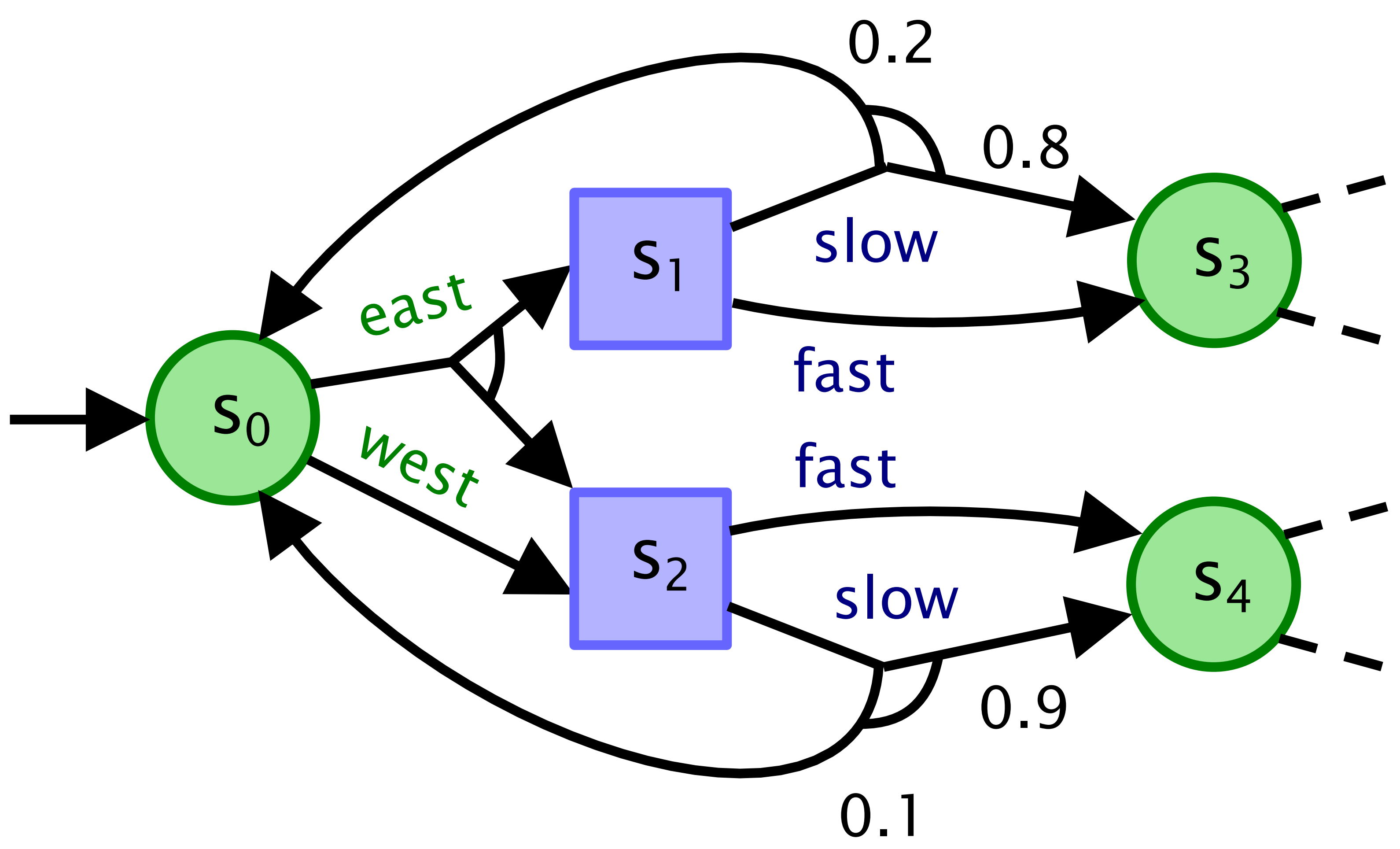



shared autonomy:
human-robot control



# Strategies for stochastic games

- Strategies (policies) for turn-based stochastic games
  - a strategy for player i is a mapping $\pi_i : (S \times A)^* \times S_i \rightarrow Dist(A)$
  - a strategy profile $(\pi_1, \pi_2)$ defines strategies for both players
- For state $s$ of game $\mathscr{G}$ and strategy profile $(\pi_1, \pi_2)$:
  - we can define probability space $Pr_s^{\pi_1,\pi_2}$, random variables $\mathbb{E}_s^{\pi_1,\pi_2}(X)$ and value functions $V^{\pi_1,\pi_2}(s)$




- Strategies
  - can again be deterministic / randomised or memoryless / history-dependent
  - $\Pi_i$ is the set of all strategies for player $i \in \{1,2\}$

# Objectives for stochastic games

- Objectives $V_1$, $V_2$ for players 1 and 2 can be distinct
  - simple, useful scenario: zero-sum (directly opposing), i.e., $V_1 = -V_2$
  - so we assume a single objective V which one player maximises and the other minimises

- Consider MaxProb for player 1 (other cases are similar):

  $\max_{\pi_1 \in \Pi_1} \min_{\pi_2 \in \Pi_2} V^{\pi_1,\pi_2}(s)$ where $V^{\pi_1,\pi_2}$ is exactly as for MDP MaxProb

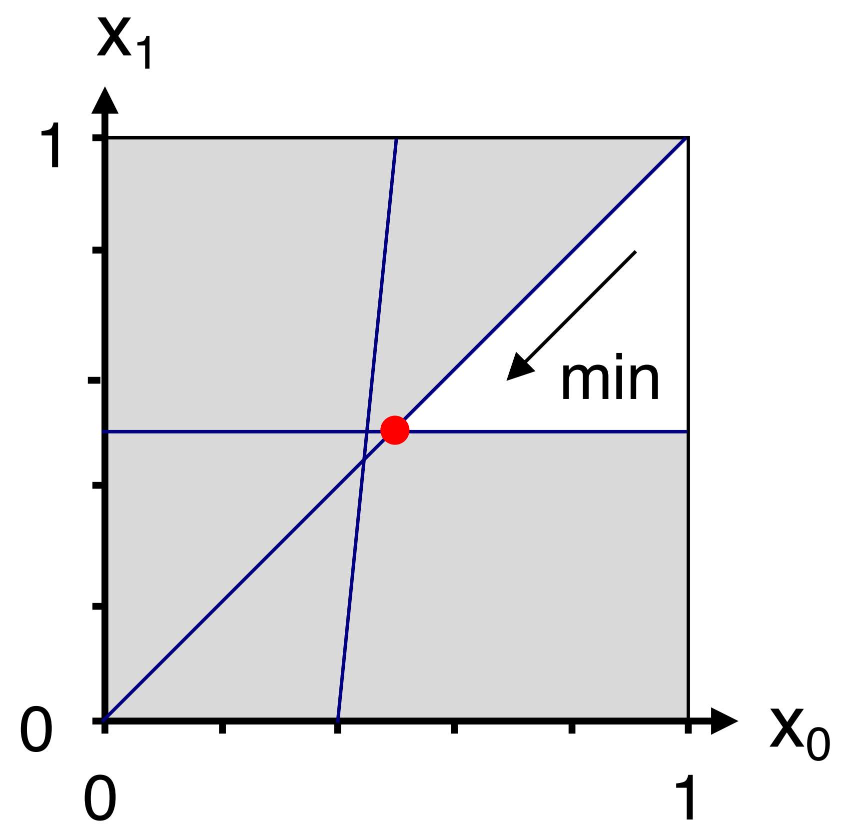
- Games are determined, i.e., for all states $s$:

$$\max_{\pi_1 \in \Pi_1} \min_{\pi_2 \in \Pi_2} V^{\pi_1,\pi_2}(s) = \min_{\pi_2 \in \Pi_2} \max_{\pi_1 \in \Pi_1} V^{\pi_1,\pi_2}(s)$$

- So we define:
  - optimal value: $V^*(s) = \max_{\pi_1 \in \Pi_1} \min_{\pi_2 \in \Pi_2} V^{\pi_1,\pi_2}(s)$
  - optimal strategy (for player 1): $\pi^* = \text{argmax}_{\pi_1 \in \Pi_1} \min_{\pi_2 \in \Pi_2} V^{\pi_1,\pi_2}(s_0)$

# Solving stochastic games

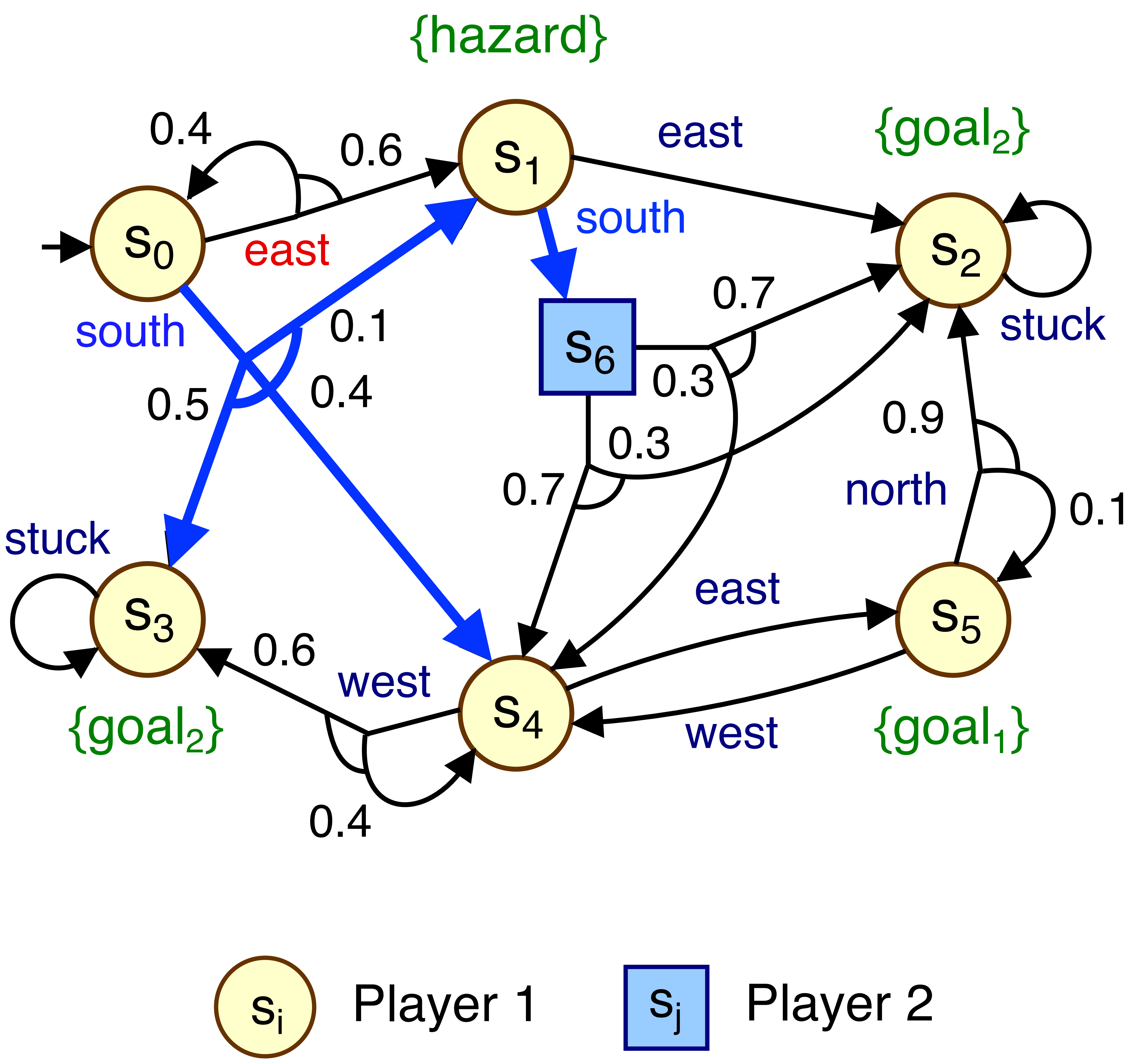
- Memoryless deterministic strategies suffice (for both players)
- Complexity worse than for MDPs: NP ∩ co-NP, rather than P
  - LP approach does not adapt (but strategy improvement is possible)
- In practice: dynamic programming (value iteration) works well
  - e.g., for MaxProb:




$$x_s^k = \begin{cases} 1 & \text{if } s \in goal \\ 0 & \text{if } s \notin goal \text{ and } k = 0 \\ \max_{a \in A(s)} \sum_{s' \in S} P_s^a(s') \cdot x_{s'}^{k-1} & \text{if } s \notin goal, s \in S_1 \text{ and } k > 0 \\ \min_{a \in A(s)} \sum_{s' \in S} P_s^a(s') \cdot x_{s'}^{k-1} & \text{if } s \notin goal, s \in S_2 \text{ and } k > 0 \end{cases}$$

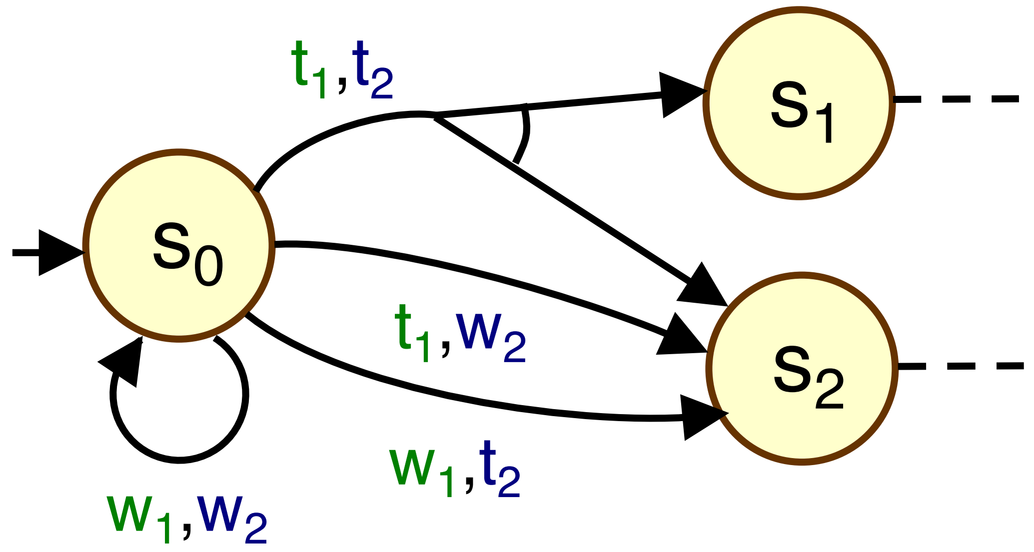
# Running example

- Optimal player 1 strategy changes:



# Zero-sum concurrent stochastic games

- Concurrent stochastic games: strategies, value functions defined similarly
  - games are still determined: $\max_{\pi_1 \in \Pi_1} \min_{\pi_2 \in \Pi_2} V^{\pi_1,\pi_2}(s) = \min_{\pi_2 \in \Pi_2} \max_{\pi_1 \in \Pi_1} V^{\pi_1,\pi_2}(s)$
  - but optimal strategies still memoryless but now randomised



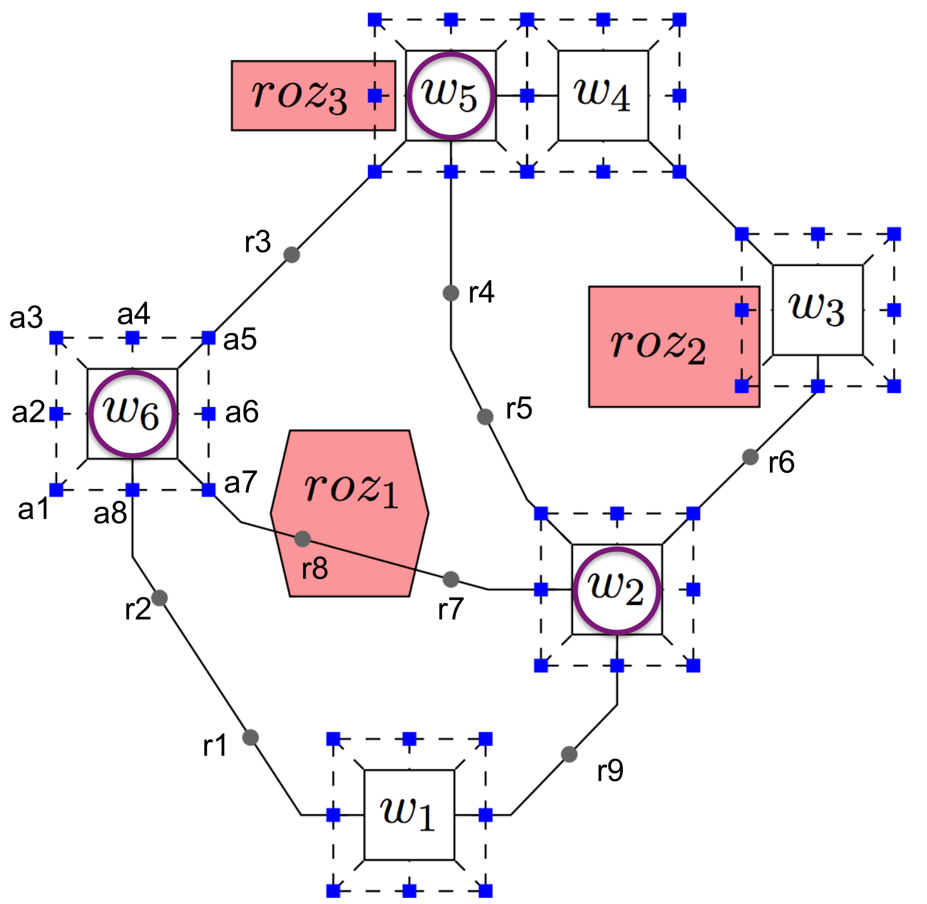

- Value iteration can be extended: $x_s^k = \begin{cases} 1 & \text{if } s \in goal \\ 0 & \text{if } s \notin goal \text{ and } k = 0 \\ val(Z) & \text{otherwise} \end{cases}$
  - where $val(Z)$ is the value of the matrix game with payoffs: $z_{a,b} = \sum_{s' \in S} P_s^{a,b}(s') \cdot x_{s'}^{k-1}$
  - solved via the linear program →
  - $p_a$ gives the probability of player 1 picking action $a$ in its optimal strategy

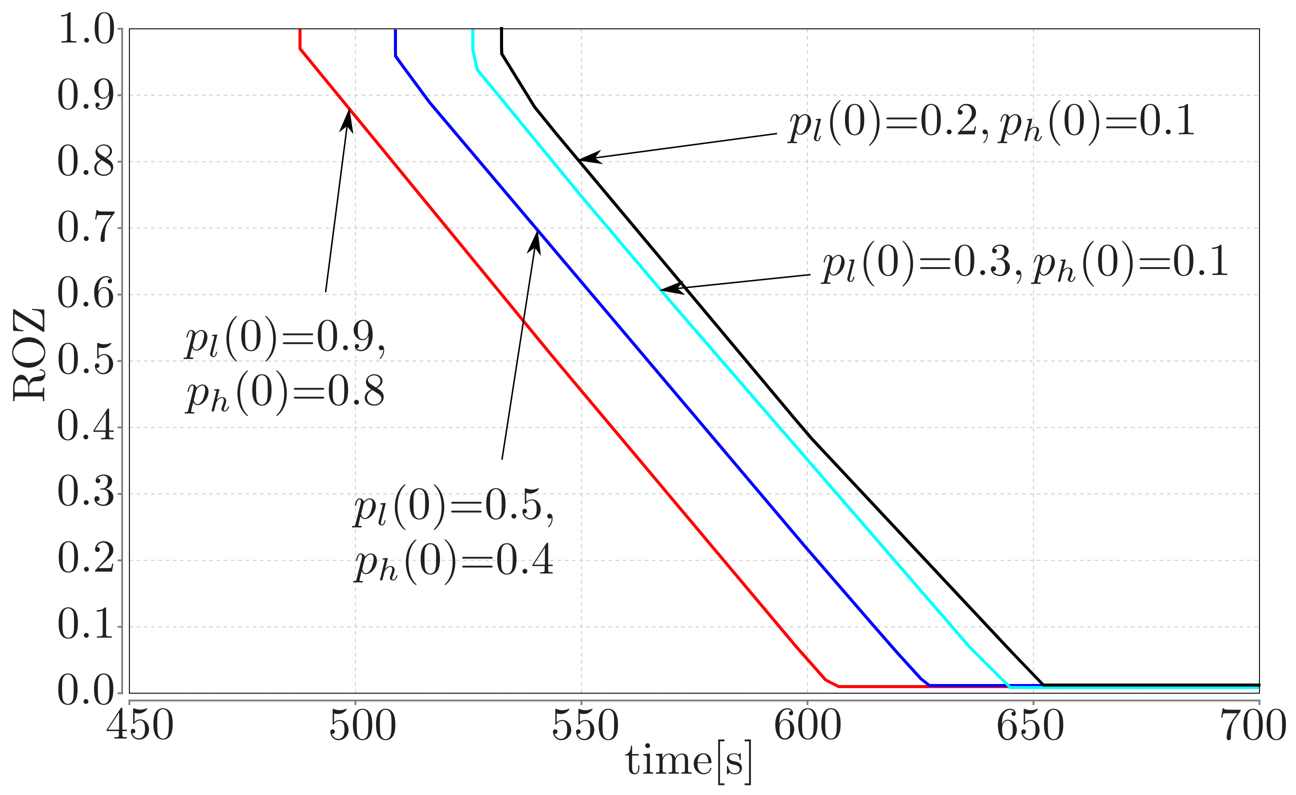
Maximise game value $v$ subject to:

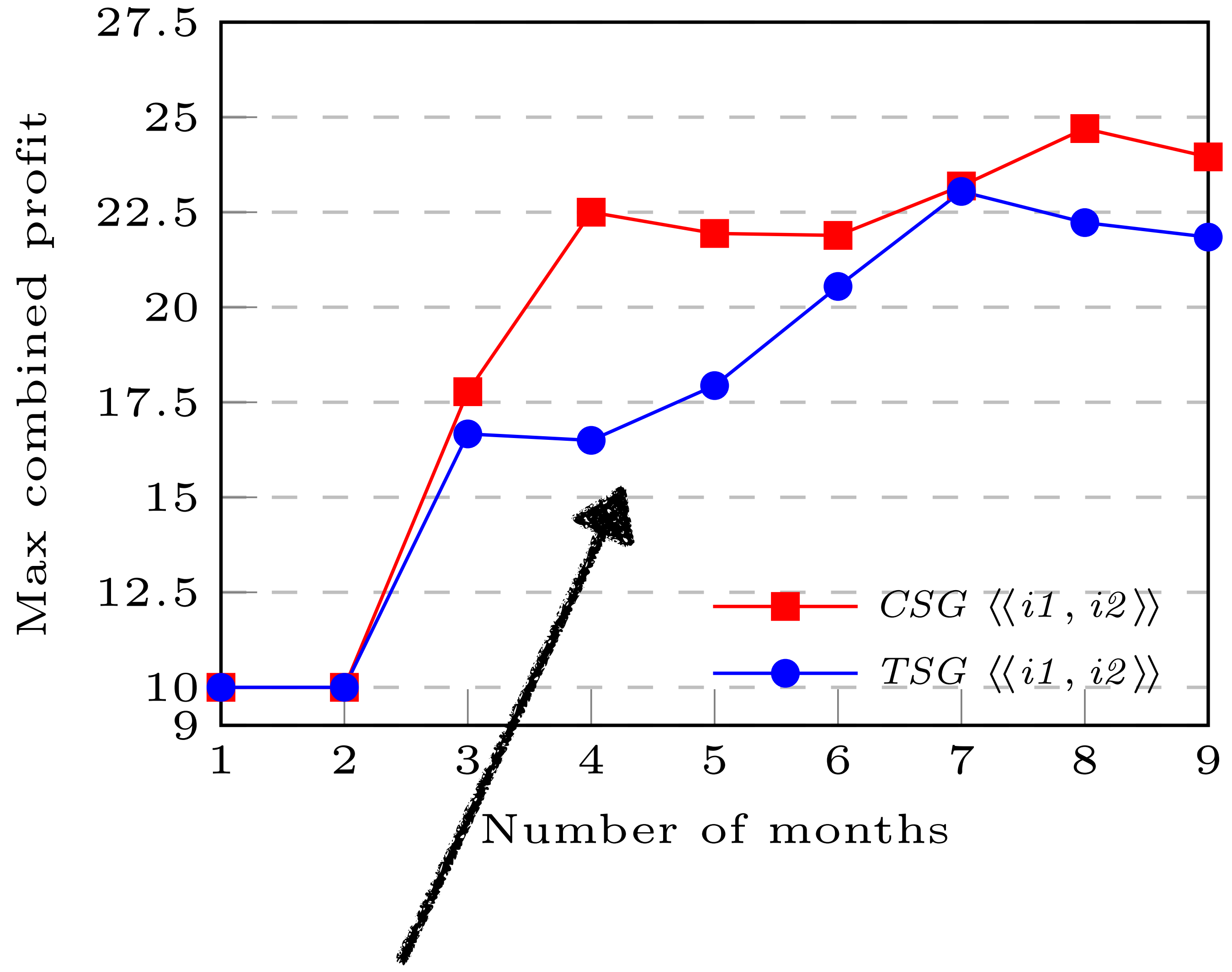
| | |
|---|---|
| $\Sigma_{a \in A_1} p_a \cdot z_{a,b} \geq v$ | for $b \in A_2$ |
| $p_a \geq 0$ | for $a \in A_1$ |
| $\Sigma_{a \in A_1} p_a = 1$ | |

# Sequential decision making with stochastic games

- UAV road surveillance
  - with partial human control (varying operator accuracy)
- Futures market investment
  - market is part stochastic, part adversarial
- Multi-robot control
  - adversarial (worst-case) vs. collaborative

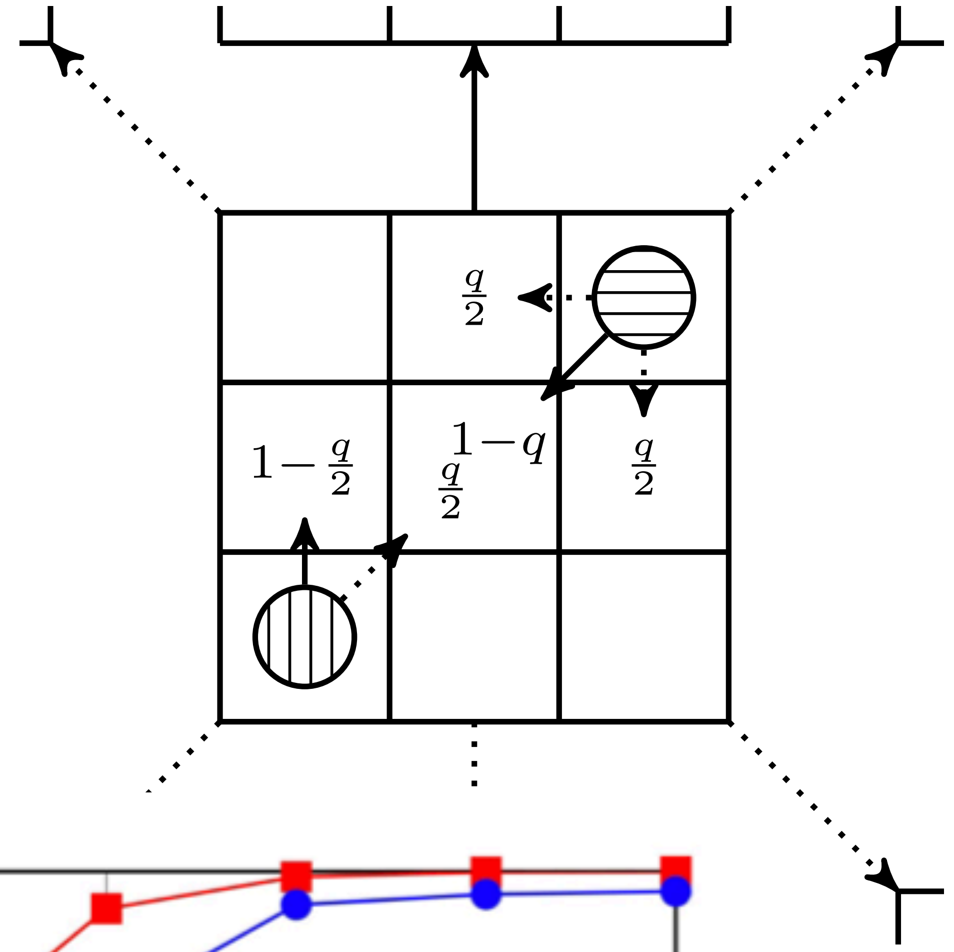


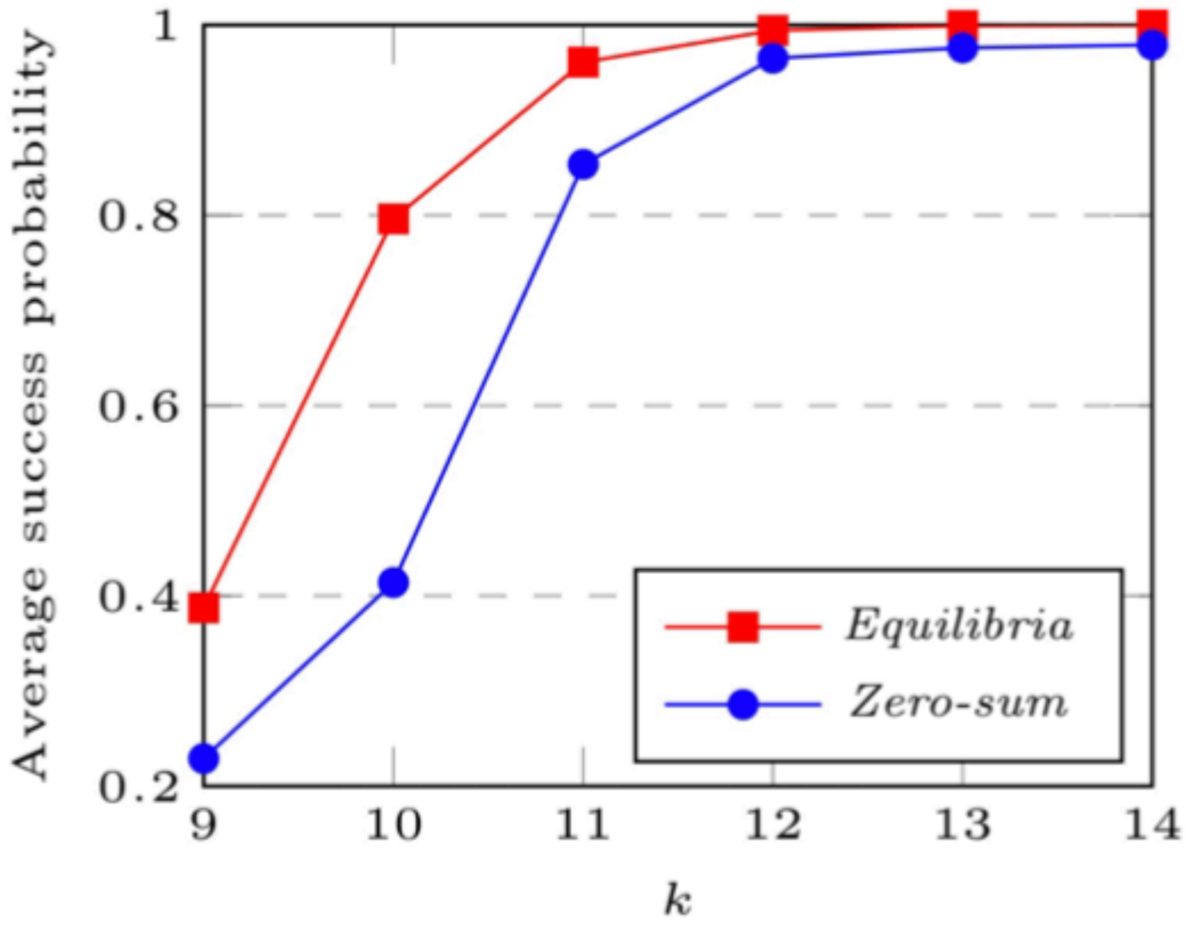




Turn-based game too pessimistic (unrealistic adversary)

PRISM-games

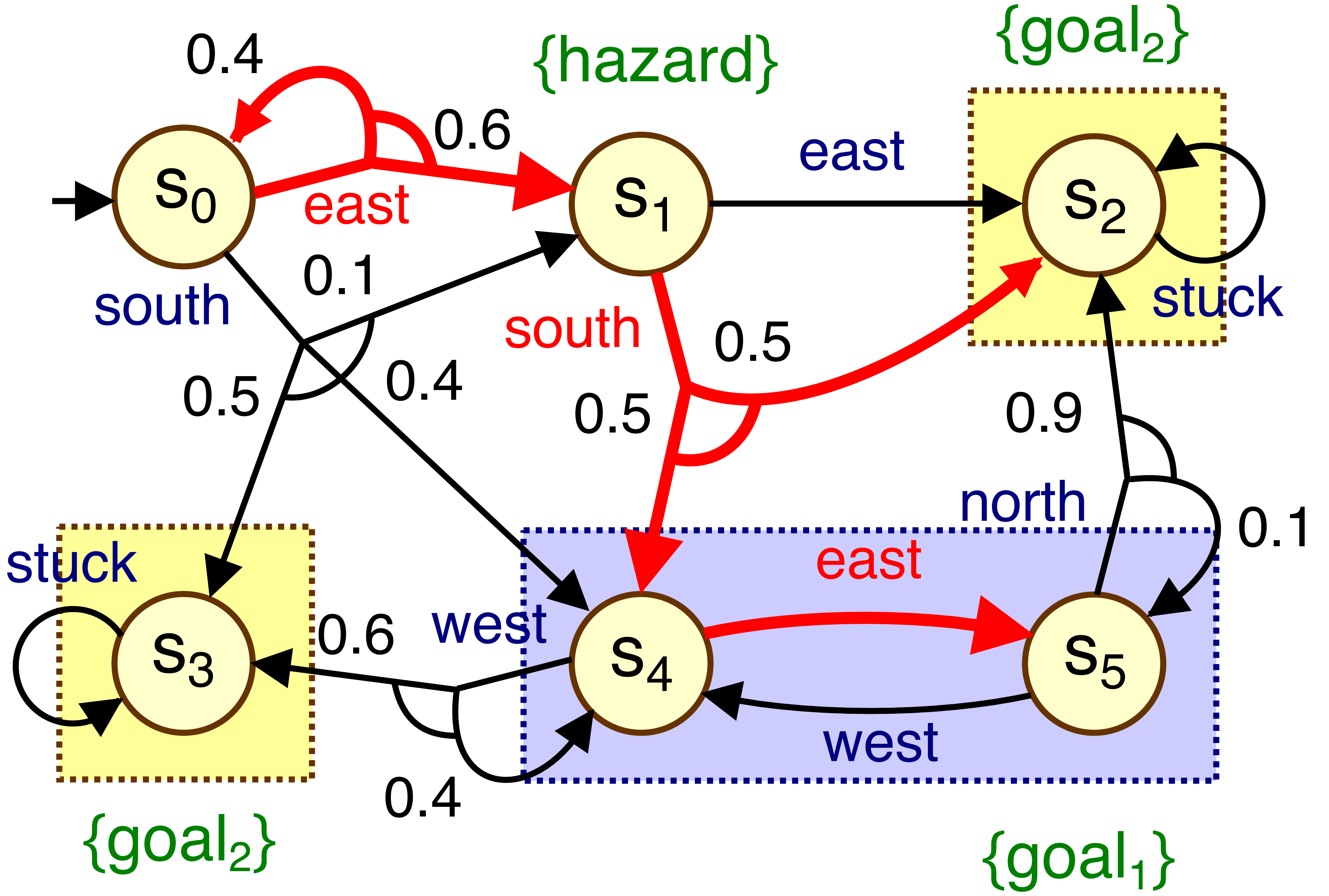


# Uncertain MDPs

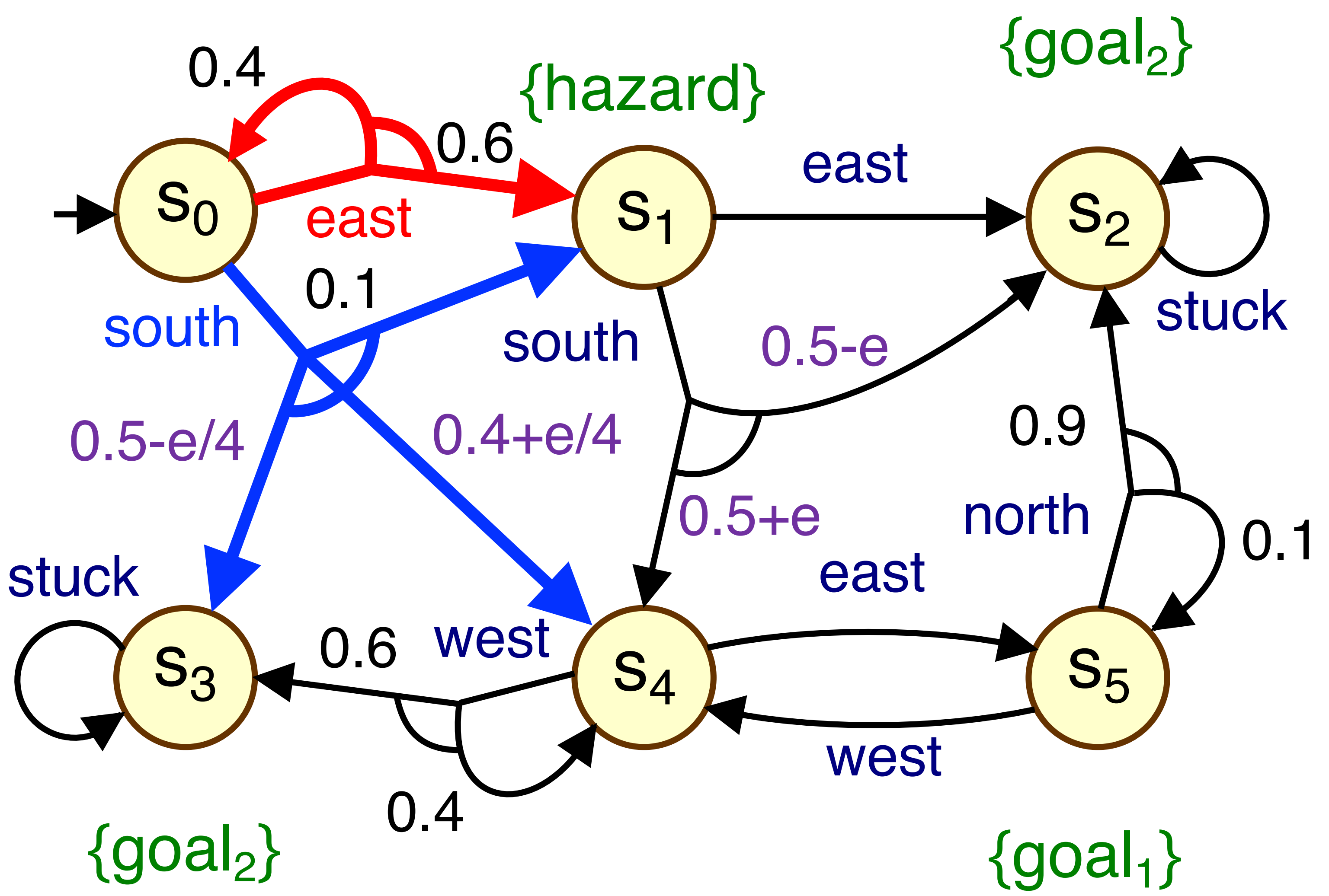
# MDPs + epistemic uncertainty

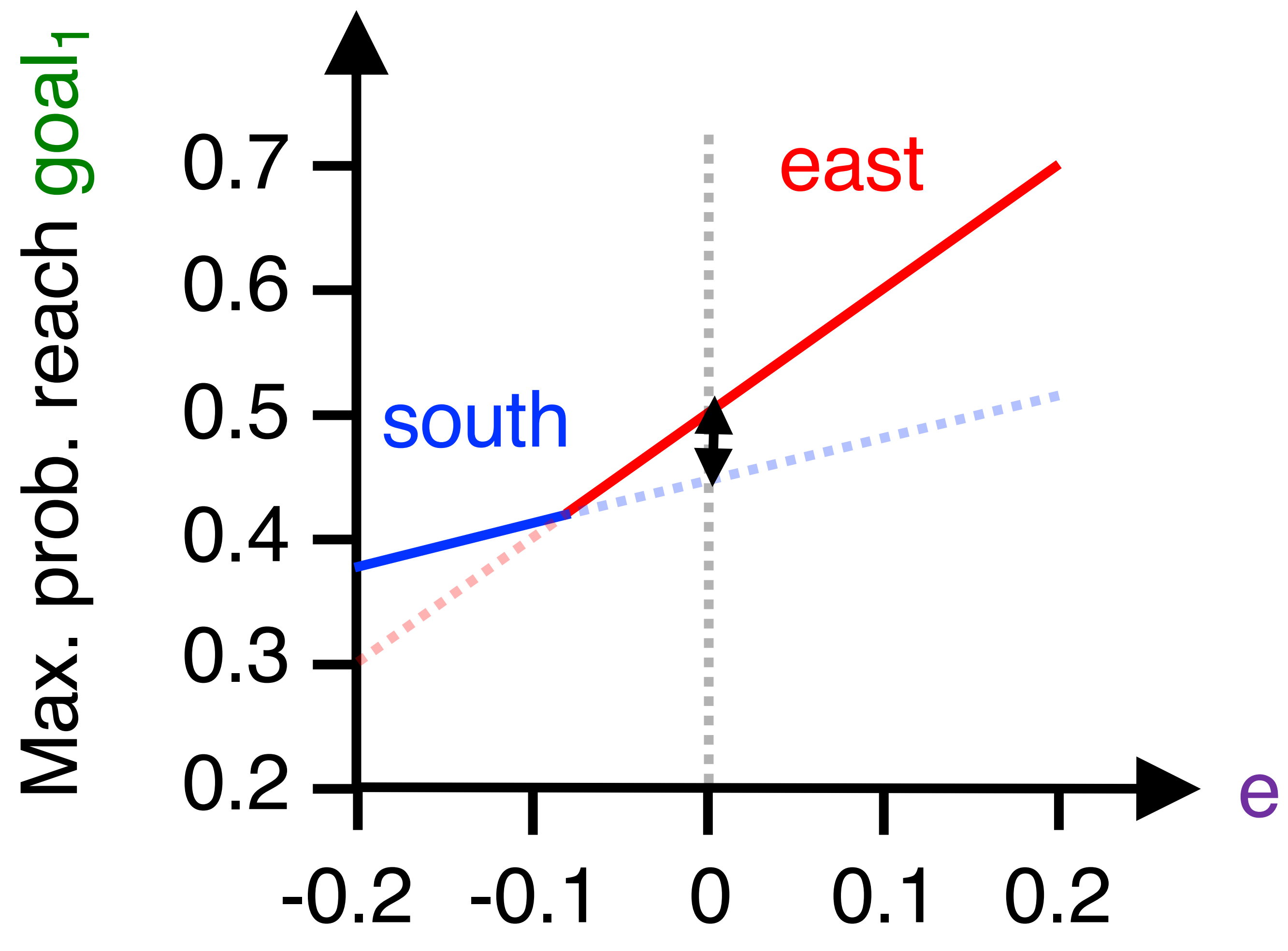
- We can use MDPs for sequential decision making under (aleatoric) uncertainty
  - modelled here using transition probabilities (often learnt from data)



# MDPs + epistemic uncertainty

- We can use MDPs for sequential decision making under (aleatoric) uncertainty
  - modelled here using transition probabilities (often learnt from data)
- Policies can be sensitive to small perturbations in transition probabilities
  - so “optimal” policies can in fact be sub-optimal



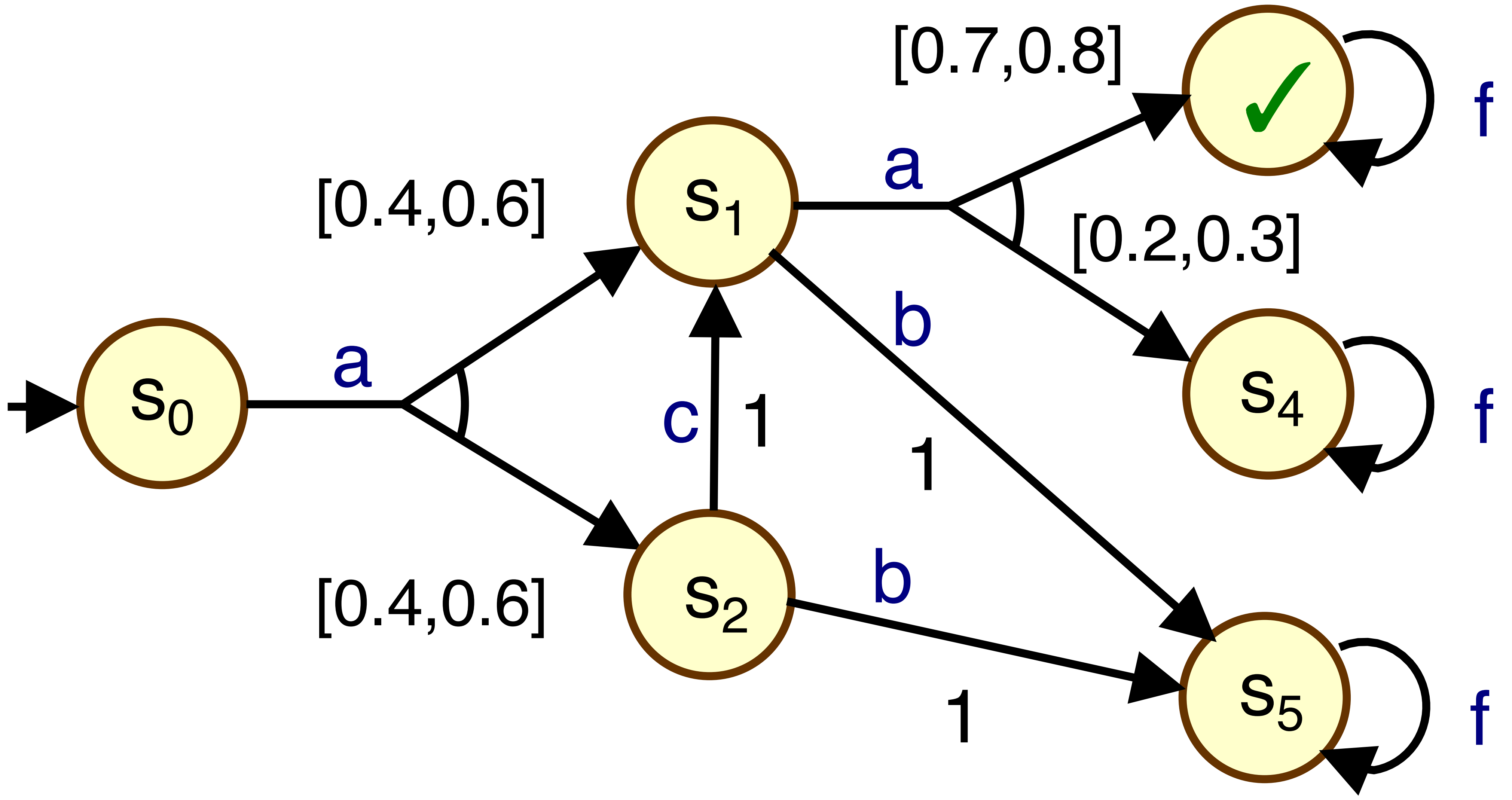
# MDPs + epistemic uncertainty

- We can use MDPs for sequential decision making under (aleatoric) uncertainty
  - modelled here using transition probabilities (often learnt from data)
- Policies can be sensitive to small perturbations in transition probabilities
  - so "optimal" policies can in fact be sub-optimal
- Uncertain MDPs: MDPs + epistemic uncertainty (model uncertainty)
  - we focus here on uncertainty in transition probabilities
- Key questions:
  - how to model (and solve for) epistemic uncertainty?
  - what guarantees do we get?
  - is it statistically accurate?
  - how computationally efficient is it?

# Uncertain MDPs

- An uncertain MDP (uMDP) takes the form $\mathcal{M} = (S, s_0, A, \mathcal{P})$ where:
  - states $S$, initial state $s_0$ and actions $A$ are as for MDPs
  - $\mathcal{P}$ is the transition function uncertainty set
    - i.e., each $P \in \mathcal{P}$ is a transition function $P : S \times A \times S \rightarrow [0,1]$

- The uncertainty set $\mathcal{P}_s^a \subseteq Dist(S)$
  - for each $s \in S$, $a \in A(s)$
  - is $\mathcal{P}_s^a = \{P_s^a \ : \ P \in \mathcal{P}\}$
  - similarly: $\mathcal{P}^a = \{P^a \ : \ P \in \mathcal{P}\}$
  - ($\mathcal{P}_s^a$ sometimes "ambiguity sets")



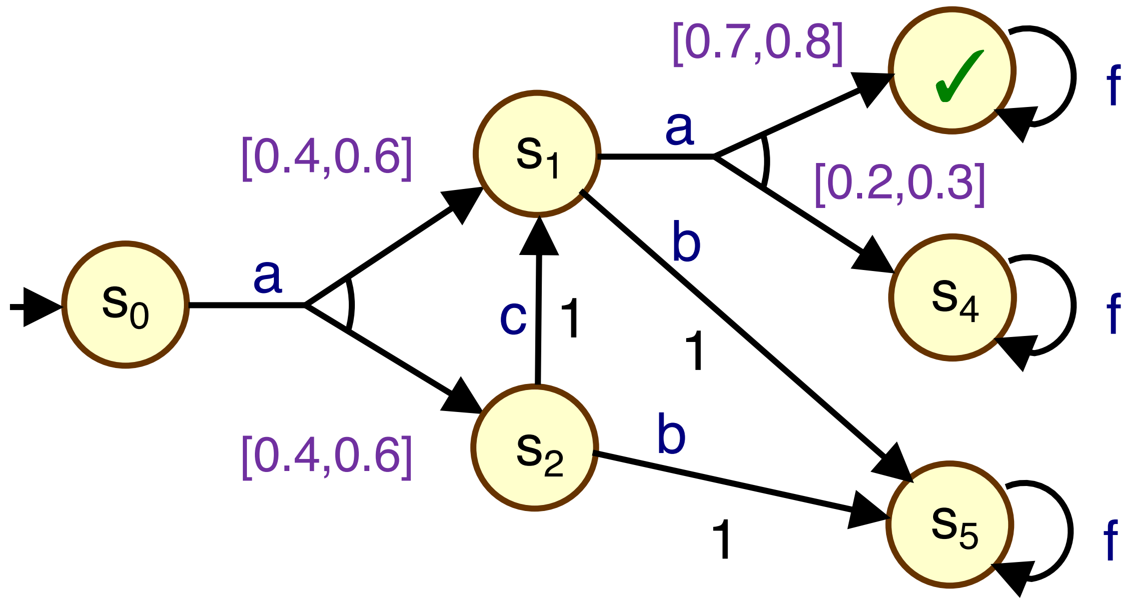
# Uncertain MDPs

- Semantics of a uMDP $\mathscr{M} = (S, s_0, A, \mathscr{P})$
  - $\mathscr{M}$ can be seen as a (usually infinite) set of MDPs: $[\![\mathscr{M}]\!] = \{\mathscr{M}[P] : P \in \mathscr{P}\}$
  - where $\mathscr{M}[P] = (S, s_0, A, P)$ is $\mathscr{M}$ instantiated with $P \in \mathscr{P}$

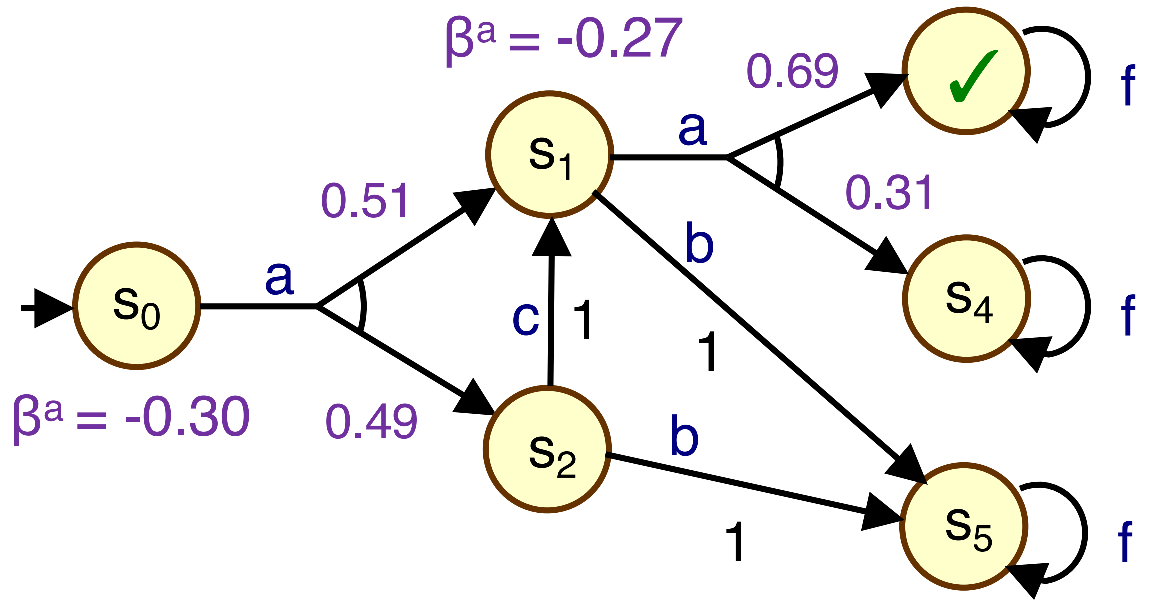
- But other views are possible
  - dynamic, Bayesian, …

- Some examples of uMDPs

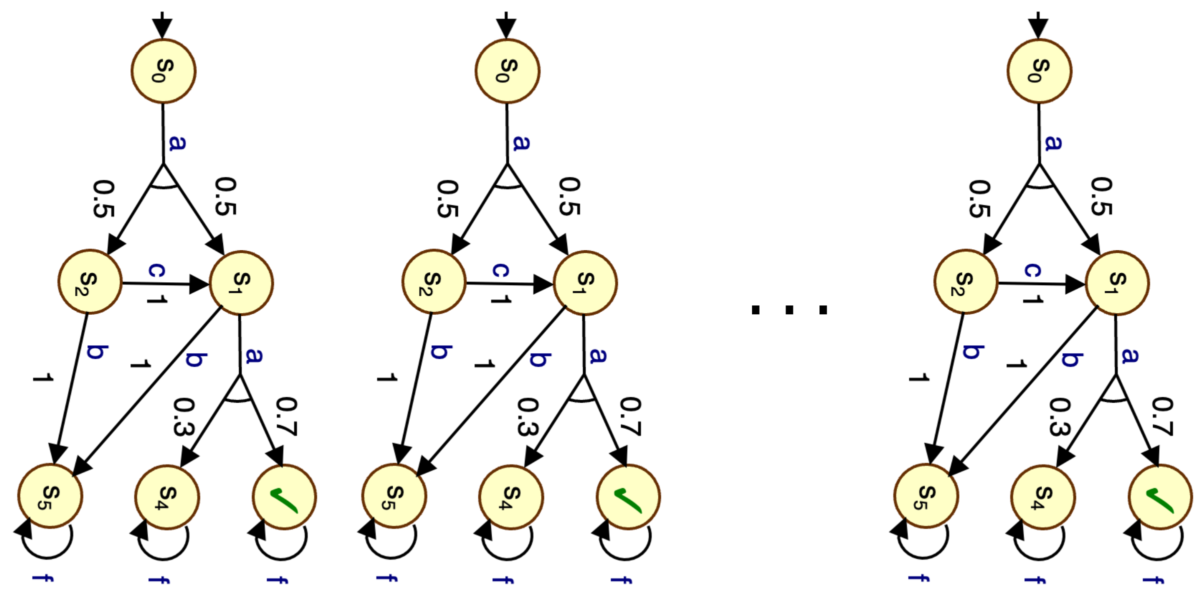
Interval MDPs (IMDPs)




Likelihood MDPs

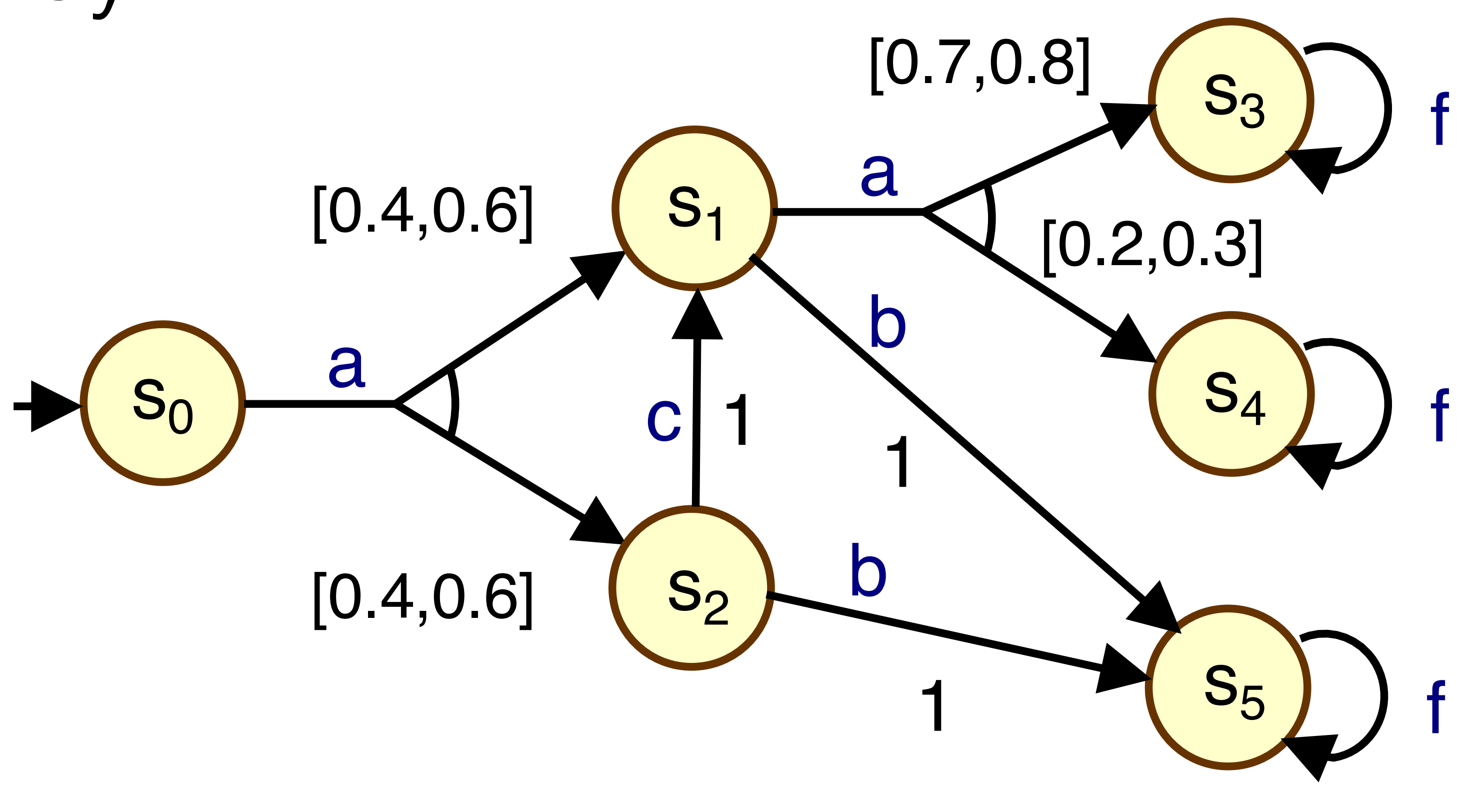



Sampled MDPs



# Uncertainty set dependencies

- Can we allow dependencies between uncertainty sets?
  - implications for computational tractability and modelling accuracy



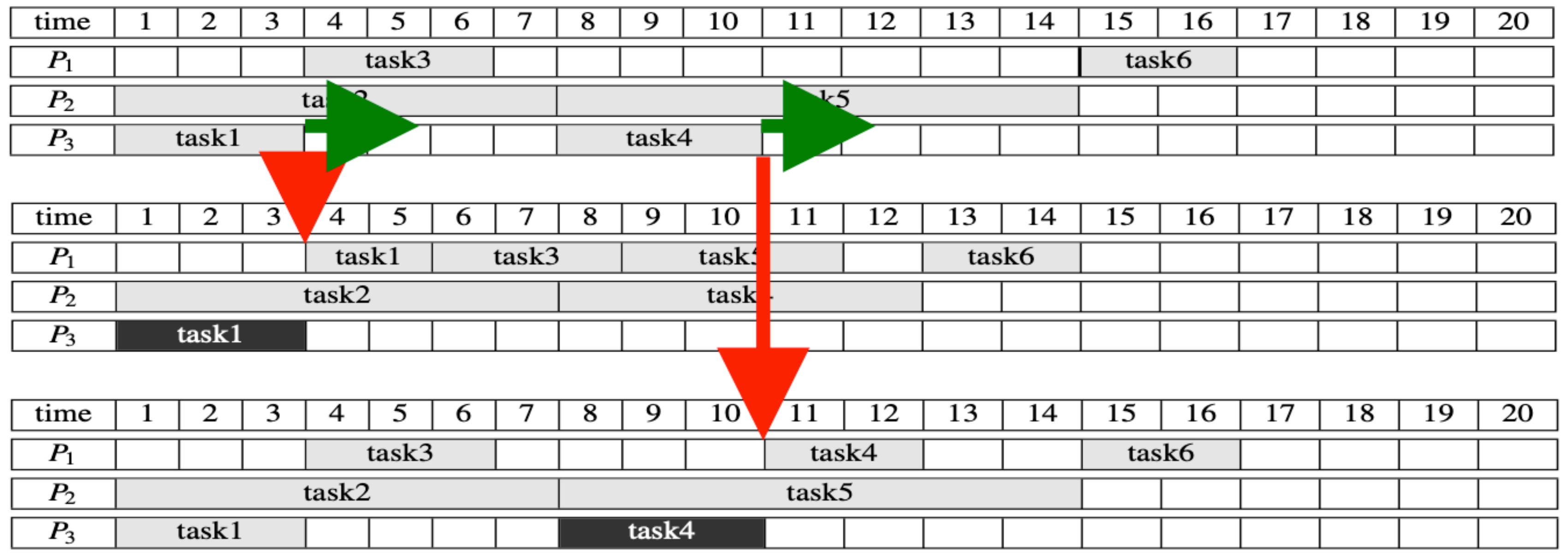

- Rectangularity
  - transition function uncertainty set $\mathscr{P}$ is (s,a)-rectangular
    - if we have $\mathscr{P} = \times_{(s,a)\in S\times A}\, \mathscr{P}^a_s$
    - i.e., if there are no dependencies between uncertainty sets for each s, a
  - interval MDPs are (s,a)-rectangular ("sampled MDPs" might not be)
  - we will assume (s,a)-rectangularity for now (and later relax it)

- We can also define s-rectangularity [Wiesemann et al.]
  - $\mathscr{P} = \times_{s\in S}\, \mathscr{P}^s$ where $\mathscr{P}_s = \{(P^a_s)_{a\in A} \ : \ P \in \mathscr{P}\}$

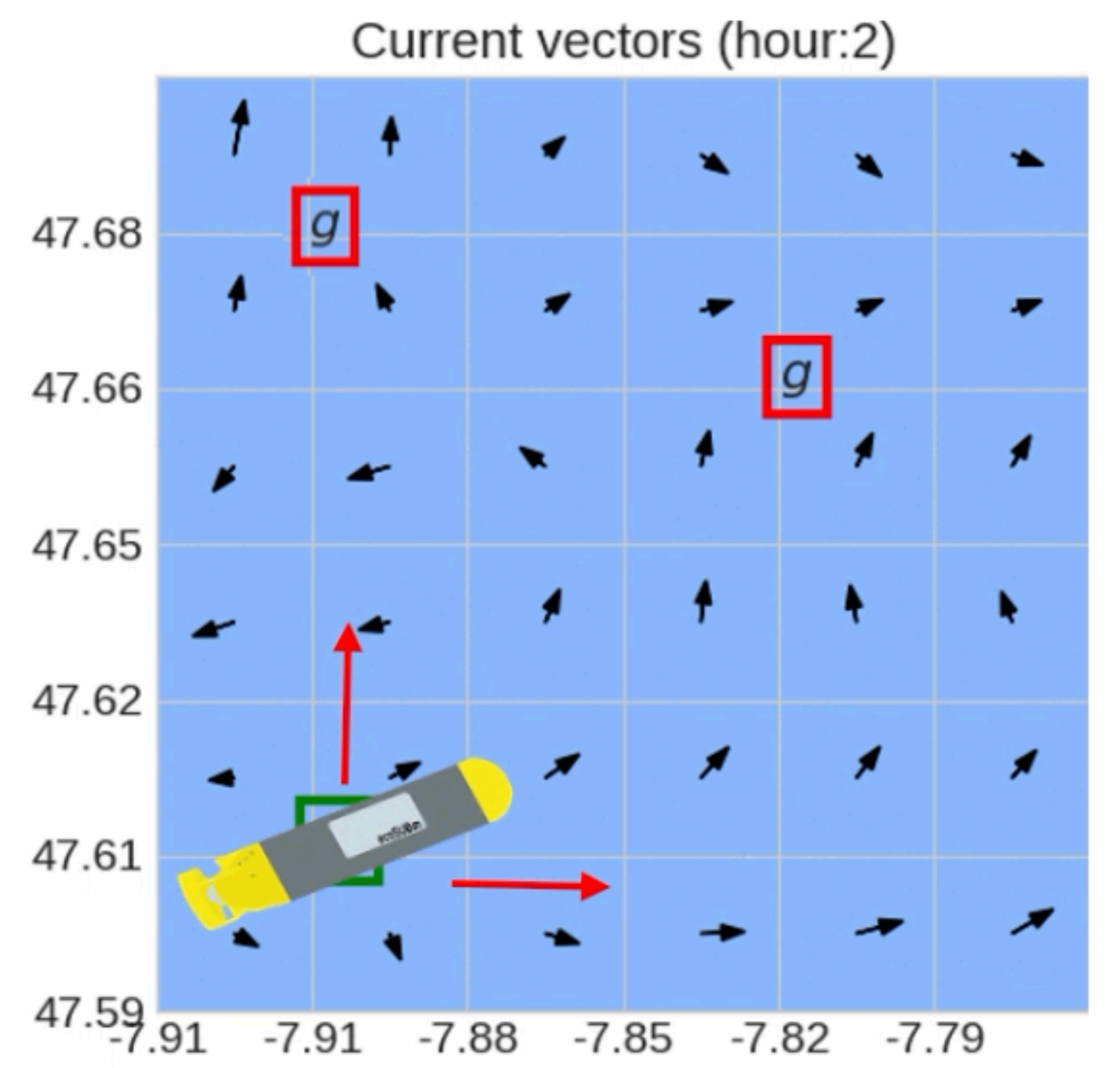
# Non-rectangular uMDPs

- When might dependences between uncertainties arise?

Task scheduling in the presence of faulty processors

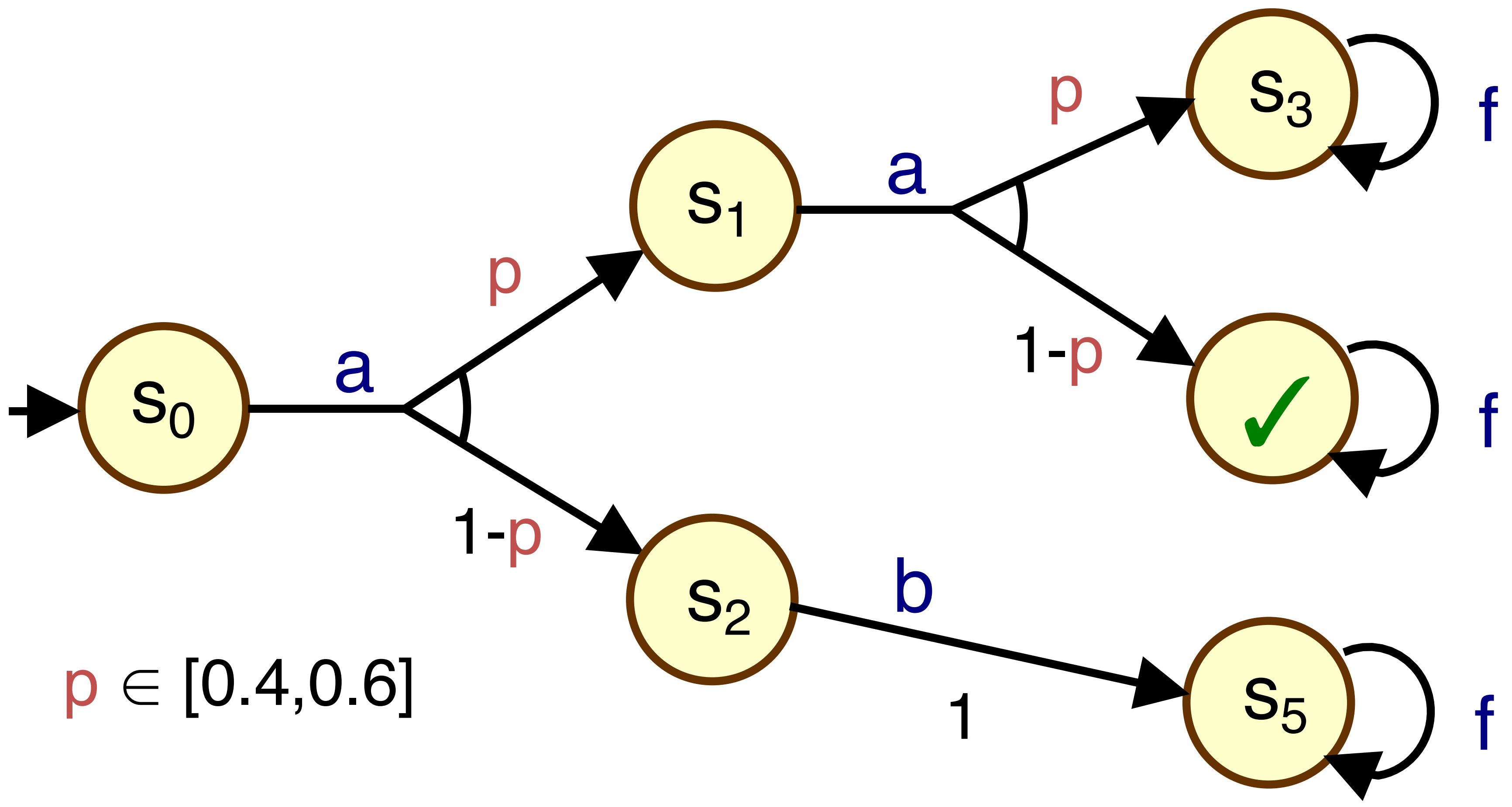



Underwater vehicle control in unknown ocean currents



# Non-rectangular uMDPs

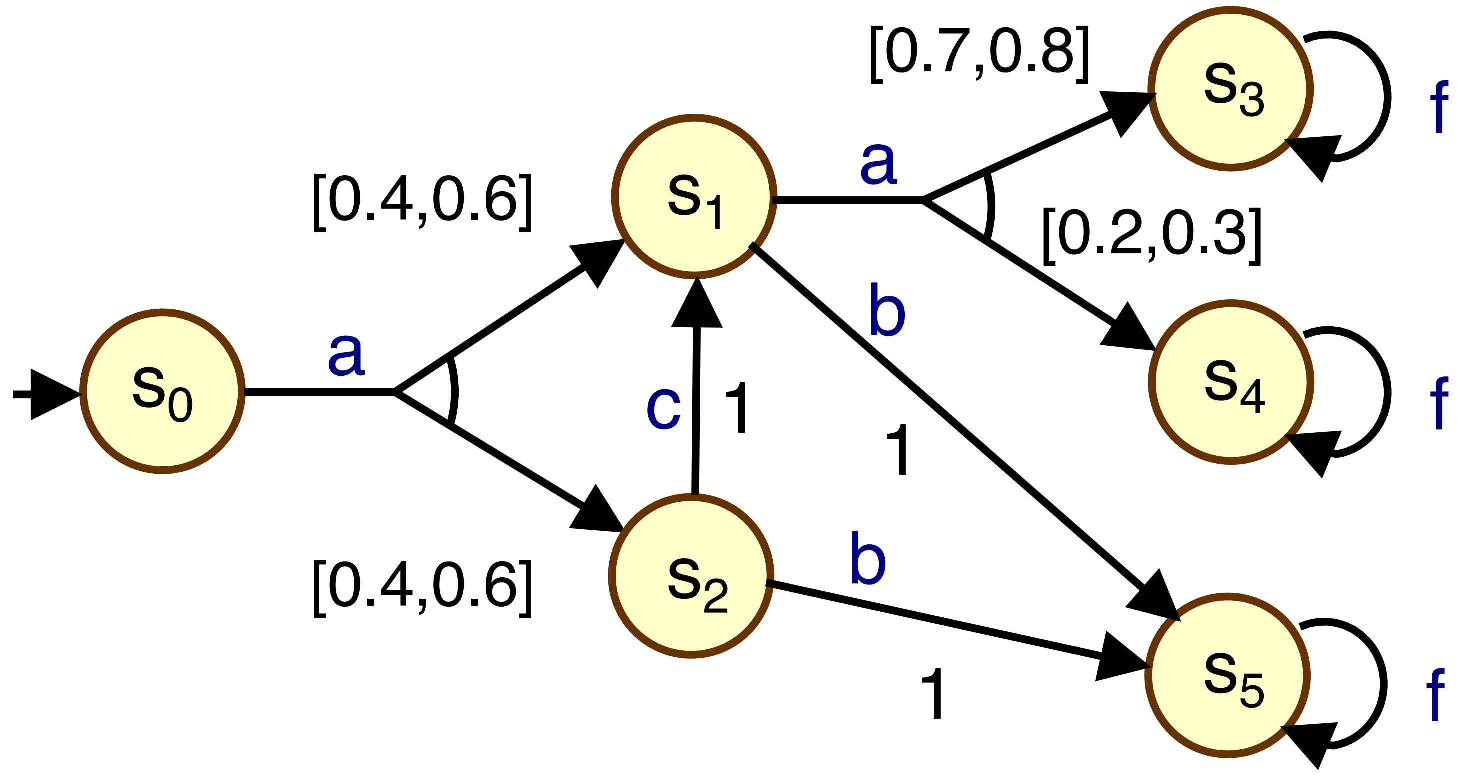
- Example MDP (in fact, just a single policy) with parameter p



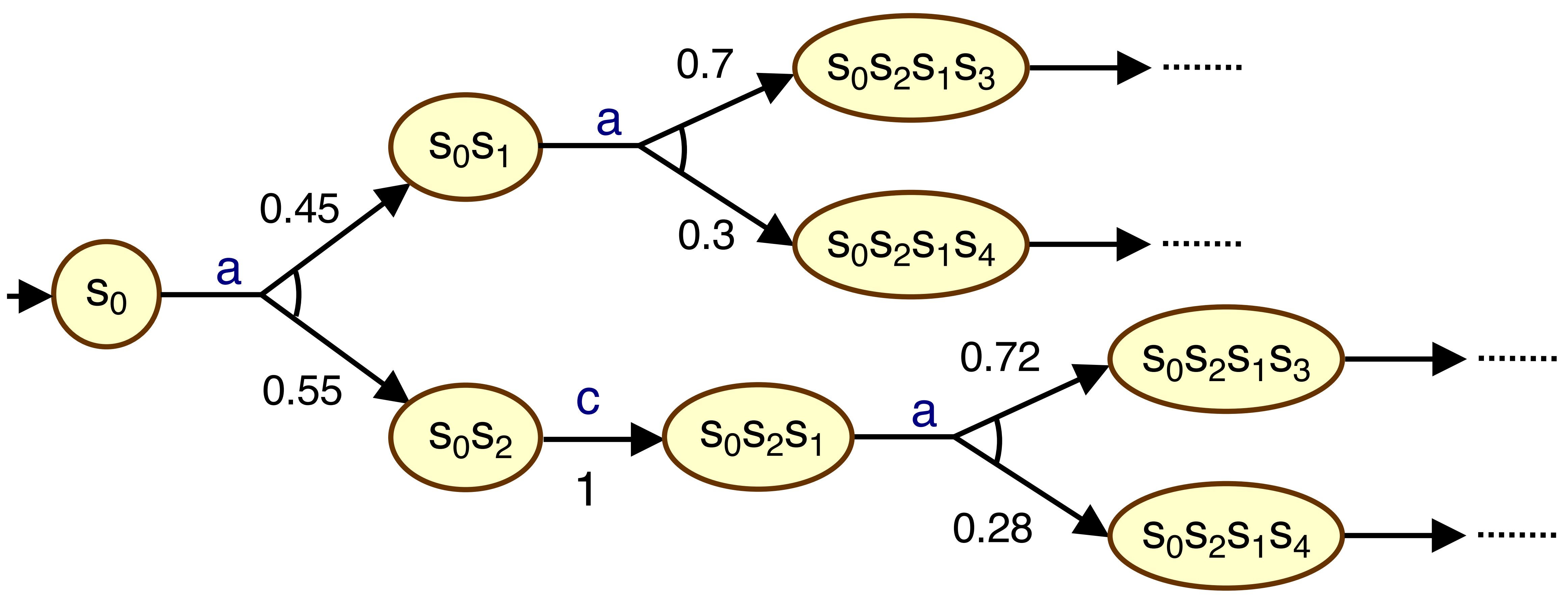

- Worst-case probability to reach ✓?
  - $\min\{p(1-p)\ :\ p \in [0.4,0.6]\} = 0.4 \cdot (1-0.4) = 0.24$
- Worst-case probability to reach ✓ under rectangularity assumptions?
  - $\min\{p_1(1-p_2)\ :\ p_1, p_2 \in [0.4,0.6]\} = 0.4 \cdot (1-0.6) = 0.16$ (too conservative)

# Policies in uMDPs

- For uMDPs, as for MDPs, we can define
  - policies $\pi : (S \times A)^* \times S \to A$, or
  - memoryless policies $\pi_m : S \to A$
  - (depending on the set $\mathscr{P}$, some care is needed to make sure policies can be applied)




- For policy $\pi \in \Pi$ and transition probabilities $P \in \mathscr{P}$:
  - we can define probability space $Pr_s^{\pi,P}$, random variables $\mathbb{E}_s^{\pi,P}(X)$ and value functions $V^{\pi,P}(s)$
  - which correspond to the MDP $\mathscr{M}[P]$

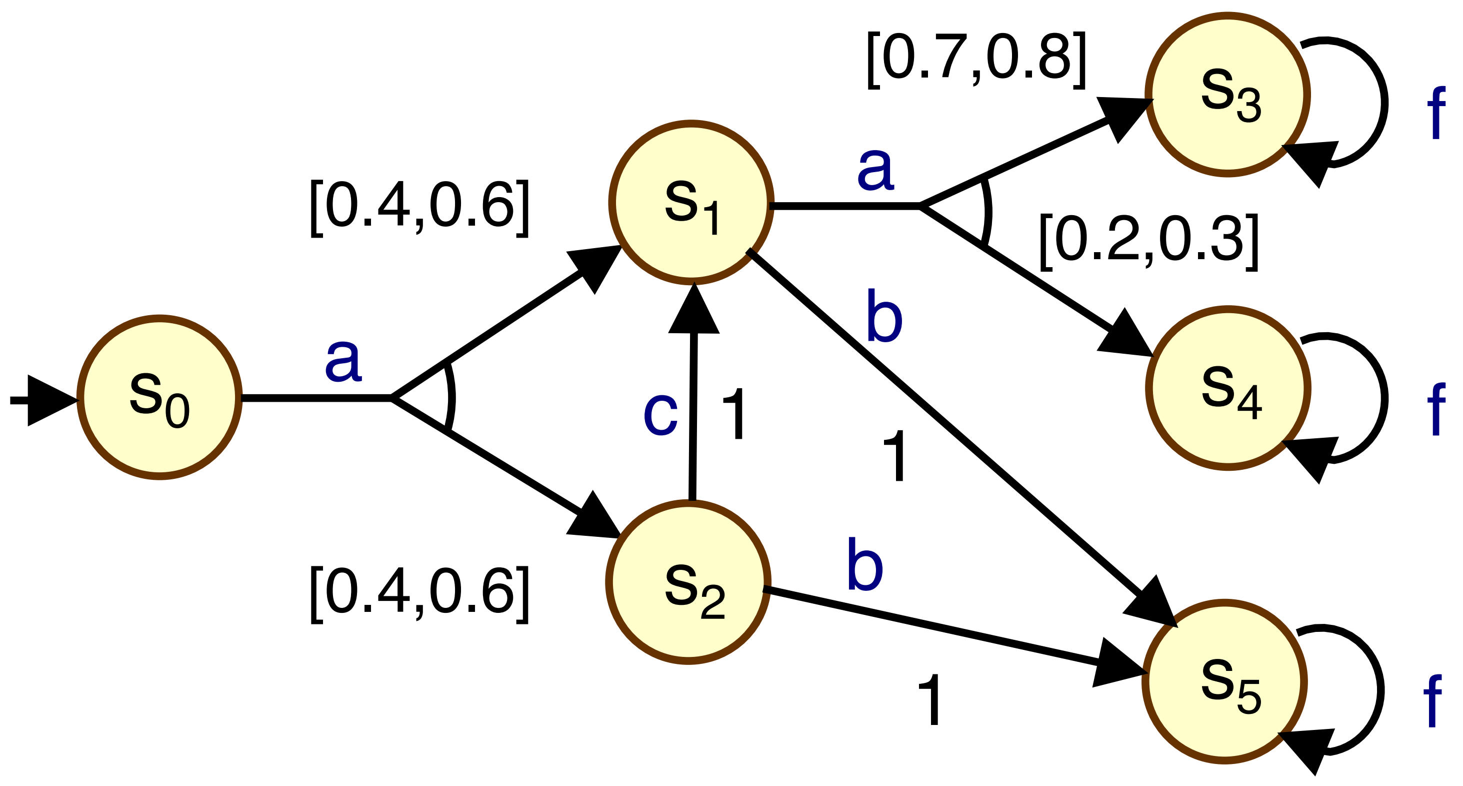


# Robust control

- For now, we consider a robust view of uncertainty
  - i.e., we focus on worst-case (adversarial, pessimistic) scenarios

- Robust policy evaluation:
  - worst-case scenario for (maximising) policy $\pi$, i.e.: $\min_{P \in \mathscr{P}} V^{\pi,P}(s)$

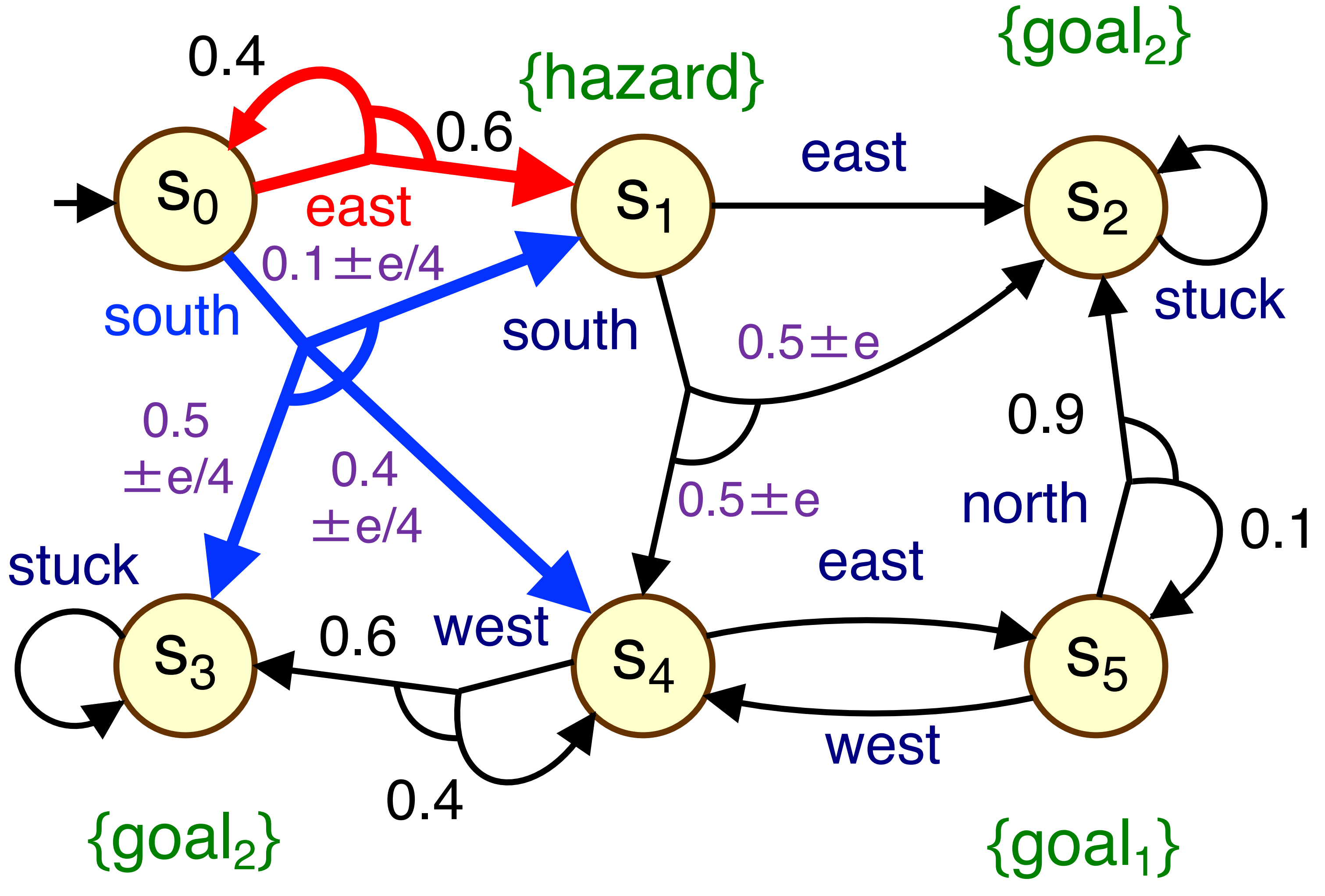
- Robust control (policy optimisation):
  - optimal worst-case value $V^*(s) = \max_{\pi \in \Pi} \min_{P \in \mathscr{P}} V^{\pi,P}(s)$
  - optimal worst-case policy $\pi^* = \mathrm{argmax}_{\pi \in \Pi} \min_{P \in \mathscr{P}} V^{\pi,P}(s)$




- Other cases:
  - for a minimising objective (e.g. SPP), we use: $V^*(s) = \min_{\pi \in \Pi} \max_{P \in \mathscr{P}} V^{\pi,P}(s)$
  - we may also consider optimistic scenarios, e.g. $V^*(s) = \max_{\pi \in \Pi} \max_{P \in \mathscr{P}} V^{\pi,P}(s)$

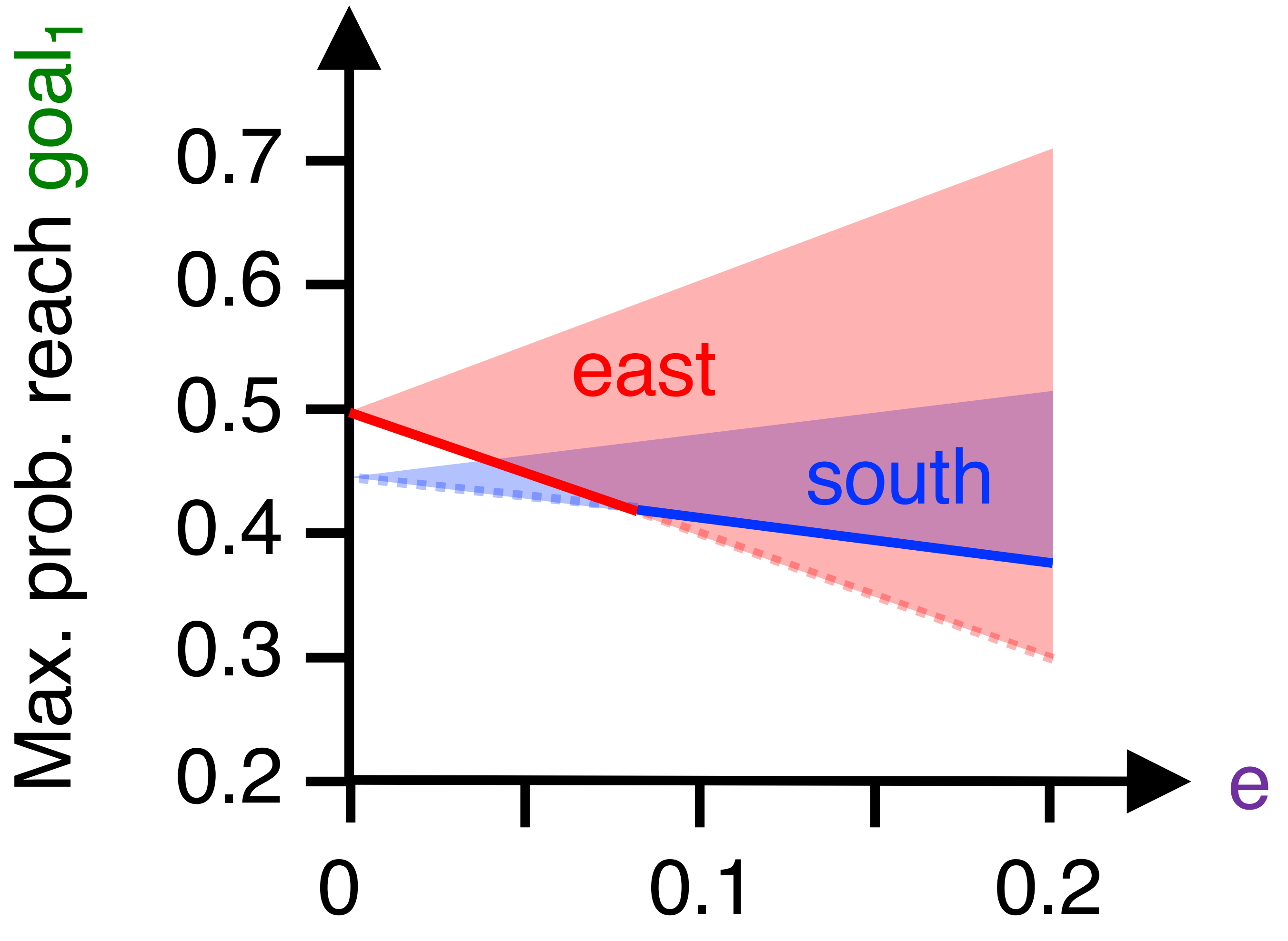
# Running example: Robust control

- An IMDP for the robot example
  - uncertainty added to two state-action pairs



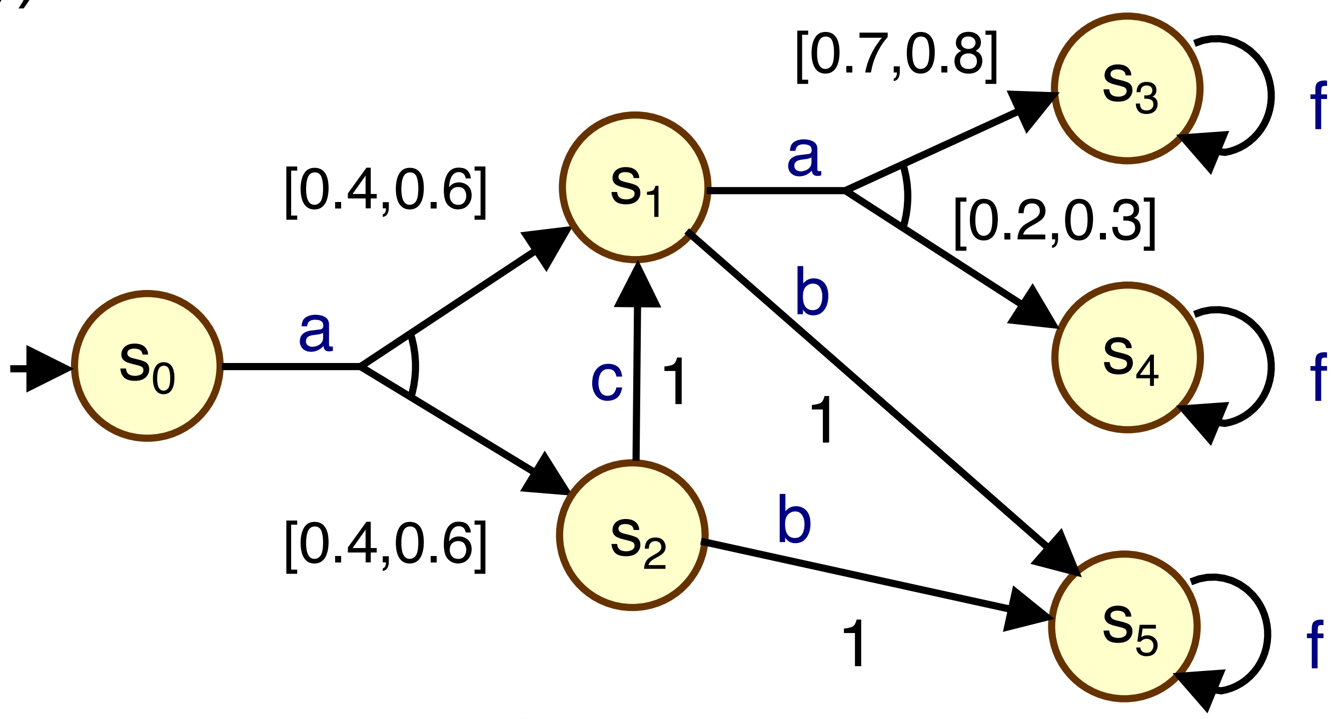

  - Note: the degree of uncertainty (e) in states $s_1$ and $s_2$ is correlated here (but the actual transition probabilities are not)

- Robust control
  - for any e, we can pick a "robust" (optimal worst-case) policy
  - and give a safe lower bound on its performance

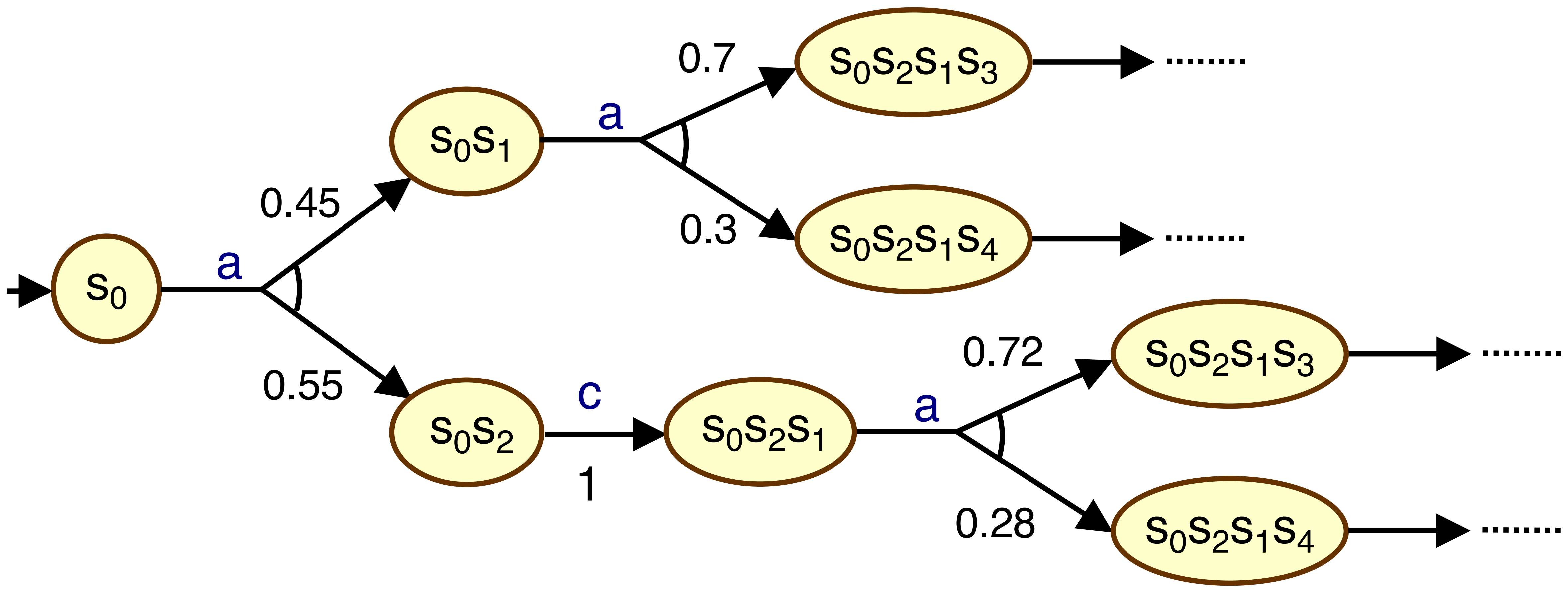


# Resolving uncertainty

- Now we consider a more dynamic approach to resolving uncertainty
  - (which we will need to extend dynamic programming to this setting)




- An environment policy (or nature policy, or adversary) $\tau \in \mathscr{T}$
  - is a mapping $\tau : (S \times A)^* \times (S \times A) \to Dist(S)$
  - such that $\tau(s_0, a_0, \ldots, s_n, a_n) \in \mathscr{P}_s^a$
  - note: this assumes (s,a)-rectangularity!




- Policies $\pi, \tau$ yield
  - a probability space $Pr_s^{\pi,\tau}$
  - random variables $\mathbb{E}_s^{\pi,\tau}(X)$
  - and value functions $V^{\pi,\tau}$

# Summary (part 2)

- Stochastic games
  - unknown parts of the system can be modelled adversarially
  - zero-sum turn-based (or concurrent) stochastic games
    - dynamic programming (value iteration) generalises
- Uncertain MDPs
  - MDPs plus epistemic uncertainty: set of transition functions
  - rectangularity (dependencies)
  - control policies + robust control
  - environment policies

# References (part 2)


- Stochastic games
  - J. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer, 1997
  - M. Kwiatkowska, G. Norman and D. Parker, *Probabilistic Model Checking and Autonomy,* Annual Review of Control, Robotics, and Autonomous Systems, 5, 2022

- Uncertain MDPs and interval MDPs
  - G. N. Iyengar, Robust dynamic programming, *Mathematics of Operations Research*, 30(2), 2005
  - A. Nilim and L. Ghaoui, Robust control of Markov decision processes with uncertain transition matrices, *Operations Research*, 53(5), 780–798, 2005
  - E. Wolff, U. Topcu, and R. Murray, Robust control of uncertain Markov decision processes with temporal logic specifications, In *Proc. 51th IEEE Conference on Decision and Control (CDC'12)*, 2012
  - W. Wiesemann, D. Kuhn and B. Rustem, Robust Markov Decision Processes, *Math. Oper. Res.,* 38(1), 153-183, 2013
  - A. Puggelli, W. Li, A. Sangiovanni-Vincentelli and S. Seshia, Polynomial-time verification of PCTL properties of MDPs with convex uncertainties, In *Proc. 25th International Conference on Computer Aided Verification (CAV'13)*, LNCS, vol. 8044, Springer, 2013