Lecture 10
Markov Decision Processes

Alessandro Abate

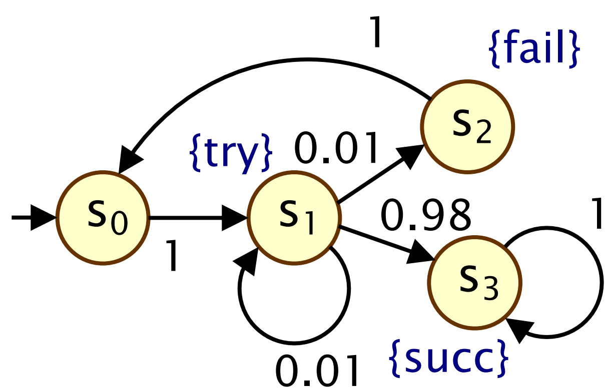
Department of Computer Science
University of Oxford

Overview

- Nondeterminism
- Markov decision processes (MDPs)
- Paths, probabilities and adversaries
- End components: long-run behaviour

Recap: DTMCs

- Discrete-time Markov chains (DTMCs)
  - discrete state space, transitions are discrete timesteps
  - from each state, choice of successor state (i.e. which transition) is determined by a discrete probability distribution



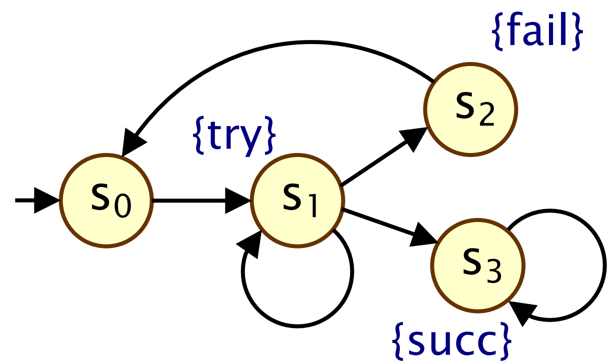
- DTMCs are fully probabilistic
  - well suited to modelling, for example, simple random algorithm or synchronous probabilistic system where components move in lock-step

Nondeterminism

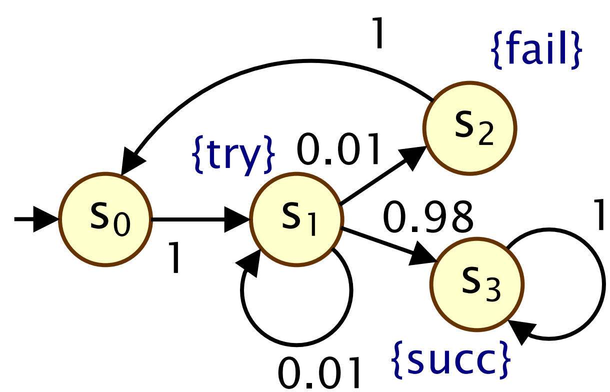
- But, some aspects of a system may not be probabilistic and should not be modelled probabilistically; for example:
- Concurrency – scheduling/composition of parallel components
  - e.g. randomised distributed algorithm– multiple probabilistic processes operating asynchronously
- Unknown environments
  - e.g. probabilistic security protocols – unknown adversary
- Underspecification – unknown model parameters
  - e.g. a probabilistic communication protocol designed for message propagation delays of between d_{min} and d_{max}
- Abstraction
  - e.g. partition DTMC into similar (but not identical) states
- Planning/Strategy Synthesis
  - Nondeterminism as action choices for an agent

Probability vs. nondeterminism

- Labelled transition system
  - (S,s_0,R,L) where $R \subseteq S \times S$
  - choice is nondeterministic



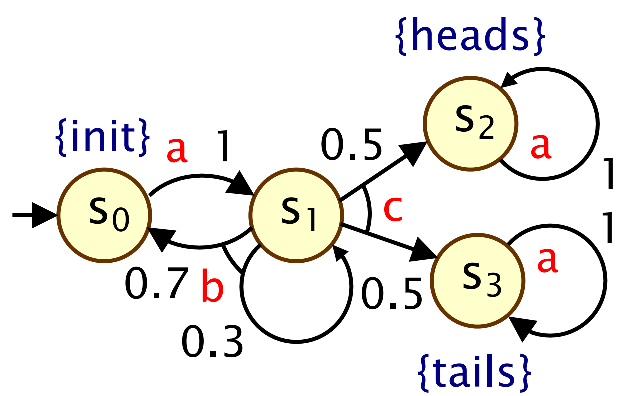
- Discrete-time Markov chain
  - (S,s_0,P,L) where $P : S \times S \rightarrow [0,1]$
  - choice is probabilistic



- How to combine the two?

Markov decision processes

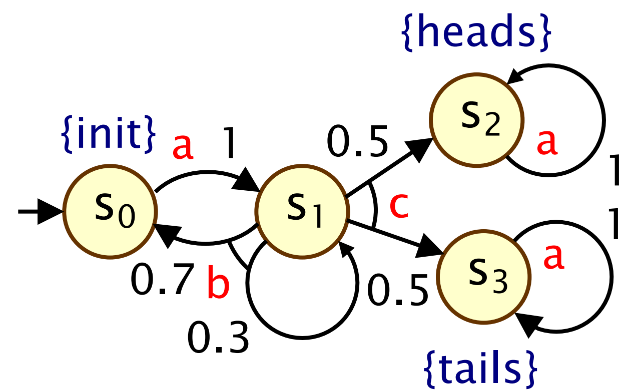
- Markov decision processes (MDPs)
  - extension of DTMCs allowing nondeterministic choices
- Like DTMCs:
  - discrete set of states representing possible configurations of the system being modelled
  - transitions between states occur in discrete time steps
- Probabilities and nondeterminism
  - in each state, a nondeterministic choice between several discrete probability distributions over successor states



Markov decision processes

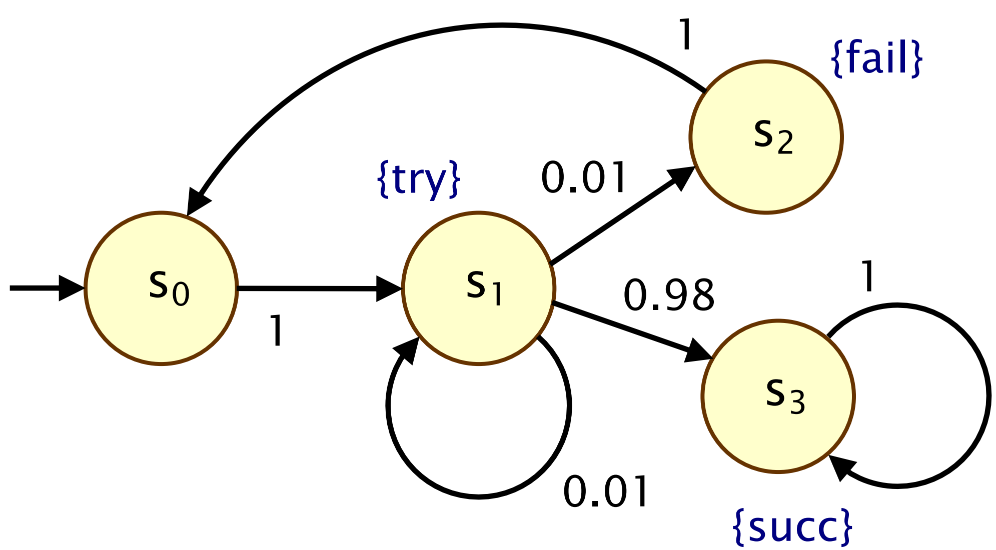
- Formally, an MDP M is a tuple $(S, s_{init}, \mathbf{Steps}, L)$ where:
  - S is a finite set of states (“state space”)
  - $s_{init} \in S$ is the initial state
  - **Steps** : $S \rightarrow 2^{Act \times Dist(S)}$ is the transition probability function, where Act is a set of actions and Dist(S) is the set of discrete probability distributions over the set S
  - $L : S \rightarrow 2^{AP}$ is a labelling with atomic propositions

- Notes:
  - Steps(s) is always non-empty, i.e. no deadlocks
  - the use of actions to label distributions can be omitted



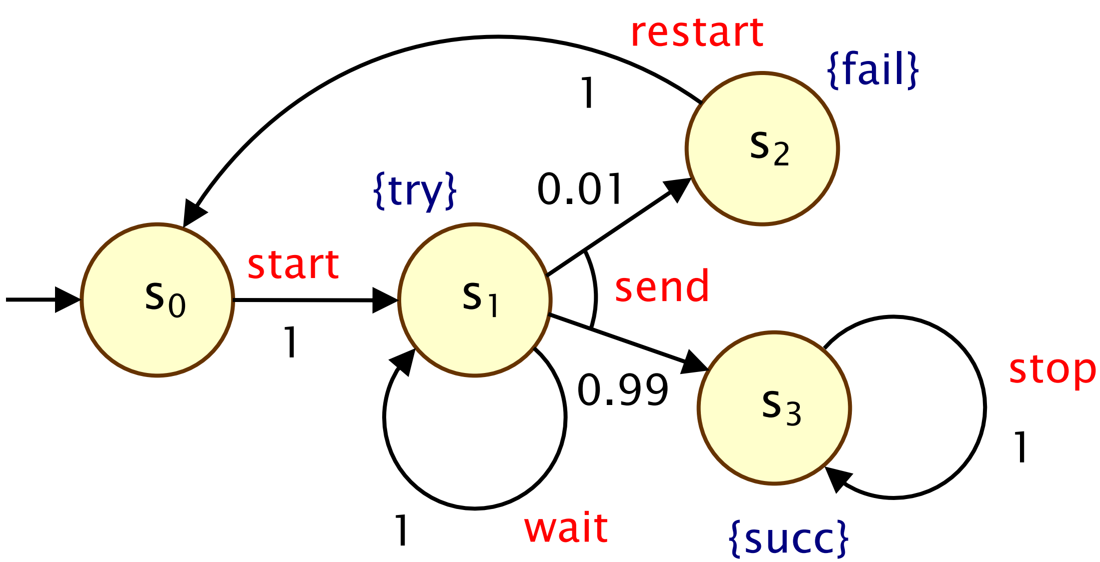
Simple DTMC example

- Modelling a very simple communication protocol
  - after one step, process starts trying to send a message
  - with probability 0.01, channel unready so wait a step
  - with probability 0.98, send message successfully and stop
  - with probability 0.01, message sending fails, restart



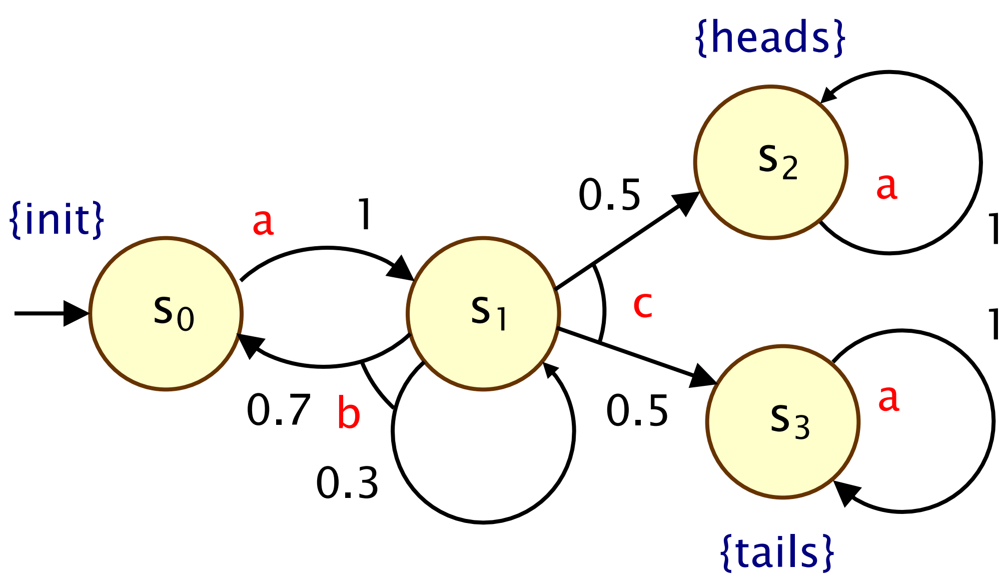
Simple MDP example

- Modification of the simple DTMC communication protocol
  - after one step, process starts trying to send a message
  - then, a nondeterministic choice between: (a) waiting a step because the channel is unready; (b) sending the message
  - if the latter, with probability 0.99 send successfully and stop
  - and with probability 0.01, message sending fails, restart



Simple MDP example 2

- Another simple MDP example with four states
  - from state s_0, move directly to s_1 (action a)
  - in state s_1, nondeterministic choice between actions b and c
  - action b gives a probabilistic choice: self-loop or return to s_0
  - action c gives a 0.5/0.5 random choice between heads/tails



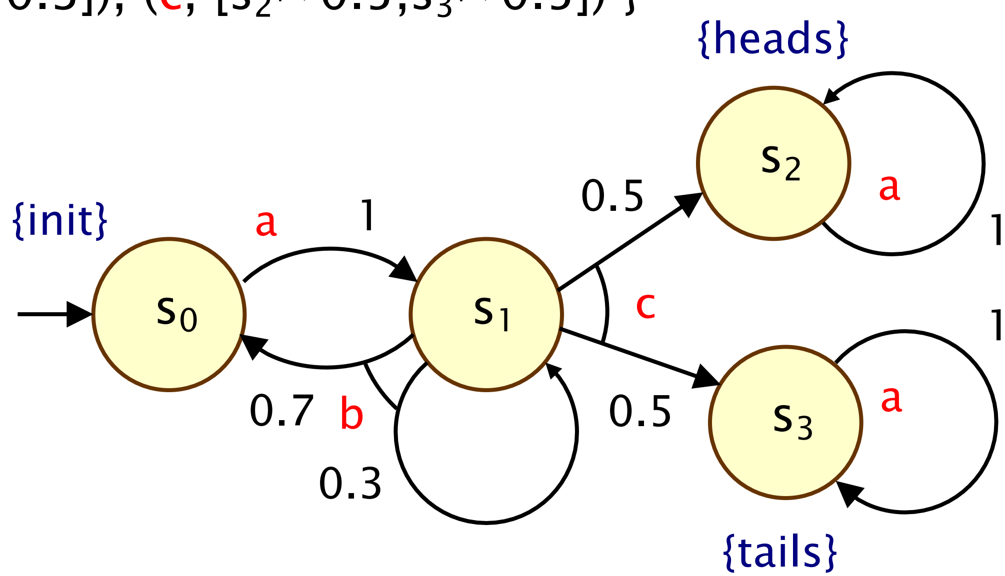
Simple MDP example 2

$M = (S, s_{init}, \mathbf{Steps}, L)$

$S = \{s_0, s_1, s_2, s_3\}$
$s_{init} = s_0$

AP = {init,heads,tails}
$L(s_0)$={init},
$L(s_1)=\varnothing$,
$L(s_2)$={heads},
$L(s_3)$={tails}

$\mathbf{Steps}(s_0) = \{ (a, [s_1 \mapsto 1]) \}$
$\mathbf{Steps}(s_1) = \{ (b, [s_0 \mapsto 0.7, s_1 \mapsto 0.3]), (c, [s_2 \mapsto 0.5, s_3 \mapsto 0.5]) \}$
$\mathbf{Steps}(s_2) = \{ (a, [s_2 \mapsto 1]) \}$
$\mathbf{Steps}(s_3) = \{ (a, [s_3 \mapsto 1]) \}$



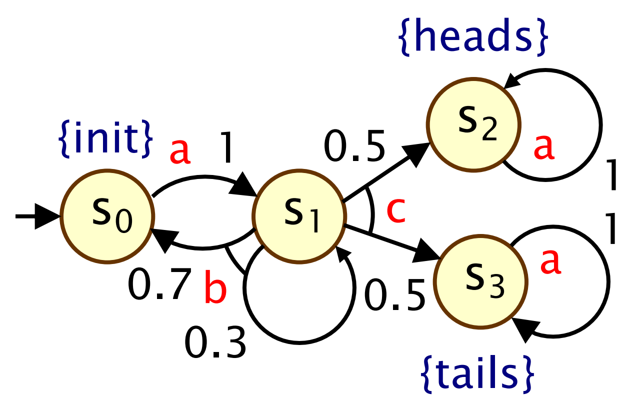
The transition probability function

- It is often useful to think of the function **Steps** as a matrix
  - non-square matrix with $|S|$ columns and $\Sigma_{s \in S}$ $|$**Steps**$(s)|$ rows

- Example (for clarity, we omit actions from the matrix)

Steps$(s_0) = \{ (a, s_1 \mapsto 1) \}$
Steps$(s_1) = \{ (b, [s_0 \mapsto 0.7, s_1 \mapsto 0.3]), (c, [s_2 \mapsto 0.5, s_3 \mapsto 0.5]) \}$
Steps$(s_2) = \{ (a, s_2 \mapsto 1) \}$
Steps$(s_3) = \{ (a, s_3 \mapsto 1) \}$

$$\mathbf{Steps} = \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ \hline 0.7 & 0.3 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array} \right]$$



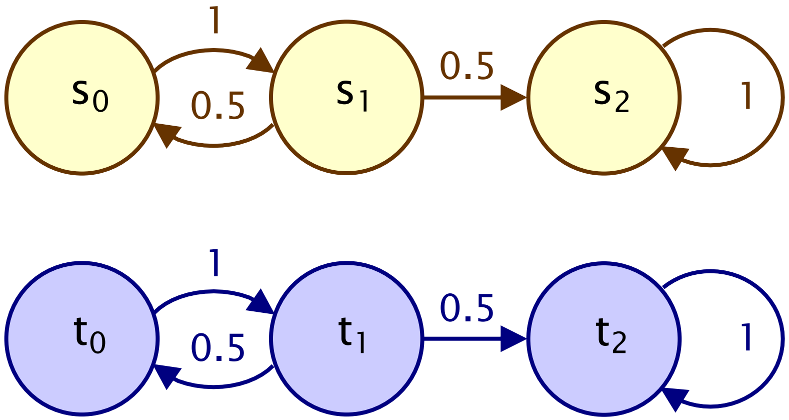
Example - Parallel composition

Asynchronous parallel composition of two 3-state DTMCs

PRISM code:

```
module M1
    s : [0..2] init 0;
    [] s=0 -> (s'=1);
    [] s=1 -> 0.5:(s'=0) + 0.5:(s'=2);
    [] s=2 -> (s'=2);
endmodule
```

```
module M2 = M1 [ s=t ] endmodule
```

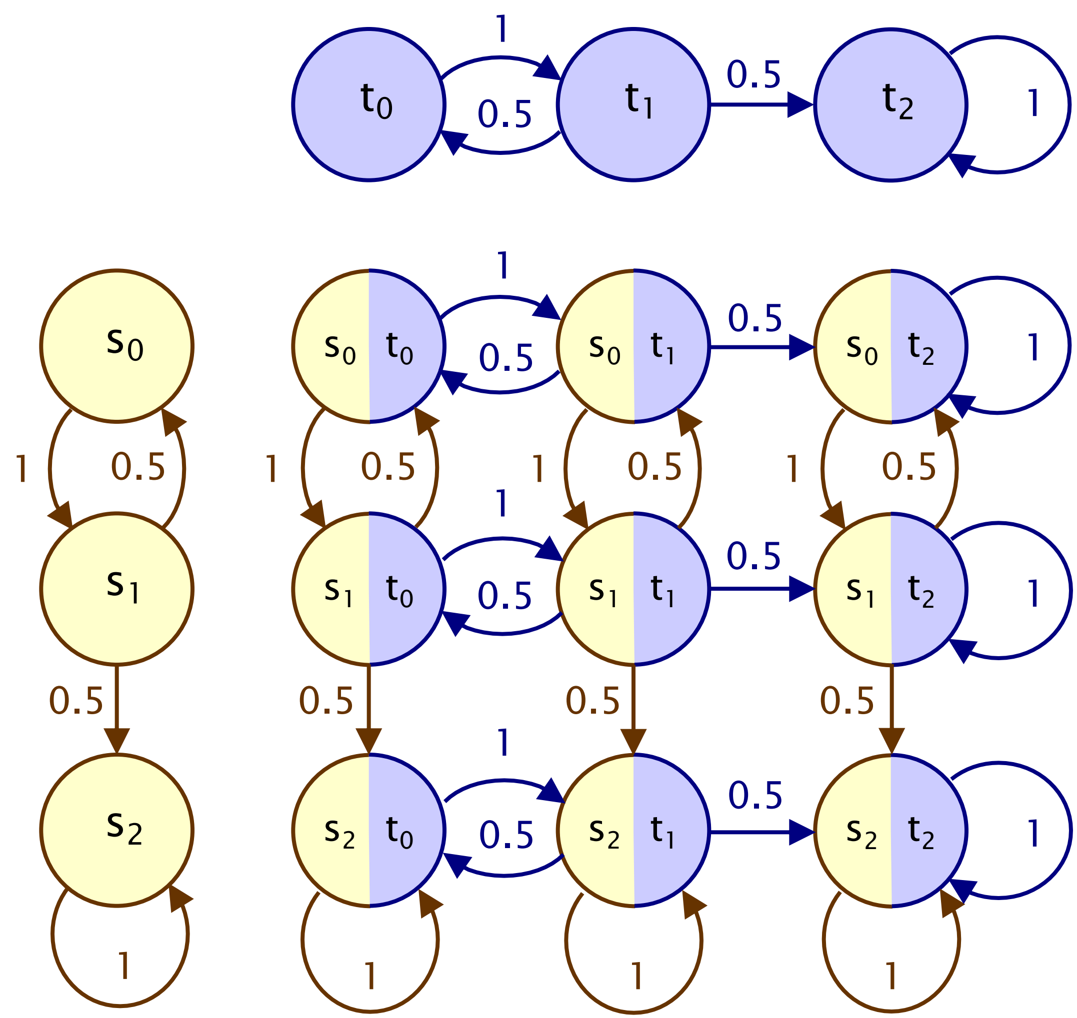


Note: no actions needed for each DTMC/module

Example – Parallel composition

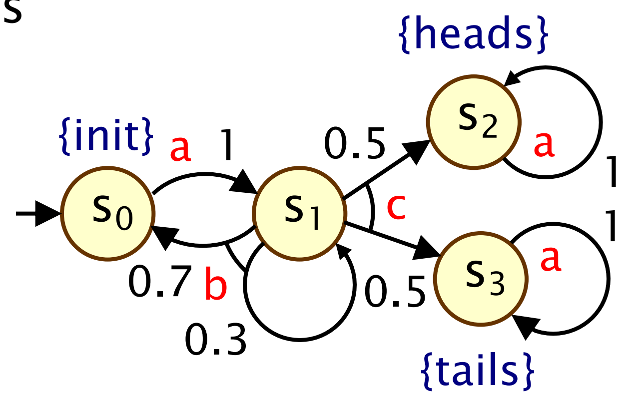
Asynchronous parallel composition of two 3-state DTMCs

Actions now needed in composition (action labels omitted here)



Paths and probabilities

- A (finite or infinite) path through an MDP
  - is a sequence of states and action/distribution pairs
  - e.g. $s_0(a_0,\mu_0)s_1(a_1,\mu_1)s_2\ldots$
  - such that $(a_i,\mu_i) \in \textbf{Steps}(s_i)$ and $\mu_i(s_{i+1}) > 0$ for all $i \geq 0$
  - represents an execution (i.e. one possible behaviour) of the system that the MDP is modelling
- Path(s) = set of all paths through MDP starting in state s
  - $\text{Path}_{\text{fin}}(s)$ = set of all finite paths from s
- Paths resolve both nondeterministic and probabilistic choices
  - how to reason about probabilities?



Adversaries

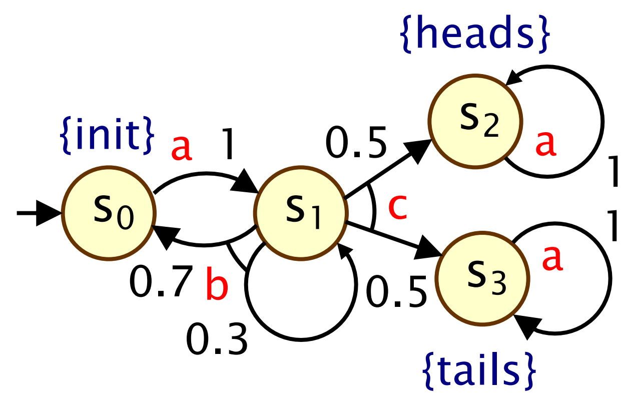
- To consider the probability of some behaviour of the MDP
  - first need to resolve the nondeterministic choices
  - this results in a DTMC, for which we can define a probability measure over paths

- An adversary resolves nondeterministic choice in an MDP
  - also known as "scheduler", "policy", "strategy", "controller"
- Formally:
  - an adversary σ for an MDP M is a function mapping every finite path $\omega = s_0(a_0,\mu_0)s_1 \ldots s_n$ to an element $\sigma(\omega)$ of **Steps**(s_n)
  - i.e. resolves nondeterminism based on execution history

- Adv (or Adv_M) denotes the set of all adversaries

Adversaries – Examples

- Consider the previous example MDP
  - note that s_1 is the only state for which $|\mathbf{Steps}(s)| > 1$
  - i.e. s_1 is the only state for which an adversary makes a choice
  - let μ_b and μ_c denote the probability distributions associated with actions b and c in state s_1

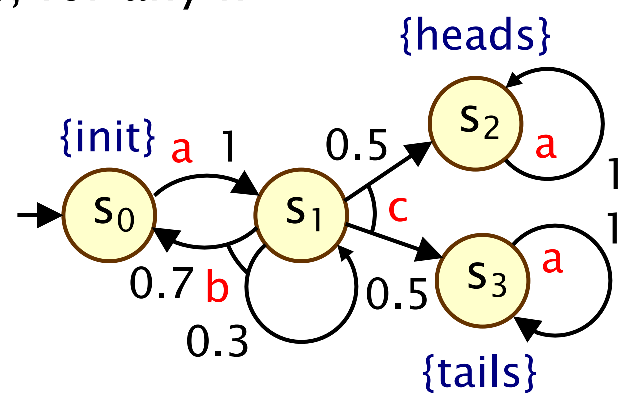


- Adversary σ_1
  - picks action c the first time
  - $\sigma_1(s_0 s_1)=(c,\mu_c)$
- Adversary σ_2
  - picks action b the first time, then c
  - $\sigma_2(s_0 s_1)=(b,\mu_b)$, $\sigma_2(s_0 s_1 s_1)=(c,\mu_c)$, $\sigma_2(s_0 s_1 s_0 s_1)=(c,\mu_c)$, …

(note: actions/distributions omitted from paths for clarity)

Adversaries and paths

- $Path^{\sigma}(s) \subseteq Path(s)$
  - (infinite) paths from s where nondeterminism resolved by σ
  - i.e. paths $s_0(a_0,\mu_0)s_1(a_1,\mu_1)s_2\ldots$
  - for which $\sigma(s_0(a_0,\mu_0)s_1\ldots s_n)) = (a_n,\mu_n)$, for any n



- Adversary σ_1
  - (picks action c the first time)
  - $Path^{\sigma_1}(s_0) = \{ s_0s_1s_2^{\omega}, s_0s_1s_3^{\omega} \}$

- Adversary σ_2
  - (picks action b the first time, then c)
  - $Path^{\sigma_2}(s_0) = \{ s_0s_1s_0s_1s_2^{\omega}, s_0s_1s_0s_1s_3^{\omega}, s_0s_1s_1s_2^{\omega}, s_0s_1s_1s_3^{\omega} \}$

Induced DTMCs

- Adversary σ for MDP induces an infinite-state DTMC D^σ

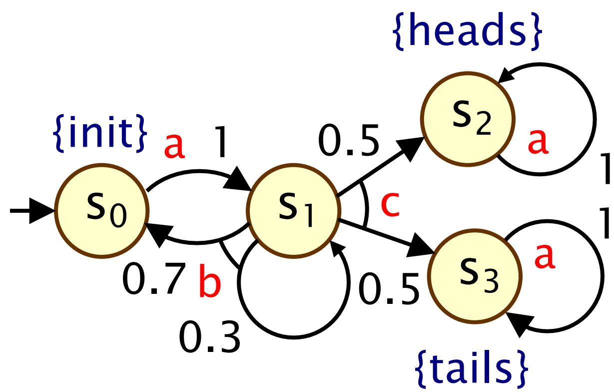
- $D^\sigma = (Path^\sigma_{fin}(s), s, \mathbf{P}^\sigma_s)$ where:
  - states of the DTMC are the finite paths of σ starting in state s
  - initial state is s (the path starting in s of length 0)
  - $\mathbf{P}^\sigma_s(\omega,\omega')=\mu(s')$ if $\omega' = \omega(a, \mu)s'$ and $\sigma(\omega)=(a,\mu)$
  - $\mathbf{P}^\sigma_s(\omega,\omega')=0$ otherwise
  - (labels omitted for simplicity)

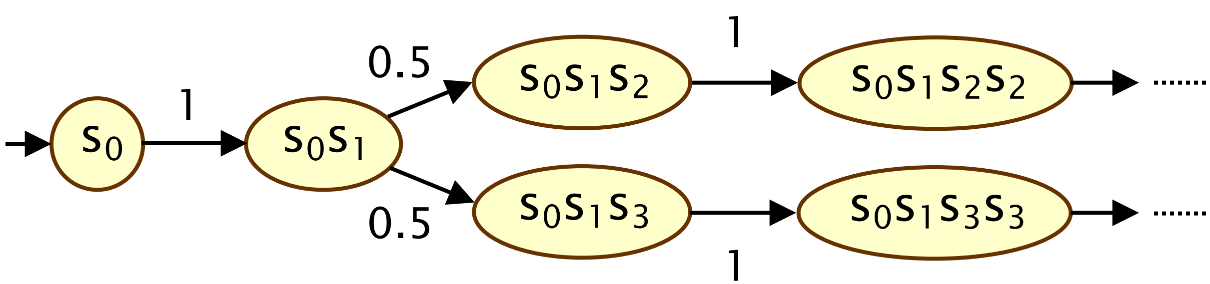
- 1-to-1 correspondence between $Path^\sigma(s)$ and paths of D^σ

- This gives us a probability measure Pr^σ_s over $Path^\sigma(s)$
  - from probability measure over paths of D^σ

Adversaries – Examples

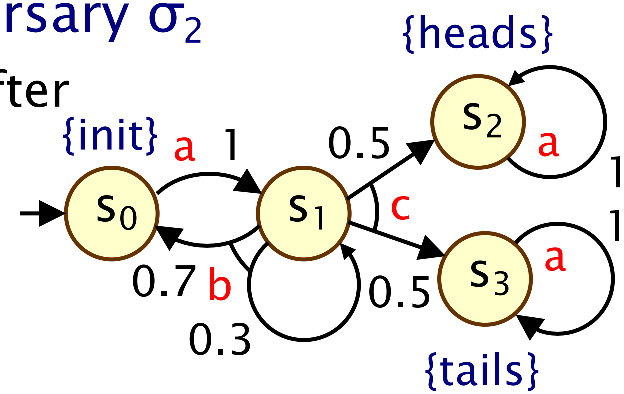
- Fragment of induced DTMC for adversary σ_1
  - σ_1 picks action c the first time

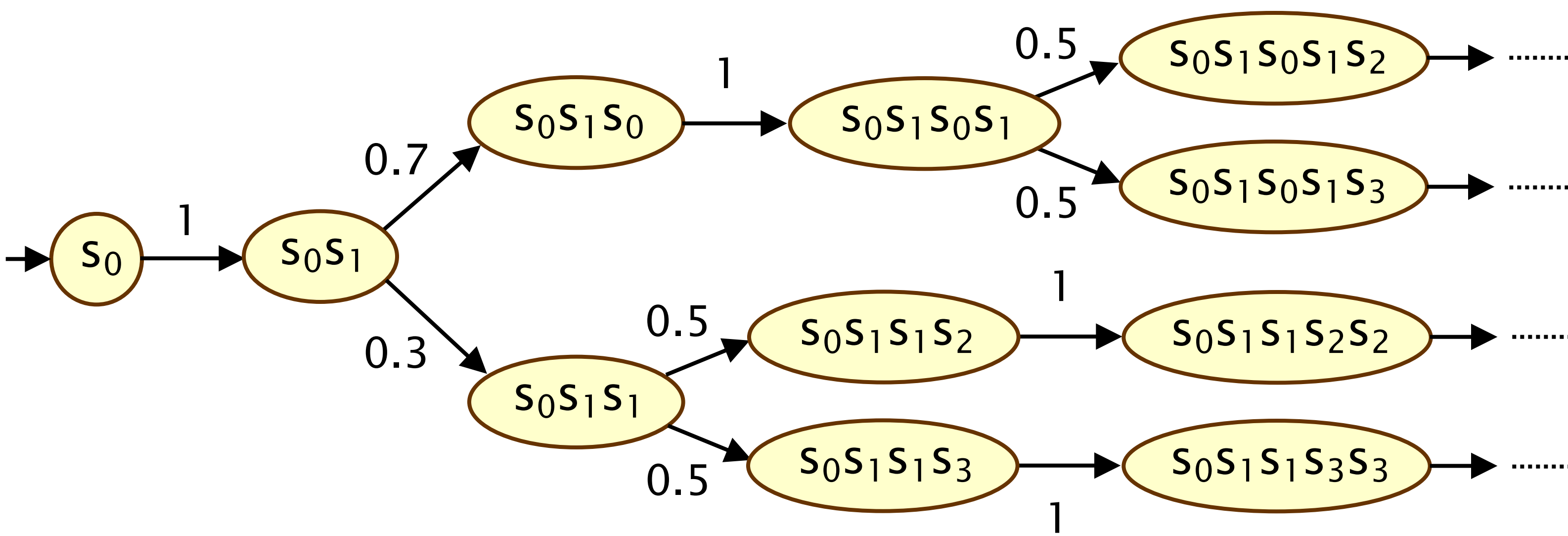




Adversaries – Examples

- Fragment of induced DTMC for adversary σ_2
  - σ_2 picks first action b, then c thereafter





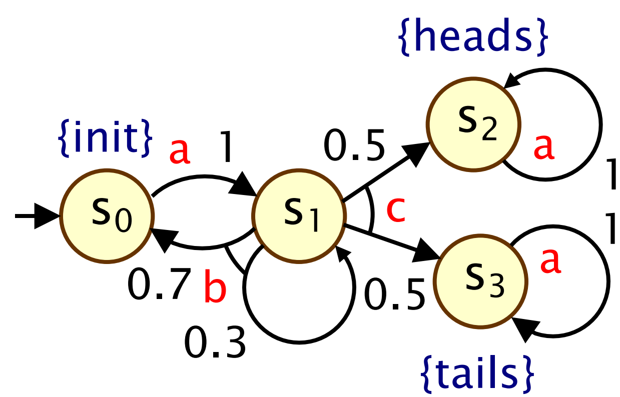
MDPs and probabilities

- $\mathrm{Prob}^{\sigma}(s, \psi) = \mathrm{Pr}^{\sigma}_{s}\ \{\ \omega \in \mathrm{Path}^{\sigma}(s) \mid \omega \vDash \psi\ \}$
  - for some path formula ψ
  - e.g. $\mathrm{Prob}^{\sigma}(s, \mathrm{F\ tails})$

- MDP provides best-/worst-case analysis
  - based on lower/upper bounds on probabilities
  - over all possible adversaries

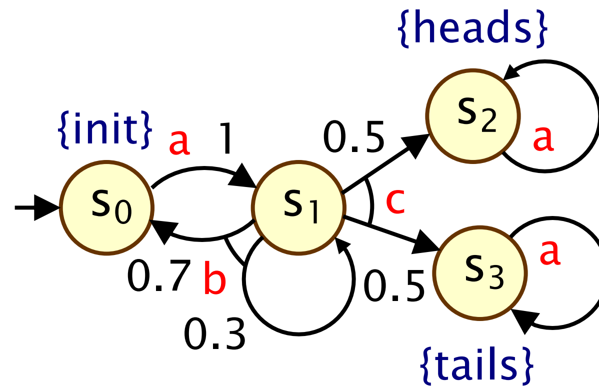
$$p_{min}(s,\psi) = \inf_{\sigma \in Adv} \mathrm{Prob}^{\sigma}(s,\psi)$$

$$p_{max}(s,\psi) = \sup_{\sigma \in Adv} \mathrm{Prob}^{\sigma}(s,\psi)$$

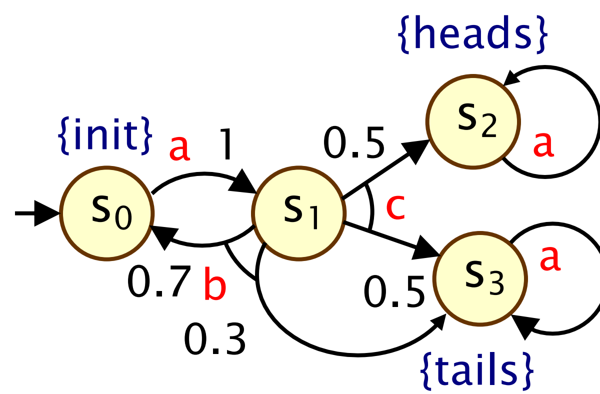


Examples

- $Prob^{\sigma 1}(s_0, F\ tails) = 0.5$
- $Prob^{\sigma 2}(s_0, F\ tails) = 0.5$
  - (where σ_i picks b i-1 times then c)
- …
- $p_{max}(s_0, F\ tails) = 0.5$
- $p_{min}(s_0, F\ tails) = 0$

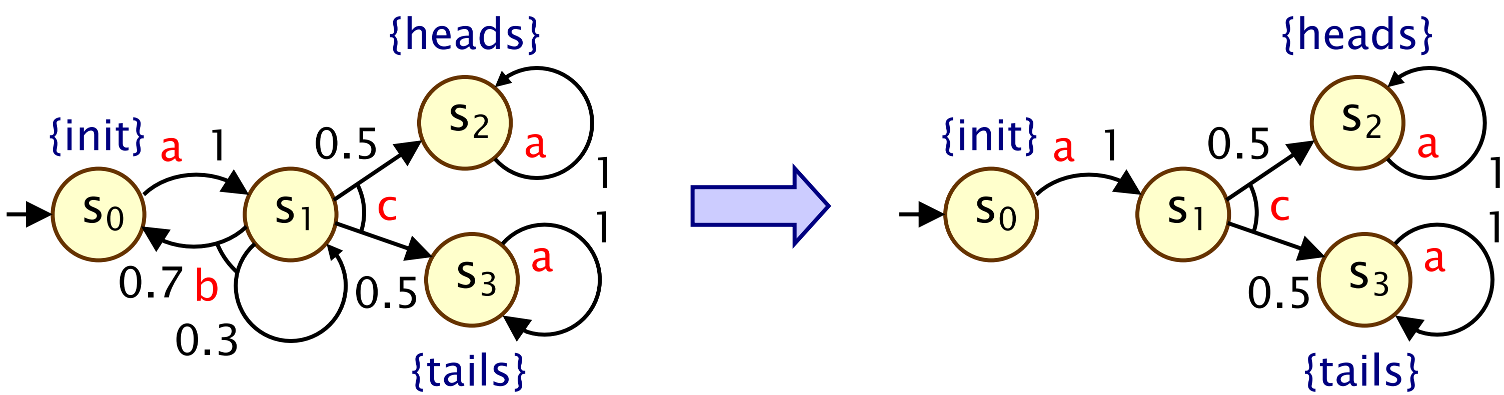


- $Prob^{\sigma 1}(s_0, F\ tails) = 0.5$
- $Prob^{\sigma 2}(s_0, F\ tails)$
  $= 0.3+0.7\cdot 0.5 = 0.65$
- $Prob^{\sigma 3}(s_0, F\ tails)$
  $= 0.3+0.7\cdot 0.3+0.7\cdot 0.7\cdot 0.5 = 0.755$
- …
- $p_{max}(s_0, F\ tails) = 1$
- $p_{min}(s_0, F\ tails) = 0.5$



Memoryless adversaries

- Memoryless adversaries always pick same choice in a state
  - also known as: Markov, simple, positional, stationary
  - formally, $\sigma(s_0(a_0,\mu_0)s_1 \ldots s_n)$ depends only on s_n
  - can write as a mapping from states, i.e. $\sigma(s)$ for each $s \in S$
  - induced DTMC can be mapped to a $|S|$-state DTMC
- From previous example:
  - adversary σ_1 (picks c in s_1) is memoryless; σ_2 is not

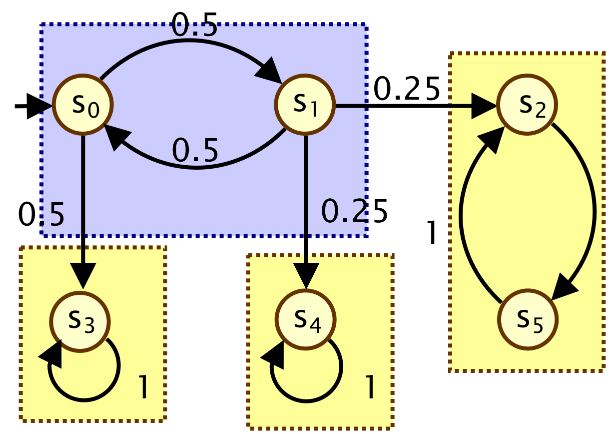


Other classes of adversaries

- Finite-memory adversary
  - finite number of modes, which can govern choices made
  - formally defined by a *deterministic finite automaton*
  - induced DTMC (for finite MDP) again mapped to finite DTMC
- Randomised adversary
  - maps finite paths $s_0(a_1,\mu_1)s_1 \ldots s_n$ in MDP to a probability distribution over element of **Steps**(s_n)
  - generalises deterministic schedulers
  - still induces a (possibly infinite-state) DTMC
- Fair adversary
  - fairness assumptions on resolution of nondeterminism

Recall: fundamental property of DTMCs

- Strongly connected component (SCC)
  - maximally strongly connected set of states
- Bottom strongly connected component (BSCC)
  - SCC T from which no state outside T is reachable from T
- With probability 1, a BSCC will be reached and all of its states visited infinitely often



- Formally:
  - Pr_s { ω ∈ Path(s) | ∃ i≥0, ∃ BSCC T such that ∀ j≥i ω(j) ∈ T and ∀ s'∈T ω(k) = s' for infinitely many k } = 1

Qualitative repeated reachability: DTMCs

- $Pr_s\{\ \omega \in Path(s) \mid \forall\ i\geq 0\ .\ \exists\ j\geq i\ .\ \omega(j) \in Sat(a)\ \} = 1$
- $P_{\geq 1}$ [GF a]

PCTL*

if and only if

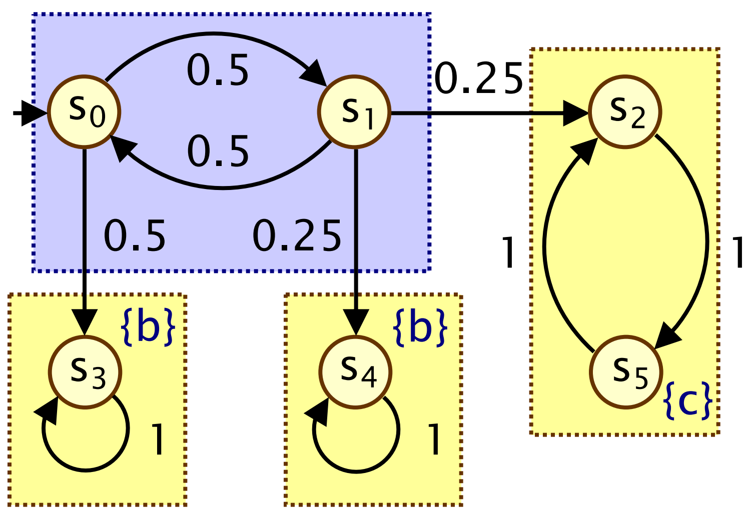
- $T \cap Sat(a) \neq \emptyset$ for all BSCCs T reachable from s

Examples:

$s_0 \models P_{\geq 1}$ [GF (b∨c)]

$s_0 \not\models P_{\geq 1}$ [GF b]

$s_2 \models P_{\geq 1}$ [GF c]



Qualitative persistence: DTMCs

- $\mathrm{Pr}_s\{\ \omega \in \mathrm{Path}(s) \mid \exists\ i{\geq}0\ .\ \forall\ j{\geq}i\ .\ \omega(j) \in \mathrm{Sat}(a)\ \} = 1$
- $\mathrm{P}_{\geq 1}$ [FG a]

if and only if

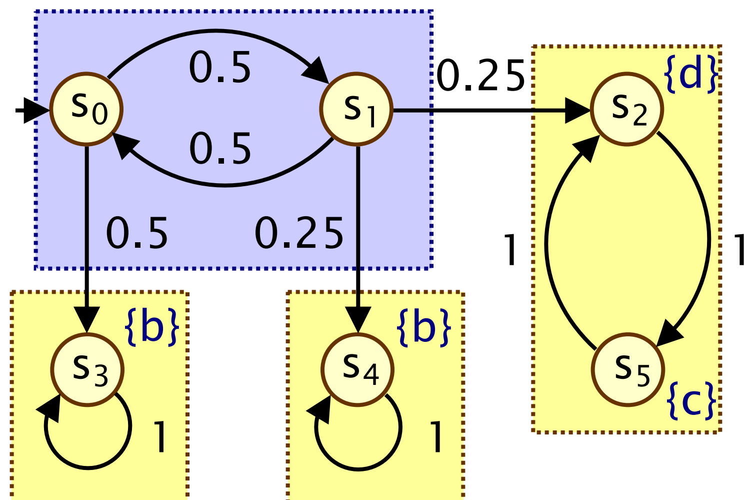
- $T \subseteq \mathrm{Sat}(a)$ for all BSCCs T reachable from s

Examples:

$s_0 \not\models \mathrm{P}_{\geq 1}$ [FG (b∨c)]

$s_0 \models \mathrm{P}_{\geq 1}$ [FG (b∨c∨d)]

$s_2 \models \mathrm{P}_{\geq 1}$ [FG (c∨d)]



Repeated reachability + persistence

- Repeated reachability and persistence are dual requirements
  - $GF\ a \equiv \neg(FG\ \neg a)$, $FG\ a \equiv \neg(GF\ \neg a)$
- Hence, for example:
  - $Prob(s, GF\ a) = 1 - Prob(s, FG\ \neg a)$

- $Prob(s, GF\ a) + Prob(s, FG\ \neg a)$

$= Prob(s, F\ T_{GFa}) + Prob(s, F\ T_{FG\neg a})$

  - T_{GFa} = union of BSCCs T with $T \cap Sat(a) \neq \varnothing$ (T intersects Sat(a))
  - $T_{FG\neg a}$ = union of BSCCs T with $T \subseteq (S \setminus Sat(a))$ (no intersection)

$= Prob(s, F\ (T_{GFa} \cup T_{FG\neg a})) = 1$ (fundamental DTMC property)

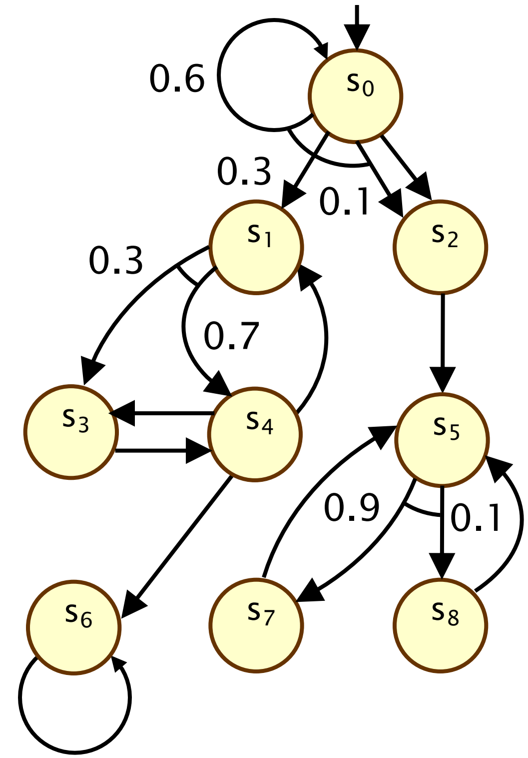
- Can we generalise this statement to MDPs?

End components of MDPs

- Consider an MDP $M = (S, s_{init}, \mathbf{Steps}, L)$

- A sub-MDP of M is a pair $(T, \mathbf{Steps}')$ where:
  - $T \subseteq S$ is a (non-empty) subset of M's states
  - $\mathbf{Steps}'(s) \subseteq \mathbf{Steps}(s)$ for each $s \in T$
  - $(T, \mathbf{Steps}')$ is closed under probabilistic branching, i.e. the set of states $\{ s' \mid \mu(s')>0 \text{ for some } (a,\mu) \in \mathbf{Steps}'(s) \}$ is a subset of T

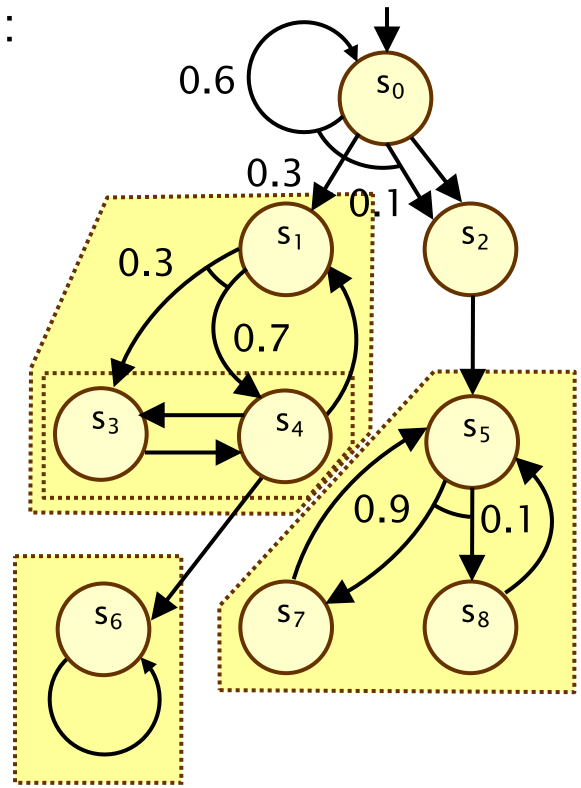
- An end component of M is a strongly connected sub-MDP



Notes:
- action labels omitted
- probabilities omitted where =1

End components – Examples

- Sub-MDPs
  - can be formed from state sets such as:
  - $\{s_2,s_5,s_7,s_8\}$, $\{s_0,s_2,s_5,s_7,s_8\}$, $\{s_5,s_7,s_8\}$,
  - $\{s_1,s_3,s_4\}$, $\{s_1,s_3,s_4,s_6\}$, $\{s_3,s_4\}$, …
- End components
  - can be formed from state sets:
  - $\{s_3,s_4\}$, $\{s_1,s_3,s_4\}$, $\{s_6\}$, $\{s_5,s_7,s_8\}$
- Note that
  - state sets do not necessarily uniquely identify end components
  - e.g. $\{s_1,s_3,s_4\}$

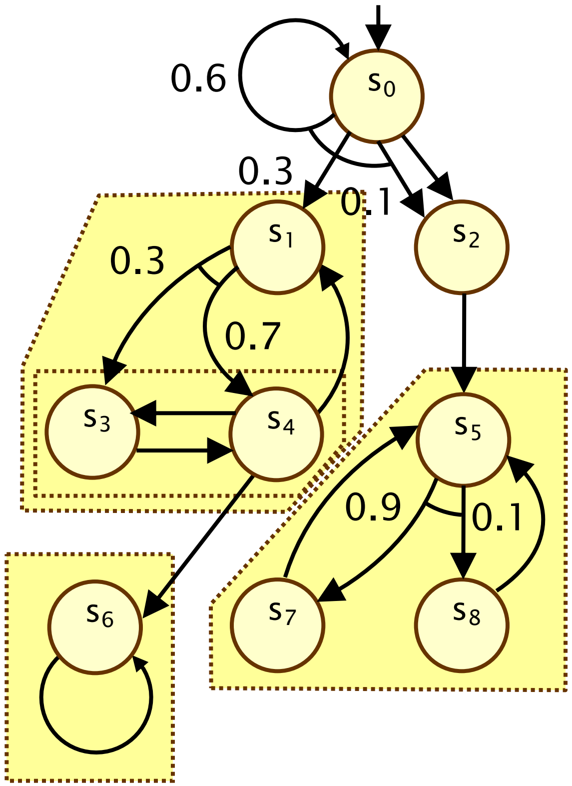


Fundamental property of MDPs

- For finite MDPs…
  - (analogue of fundamental property of finite DTMCs)

1. For every end component, there is a (finite-memory) adversary σ which, with probability 1, forces the MDP starting in the end component to remain there and visit all its states infinitely often

2. Under any adversary σ, with probability 1 an end component will be reached



Qualitative repeated reachability – MDPs

- Repeated reachability (GF) for MDPs
  - special case of more general limiting properties
  - need to distinguish between max and min
  - consider first the case of maximum probabilities…
  - $p_{max}(s, GF\ a)$

- First, a simple qualitative property:
  - $Prob^{\sigma}(s, GF\ a) > 0$ for some adversary σ, i.e. $p_{max}(s, GF\ a) > 0$

    $\Leftrightarrow$
  - $T \cap Sat(a) \neq \emptyset$ for some end component T reachable from s

- Can reason via reachability ($F\ T_{GFa}$), as earlier for DTMCs
  - see next slide for justification…

Repeated reachability – MDPs (max)

- For the qualitative property given earlier:
  - $\mathrm{Prob}^{\sigma}(s, \mathrm{GF}\ a) > 0$ for some adversary σ
    - $\Leftrightarrow\ p_{max}(s, \mathrm{GF}\ a) > 0$
    - $\Leftrightarrow\ p_{max}(s, \mathrm{F}\ T_{GFa}) > 0$
    - $\Leftrightarrow\ \mathrm{Prob}^{\sigma}(s, \mathrm{F}\ T_{GFa}) > 0$ for some adversary σ
    - $\Leftrightarrow\ s \vDash \mathrm{EF}\ T_{GFa}$
    - $\Leftrightarrow\ T \cap Sat(a) \neq \varnothing$ for some E.C. T reachable from s

- Another qualitative property:
  - $\mathrm{Prob}^{\sigma}(s, \mathrm{GF}\ a) = 1$ for some adversary σ
    - $\Leftrightarrow\ p_{max}(s, \mathrm{GF}\ a) = 1$
    - $\Leftrightarrow\ p_{max}(s, \mathrm{F}\ T_{GFa}) = 1$

Repeated reachability – MDPs (min)

- Repeated reachability for MDPs – minimum probabilities
  - $p_{min}(s, GF\ a)$

- First, a useful qualitative property:

  - $Prob^{\sigma}(s, GF\ a) = 1$ for all adversaries σ

    $\Leftrightarrow$

  - $s \vDash P_{\geq 1}$ [GF a] PCTL*

    $\Leftrightarrow$

  - $T \cap Sat(a) \neq \varnothing$ for all end components T reachable from s

Summing up…

- Nondeterminism
  - concurrency, unknown environments/parameters, abstraction
- Markov decision processes (MDPs)
  - discrete-time + probability and nondeterminism
  - nondeterministic choice between multiple distributions
- Adversaries
  - resolution of nondeterminism only
  - induced set of paths and (infinite state DTMC)
  - induces DTMC yields probability measure for adversary
  - best-/worst-case analysis: minimum/maximum probabilities
  - memoryless adversaries
- Long-run behaviour
  - Limiting properties via reachability of end components