Sequential decisions

- A sequential decision problem is a sequence of decisions, where for each decision we consider:
  - what actions are available to the agent
  - what information is, or will be, available to the agent when it will perform the action
  - effects of the actions
  - desirability of the actions

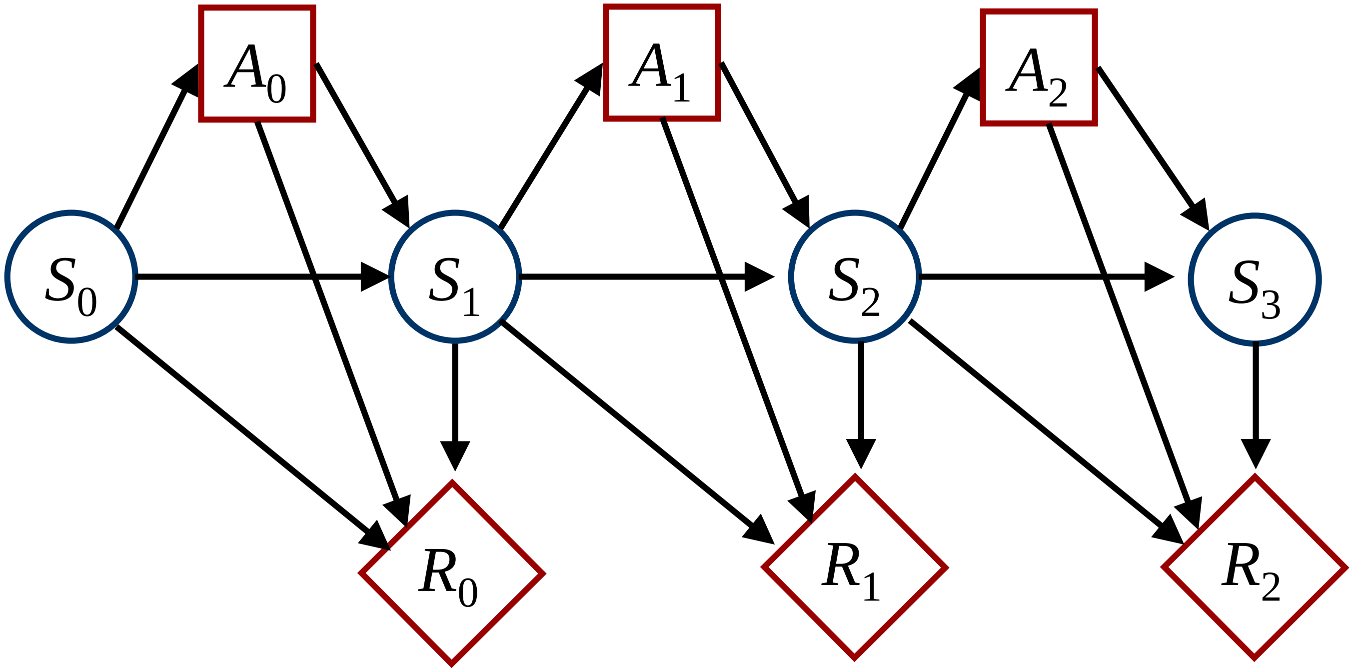
Decision processes

- Indefinite and infinite horizon problems
  - ongoing processes or it is unknown how many actions are required
- Wide range of applications
  - robotics (e.g., control)
  - investments (e.g., portfolio management)
  - computational linguistics (e.g., dialogue management)
  - operations research (e.g., inventory management)

Markov decision process (MDP)

- Definition
  - Set of states: S
  - Set of actions (i.e., decisions): A
  - Transition model: $P(S_t \mid A_{t-1}, S_{t-1})$
  - Reward model (i.e., utility): $R(S_t, A_{t-1}, S_{t-1})$
  - Discount factor: $0 \leq \gamma \leq 1$
  - Horizon (i.e., # of time steps): h
- Goal: find optimal policy

Decision network representing a finite part of an MDP



Transition model

- Markov assumption

  $$P(S_{t+1} \mid S_t, \ldots, S_0) \;=\; P(S_{t+1} \mid S_t)$$

- Stationary: the transition probabilities are the same for each time point

  $P(S_0)$ specifies initial conditions

  $P(S_{t+1} \mid A_t, S_t)$ specifies the dynamics, which is the same for each $t \geq 0$

Reward model

- Why so many utility nodes in decision network?
- $U(S_0, S_1, S_2, \ldots)$
  - infinite process $\rightarrow$ infinite utility function
- Solution: additive preferences
  - $R(S_t, A_{t-1}, S_{t-1})$
    immediate reward from doing action A_{t-1} and transitioning from state S_{t-1} to state S_t
  - $U(S_0, S_1, S_2, \ldots) = \Sigma_t R(S_t, A_{t-1}, S_{t-1})$

Discounted Rewards

- If process infinite, isn't Σ_t R(S_t, $A_{t\text{-}1}$, $S_{t\text{-}1}$) infinite?

- Solution: discounted rewards
  - Discount factor: $0 \leq \gamma \leq 1$
  - Finite utility: $\Sigma_t \gamma^t$ R(S_t, $A_{t\text{-}1}$, $S_{t\text{-}1}$) is a geometric sum
  - γ is like an inflation rate
  - Intuition: prefer utility sooner than later

Policy

- Choice of action at each time step
- Formally:
  - Mapping from states to actions
  - i.e., $\delta(\mathit{state}) = \mathit{action}$
  - Assumption: fully observable states
    - allows next action to be chosen only based on current state
- Optimal policy:
  - Policy δ^* with highest expected utility
  - $EU(\delta) \leq EU(\delta^*)$ for all δ

Example: Inventory Management

- Markov decision process
  - States: inventory levels
  - Actions: {doNothing, orderWidgets}
  - Transition model: stochastic demand
  - Reward model: Sales – Costs (Storage)
  - Discount factor: 0.999
  - Horizon: ∞
- Tradeoff: increasing supplies decreases odds of missed sales but increases storage costs

Other representations for decision processes

- Dynamic decision networks
  - MDP is a state-based representation
  - here: describe states in terms of random variables
- Partially observable Markov decision process (POMDP)
  - states are not fully observable
  - partial/noisy observations of the state