Decision-Theoretic Planning: Markov Decision Processes (MDPs)

Computer Science cpsc322, Lecture 36

(Textbook Chpt 9.5)

April, 6, 2009

Combining ideas for Stochastic planning

- What is a key limitation of decision networks?

Represent (and optimize) only a fixed number of decisions

- What is an advantage of Markov models?

The network can extend indefinitely

Goal: represent (and optimize) an indefinite sequence of decisions

Planning in Stochastic Environments

Problem		Environment: Deterministic	Environment: Stochastic
Static	Constraint Satisfaction	Arc Consistency *Vars + Constraints* Search → for CSP SLS	
	Query	*Logics* CSP Search → for Inference	***Belief Nets*** Var. Elimination *Markov Chains and HMMs*
Sequential	Planning	***STRIPS*** → CSP Search → for complex planning	*Decision Nets* Var. Elimination *Markov Decision Processes* Value Iteration

Representation

Reasoning Technique

Lecture Overview

- Recap: Markov Models
- Decision Processes: MDP
- MDP Example
- Reward and Optimal Policies

Markov Models

Simpler

Markov Chains

Noisy observations

Hidden Markov Model

Partially observable
POMDP

CPSC 422

Noisy Actions Rewards

Markov Decision Processes (MDPs)

Recap: Markov Models

t_0 t_1

Tables

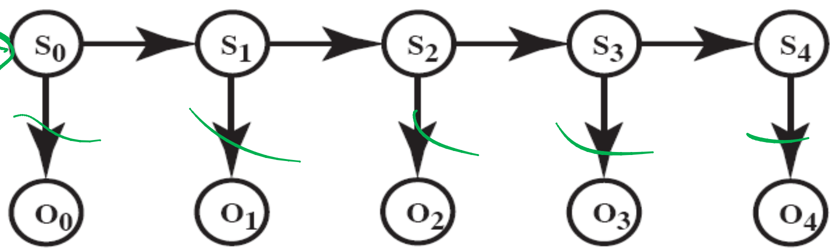
$|dom(S)| = n$

$P(S_0)$ 1:n

$P(S_{t+1} | S_t)$ n:n

$|dom(O)| = K$

HMM

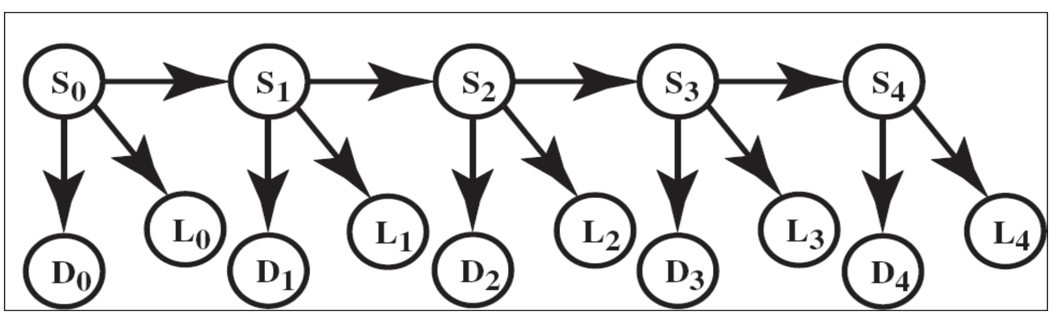


$P(O_t | S_t)$ n:K

extended HMM

multiple sensors

"sensor fusion"



Lecture Overview

- Recap: Markov Models
- **Decision Processes: MDP**
- MDP Example
- Reward and Optimal Policies

Decision Processes

Often an agent needs to go beyond a fixed set of decisions – Examples?

- Would like to have an **ongoing decision process**

Infinite horizon problems: process does not stop

robot surviving on a planet

Indefinite horizon problem: the agent does not know when the process may stop

reaching location

Finite horizon: the process must end at a give time N

in N steps

How can we deal with indefinite/infinite processes?

We make the same two assumptions we made for….

The action outcome depends only on the current state Markov

$P(S_{t+1} \mid S_t, A_t, S_{t-1}, A_{t-1}, \ldots)$

Let S_t be the state at time t …

$P(S_{t+1} \mid S_t, A_t)$

The process is *stationary*… the same for all t

We also need a more flexible specification for the utility. How?

- Defined based on a reward/punishment *R(s)* that the agent receives in each state *s*

$s_0 \; s_1 \; \ldots \; s_n$

e.g. $\sum \; r_0 \; r_1 \; \ldots \; r_n$

MDP: formal specification

For an MDP you specify:

- set S of states and set A of actions
- the process' dynamics (or ***transition model***)

  $P(S_{t+1}|S_t, A_t)$
- The **reward function**

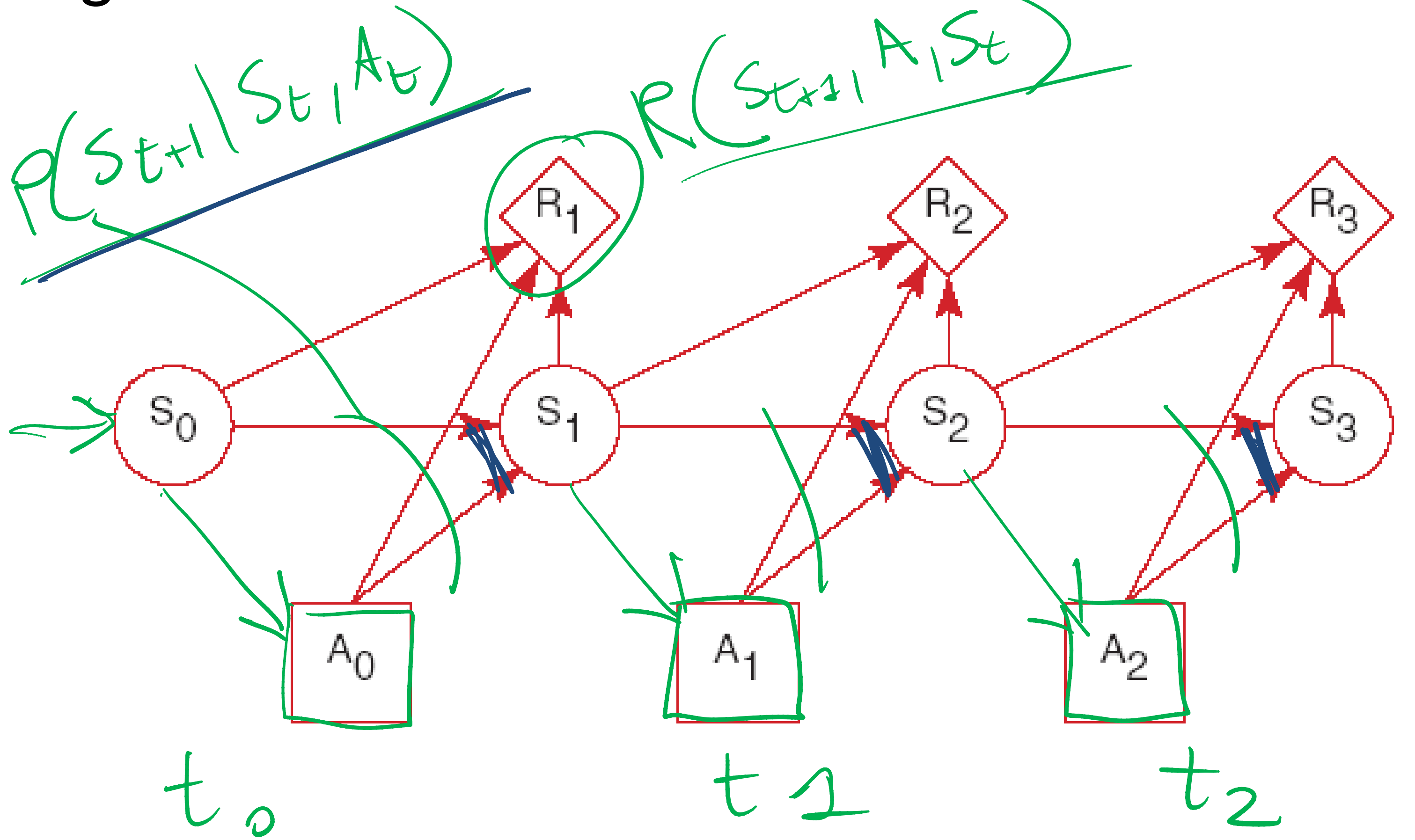
  $R(s, a, s')$

  describing the reward that the agent receives when it performs action *a* in state *s* and ends up in state *s'*
- R(s) is used when the reward depends only on the state s and not on how the agent got there
- **Absorbing/stopping/terminal state** S_{ab}

for all action $P(S_{ab} | a, S_{ab}) = 1$ $R(S_{ab}, a, S_{ab}) = 0$

MDP graphical specification

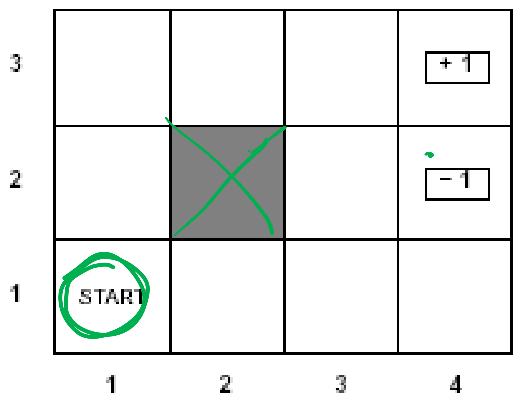
Basically a MDP augments a ……Markov Chain……
augmentetd with actions and values

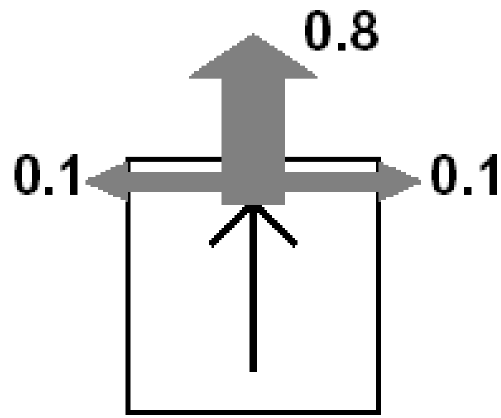


Lecture Overview

- Recap: Markov Models
- Decision Processes: MDP
- MDP Example
- Reward and Optimal Policies

Example MDP: Scenario and Actions





Agent moves in the above grid via actions *Up, Down, Left, Right*

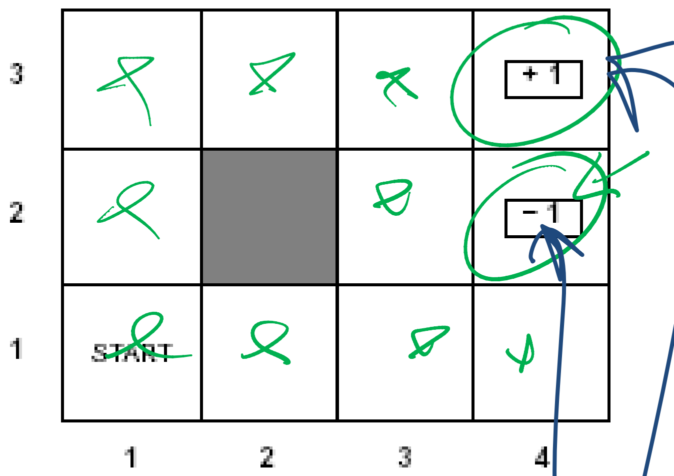
Each action has:

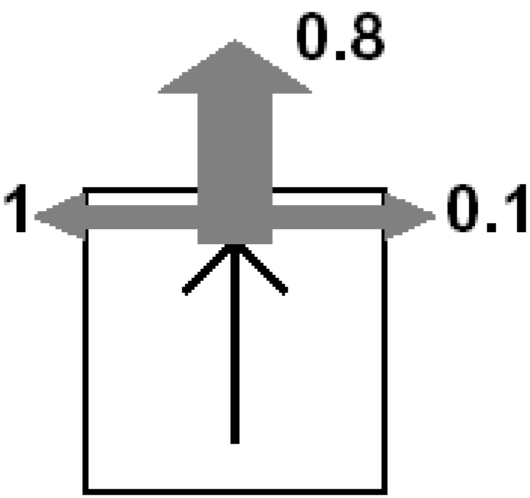
- 0.8 probability to reach its intended effect
- 0.1 probability to move at right angles of the intended direction
- If the agents bumps into a wall, it says there

How many states? 11 (11, 12, 13 . . . .)

There are two terminal states (3,4) and (2,4)

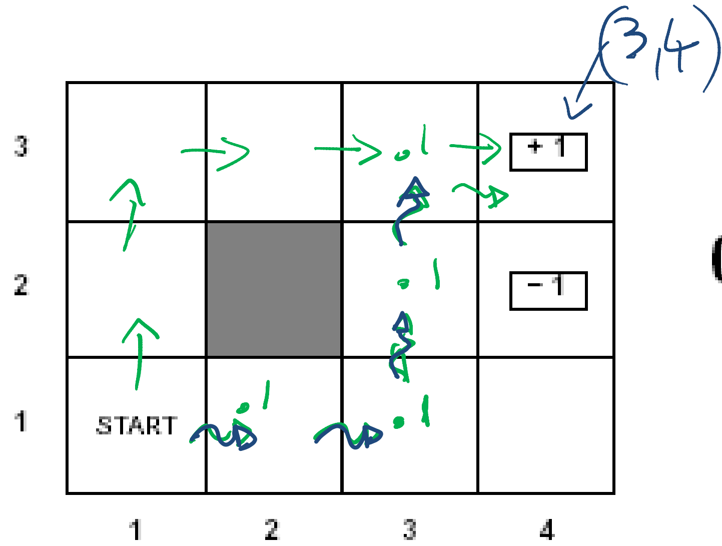
Example MDP: Rewards

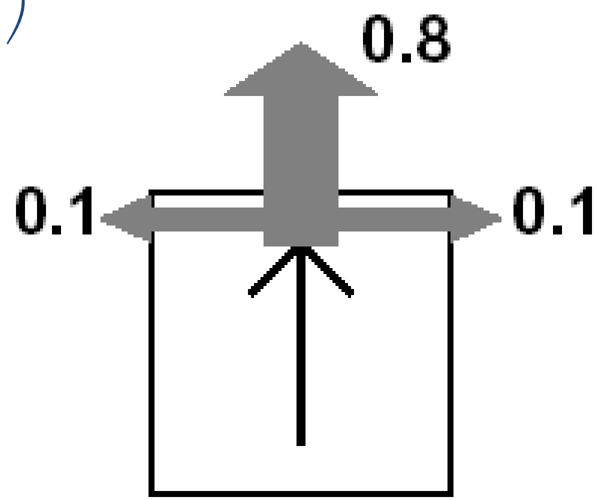




$$R(s) = \begin{cases} -0.04 & \text{(small penalty) for nonterminal states} \\ \pm 1 & \text{for terminal states} \end{cases}$$

Example MDP: Sequence of actions





Can the sequence [*Up, Up, Right, Right, Right*] take the agent in terminal state (3,4)?

$(.8)^5$

Can the sequence reach the goal in any other way?

$(.1)^4 \; .8$ ← with prob

yes

Lecture Overview

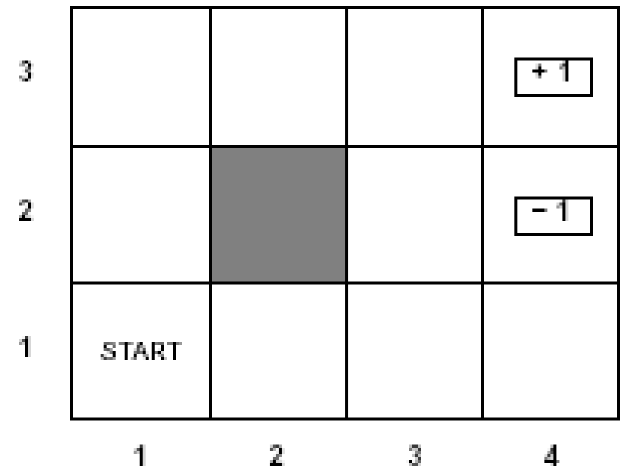
- Recap: Markov Models
- Decision Processes: MDP
- MDP Example STOP HERE TODAY
- Reward and Optimal Policies

MDPs: Policy

- So what is the best sequence of actions for our Robot?
- There is no best sequence of actions!
- As we saw for decision networks, our aim is to find an **optimal policy**: a set of δ_1 ,……, δ_n decision functions

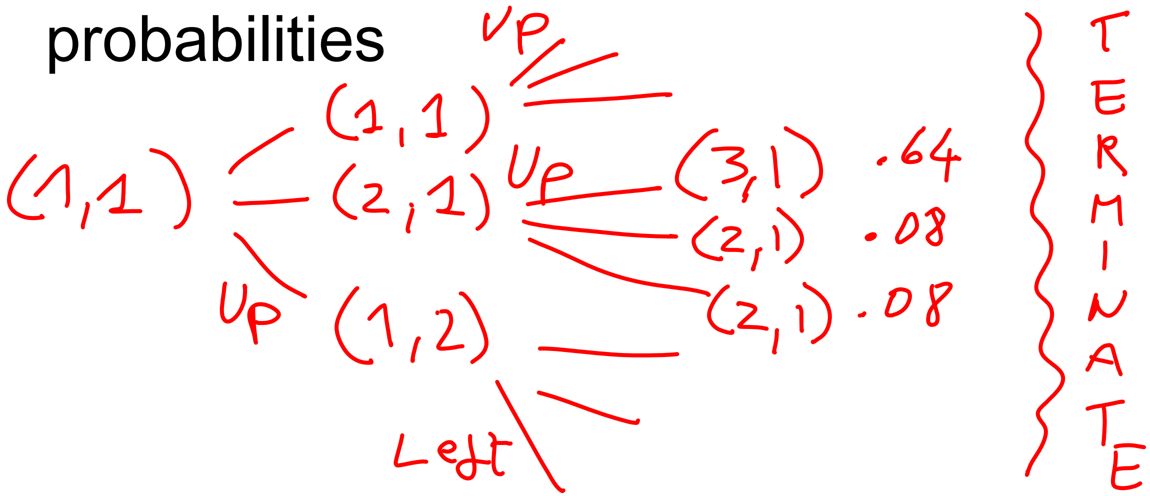
- But in an MDP the decision to be made is always…..
- Given the current state what should I do?

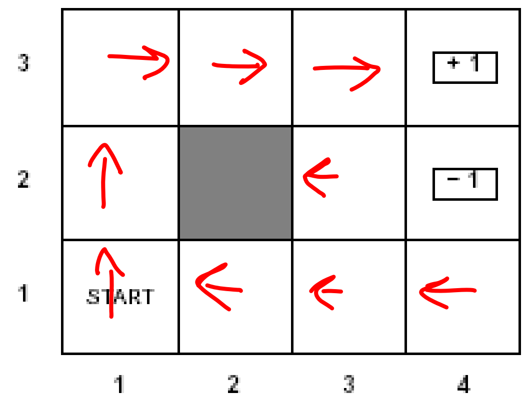
- So a policy for an MDP is a single decision function $\pi(s)$ that specifies what the agent should do for each state s



MDPs: optimal policy

Because of the stochastic nature of the environment, a policy can generate a set of environment histories with different probabilities





Optimal policy maximizes *expected total reward,* where

- Each environment history associated with that policy has a given amount of total reward
- Total reward is a function of the rewards of its individual states

Learning Goals for today's class

You can:

- Effectively represent indefinite/infinite decision processes
- Compute the probability of a sequence of actions in an Markov Decision Process (MDP)
- Compute the utility of a policy for an MDP

TODAY

TAs evaluation form

- Evaluations are not obligatory
- Please evaluate only TAs you interacted with
- TAs and Instructor won't see the evaluations until after marks are submitted
- Keep your comment specific and constructive

Next Class

- Finish MDPs – Last Class

Announcements

- Assign4 due on Wed