6/21/2021

**Lecture 2: Markov Decision Processes**

**1/ Introduction to MDPs:**

* MDP formally describe an environment for RL
* Where the environment is **fully observable**
* The current state completely characterizes the process
* Almost all RL problems can be formalize as MDP:

+ Optimal control primarily deals with continuous MDPs

+ Partially observable problems can be converted into MDPs

**+ Bandits are MDPs with one state**

**2/ Markov Property:**

* The future is independent of the past given the present.
* The state captures all relevant info from the history
* Once the state is known, the history may be thrown away meaning that the state is a sufficient statistic of the future.

**3/ State Transition Matrix:**

* For a Markov state s and successor state s’, the state transition probability is defined by:

Text

Description automatically generated

* State transition matrix P defines transition probabilities from all states s to all successor states s’

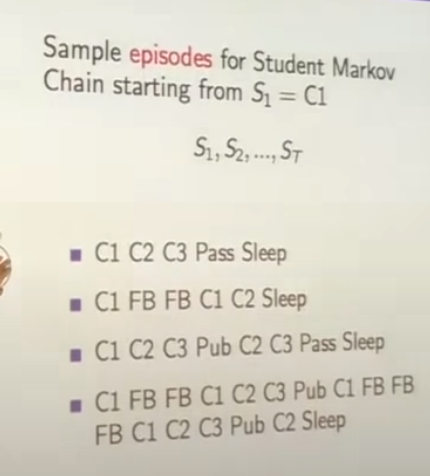
**4/ Markov Process:**

* A Markov process is a memoryless random process, is a sequence of random states S1, S2,…. With the Markov property.

Graphical user interface, text

Description automatically generated

Diagram

Description automatically generated

Diagram

Description automatically generated

* Transition matrix = what’s the probability change from 1 state to another.

**5/ Markov Reward Process:**

* A Markov reward process = a Markov chain with values

Text

Description automatically generated

* R = if we are in a state, how much reward do we get from that state.

a/ Return - reward:

* The return G\_t is the total discounted reward from time-step t. G = Goal

A picture containing scatter chart

Description automatically generated

* The discount gamma belong [0, 1] is the present value of future rewards.
* The value of receiving reward R after k+1 time-steps if gamma^k\*R
* This values immediate reward above delayed reward

+ gamma close to 0 leads to “myopic” evaluation – care about the latest reward

+ gamma close to 1 leads to “far sighted” evalution – care about previous reward

* **Why we do discount?**

+ Mathematically convenient to discount rewards

+ Avoid infinite returns in cyclic Markov Processes

+ Represent the fact that we do not have a perfect model.

+ Uncertainty about the future may not be fully represented

+ If the reward is financial, immediate rewards may earn more interest than delayed rewards (interested rate)

+ Animal/human behaviours shows preference for immediate reward

+ It is possible sometimes to use undiscounted Markov reward processes (gamma = 1) if all sequences terminate.

b/ Value Function:

* The value function v(s) gives the long-term value of state s = what’s the total reward you get

Text

Description automatically generated

–