Markov Decision Process

Lecture 2.1

Alina Vereshchaka

CSE4/510 Reinforcement Learning
Fall 2019

avereshc@buffalo.edu

August 29, 2019

*Slides has been modified from David Silver's RL course

Overview

1. Learning
2. Definition
3. Markov Decision Processes (MDP)

Table of Contents

Why do we need to learn?

Why do we need to learn?

There are (at least) two distinct reasons to learn:

1. Find previously unknown solutions. E.g., a program that can play Go better than any human, ever

Why do we need to learn?

There are (at least) two distinct reasons to learn:

1. Find previously unknown solutions. E.g., a program that can play Go better than any human, ever
2. Find solutions online, for unforeseen circumstances. E.g., a robot that can navigate terrains that differ greatly from any expected terrain

Why do we need to learn?

There are (at least) two distinct reasons to learn:

1. Find previously unknown solutions. E.g., a program that can play Go better than any human, ever
2. Find solutions online, for unforeseen circumstances. E.g., a robot that can navigate terrains that differ greatly from any expected terrain

Reinforcement learning seeks to provide algorithms for both cases
Note that the second point is not (just) about generalization — it is about learning efficiently online, during operation.

Why do we need to learn?

Science of learning to make decisions from interaction. This requires us to think about

- ...time

Why do we need to learn?

Science of learning to make decisions from interaction. This requires us to think about

- ...time
- ...(long-term) consequences of actions

Why do we need to learn?

Science of learning to make decisions from interaction. This requires us to think about

- ...time
- ...(long-term) consequences of actions
- ...actively gathering experience

Why do we need to learn?

Science of learning to make decisions from interaction. This requires us to think about

- ...time
- ...(long-term) consequences of actions
- ...actively gathering experience
- ...predicting the future

Why do we need to learn?

Science of learning to make decisions from interaction. This requires us to think about

- ...time
- ...(long-term) consequences of actions
- ...actively gathering experience
- ...predicting the future
- ...dealing with uncertainty

Examples of decision problems

Examples:

- Fly a helicopter
- Manage an investment portfolio
- Control a power station
- Make a robot walk
- Play video or board games

These are all reinforcement learning problems (no matter which solution method you use)

Table of Contents

Core concepts

Core concepts of a reinforcement learning system are:

- Environment

Core concepts

Core concepts of a reinforcement learning system are:

- Environment
- Reward signal

Core concepts

Core concepts of a reinforcement learning system are:

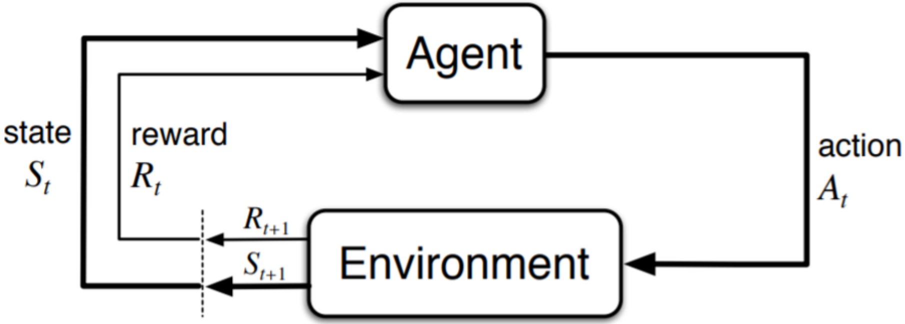
- Environment
- Reward signal
- Agent, containing:
  - Agent state
  - Policy
  - Value function (probably)
  - Model (optionally)

Definition

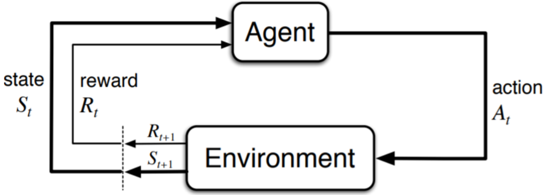
The agent is acting in an environment. How the environment reacts to certain actions is defined by a model which we may or may not know. The agent can stay in one of many states ($s \in S$) of the environment, and choose to take one of many actions ($a \in A$) to switch from one state to another. Which state the agent will arrive in is decided by the transition probabilities between states $P(s'|s, a)$. Once an action is taken, the environment delivers a reward ($r \in R$) as a feedback.

Table of Contents

Markov Decision Processes (MDP)



Finite Markov Decision Processes (MDP)



At each step t the agent:

- Receives state S_t / observation O_t and reward R_t
- Executes action A_t

The environment:

- Receives action A_t
- Emits state S_{t+1} / observation O_{t+1} and reward R_{t+1}

Finite Markov Decision Processes (MDP)

Markov property:

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, S_2, \ldots, S_t]$$

Finite Markov Decision Processes (MDP)

Markov property:

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, S_2, \ldots, S_t]$$

"The future is independent of the past given the present"

Daily life trajectory:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots, S_T$$

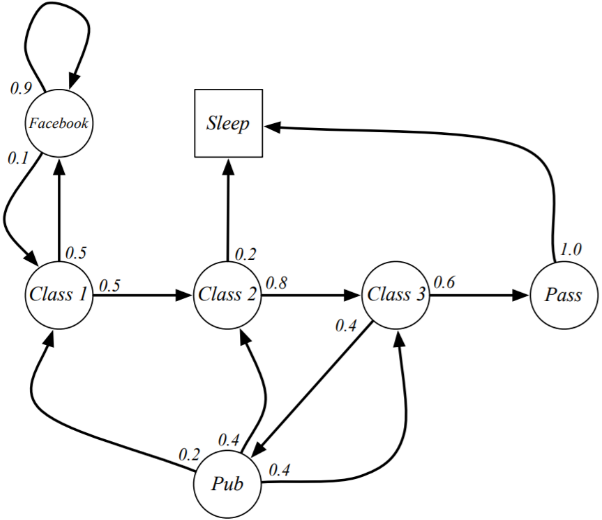
Definition

A *Markov Chain* is a tuple $\langle S, P \rangle$

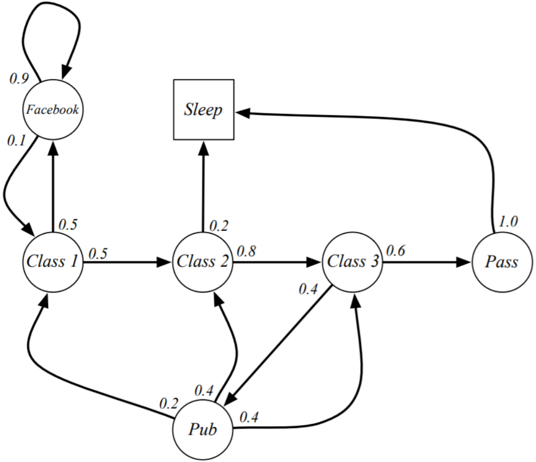
- S is a set of states
- P is a state transition probability matrix

$$P_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s] \tag{1}$$

Markov Chain - Student Example



Markov Chain - Student Example



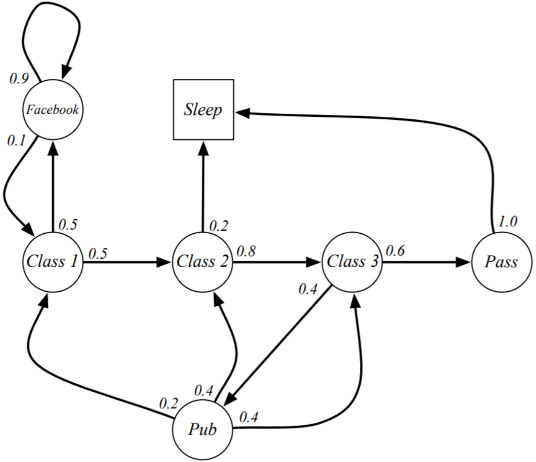
Sample episodes for Student Markov Chain starting from $S_1 = C1$.

Episode: $S_1, S_2, ..., S_T$

Episodes:

- C1 C2 C3 Pass Sleep
- C1 FB FB C1 C2 Sleep
- C1 C2 C3 Pub C2 C3 Pass Sleep
- C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

Markov Chain - Student Example



$$
\mathcal{P} = \begin{array}{c} \\ C1 \\ C2 \\ C3 \\ Pass \\ Pub \\ FB \\ Sleep \end{array}
\begin{array}{c}
\begin{array}{cccccc} C1 & C2 & C3 & Pass & Pub & FB \end{array} \\
\left[\begin{array}{cccccc}
 & 0.5 & & & & 0.5 \\
 & & 0.8 & & & \\
 & & & 0.6 & 0.4 & \\
 & & & & & \\
0.2 & 0.4 & 0.4 & & & \\
0.1 & & & & & 0.9 \\
 & & & & &
\end{array}\right.
\end{array}
$$

Markov Reward Process (MRP)

Markov reward process is a Markov chain with values.

Definition

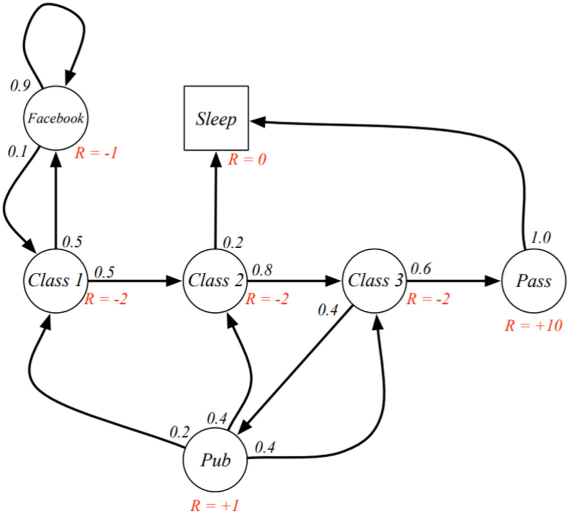
A *Markov Reward Process* is a tuple $\langle S, P, R, \gamma \rangle$

- S is a set of states
- P is a state transition probability matrix

$$P_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s] \tag{2}$$

- R is a reward function, $R_s = \mathbb{E}[R_{t+1} | S_t = s]$
- γ is a discount factor, $\gamma \in [0, 1)$

Markov Reward Process - Student Example



Discount Factor γ

The discounting factor $\gamma \in [0, 1)$ penalize the rewards in the future.
Reward at time *k* worth only γ^{k-1}

Motivation:

- The future rewards may have higher uncertainty (stock market)

Discount Factor γ

The discounting factor $\gamma \in [0, 1)$ penalize the rewards in the future.
Reward at time *k* worth only γ^{k-1}

Motivation:

- The future rewards may have higher uncertainty (stock market)
- The future rewards do not provide immediate benefits (As human beings, we might prefer to have fun today rather than 5 years later ;)

Discount Factor γ

The discounting factor $\gamma \in [0, 1)$ penalize the rewards in the future.
Reward at time k worth only γ^{k-1}

Motivation:

- The future rewards may have higher uncertainty (stock market)
- The future rewards do not provide immediate benefits (As human beings, we might prefer to have fun today rather than 5 years later ;)
- Discounting provides mathematical convenience (we don't need to track future steps infinitely to compute return)

Discount Factor γ

The discounting factor $\gamma \in [0, 1)$ penalize the rewards in the future.
Reward at time k worth only γ^{k-1}

Motivation:

- The future rewards may have higher uncertainty (stock market)
- The future rewards do not provide immediate benefits (As human beings, we might prefer to have fun today rather than 5 years later ;)
- Discounting provides mathematical convenience (we don't need to track future steps infinitely to compute return)
- It is sometimes possible to use undiscounted Markov reward processes (i.e. $\gamma = 1$) e.g. if all sequences terminate.

Return G_t

Definition

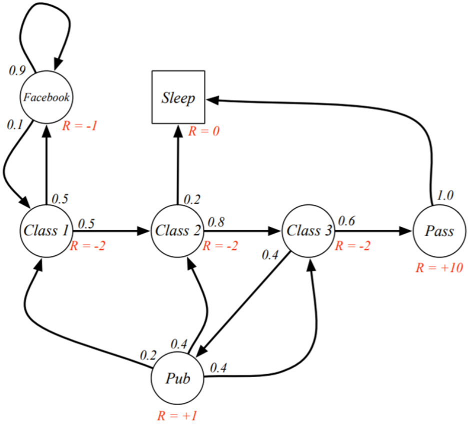
The *return* G_t is the total discounted reward from time-step t.

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{3}$$

- γ is a discount factor ($\gamma \in [0, 1)$)
- R is the reward
- The value of receiving reward R after $k + 1$ time-steps is $\gamma^k R$

Return - Student Example

Sample returns for Student MRP:
Starting from $S_1 = C1$ with $\gamma = 0.5$
$G_1 = R_2 + \gamma R_3 + \cdots + \gamma^{T-2} R_T$



Return - Student Example

Sample returns for Student MRP:
Starting from $S_1 = C1$ with $\gamma = 0.5$

$$G_1 = R_2 + \gamma R_3 + \dots + \gamma^{T-2} R_T$$

C1 C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 10 * \frac{1}{8}$	$=$	-2.25
C1 FB FB C1 C2 Sleep	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16}$	$=$	-3.125
C1 C2 C3 Pub C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 1 * \frac{1}{8} - 2 * \frac{1}{16} \dots$	$=$	-3.41
C1 FB FB C1 C2 C3 Pub C1 ...	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16} \dots$	$=$	-3.20
FB FB FB C1 C2 C3 Pub C2 Sleep			

Markov Decision Process (MDP)

Definition

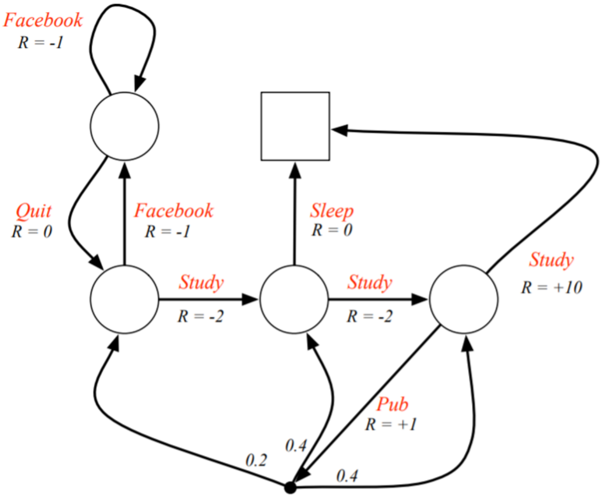
A *Markov Decision Process (MDP)* is a tuple $\langle S, A, P, R, \gamma \rangle$

- S is a set of states
- A is a set of actions
- P is a state transition probability matrix

$$P_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a] \tag{4}$$

- R is a reward function, $R_s = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$
- γ is a discount factor, $\gamma \in [0, 1)$

Markov Decision Process - Student Example



Summary So Far

- Mains reasons to learn (not just for agents) are to find previously unknown solutions and to the find solutions in unforeseen circumstances
- Core parts of a reinforcement learning are: Environment, Reward, Agent
- Markov property: The future is independent of the past given the present
- The discounting factor $\gamma \in [0, 1)$ penalize the rewards in the future
- Markov Decision Process (MDP) defined as a tuple $\langle S, A, P, R, \gamma \rangle$