David Silver's RL Intro （分解版）

by 深度碎片

早期版本都会是英文，全部完结后，会在英文版基础上做中文版；
现在就想一起学的小伙伴可以带着文字版的问题和回答去看分解视频；
PDF文字版本，可在这里下载；

David Silver Reinforcement learning (study use) - YouTube
David Silver Reinforcement learning (study use) - Bilibili

Why study this course?

- This is a short introduction course to RL
- David Silver is a great lecturer to beginners

How do I study this course?

- break videos into small questions
- each video is to address an important question to me
- provide one line answer to each question

Lecture 1

P1 1.1 admin textbook lecture notes

Lecture 1 tries to answer 3 questions:

1. What does a RL problem look like?
2. How to solve a RL problem?
3. What are the challenges to RL?

Sutton's book is comprehensive, but second book is shorter and mathematical

P2 1.2 RL lays at the core of many sciences

- RL are crucial and fundamental to many key sciences

P3 1.3 what special about RL

1. No supervisor, just trial and error with reward signal
2. Feedback is delayed, but reward is instantaneous, I guess
3. time data is sequential and dynamic vs (wholesale, fixed) iid data
4. sequential data is affected by agent action at each time step

P4 1.4 RL examples and demos

apply above unique differences to the following examples

- fly stunt maneuver in helicopter
- fight backgammon
- investment portfolio
- control power station

P5 1.5 define Reward in RL problem

How to understand reward at each step and goal in the end?

- all goals can be defined with maximization of all rewards at each time
- "no intermediate reward" can be resolved with this hypothesis above
- goal in terms of speed can also be resolved as well

P6 1.6 reward examples

How to understand the reward and goal in real world examples?

- +ve = positive, -ve = negative
- helicopter: +ve reward each step, large -ve reward at the end if crashed
- Backgammon: no intermediate reward, just end game reward
- power station: +ve reward as power increase, -ve reward if exceeding saftey threshold

P7 1.7 maximizing reward can be tricky

Why maximize reward is not as straightforward as one might think?

- maximizing by adding rewards incrementally at each time step
- maximizing by taking strategic moves sacrifice reward now for bigger reward later

P8 1.8 define agent and environment why reward scalar

What are the training data for RL in terms of agent and environment?

- Agent is the core of RL model, which receive observation and reward and take an action at each time step
- Environment, RL model has no control, which is influenced by action and emit reward and observation for agent

Why reward can be a scalar when facing conflicting needs?

- no matter what, as long as an action to be taken, scalar can do the job

P9 1.9 define history and state in RL problem

What is history?

- all (a stream of) observations, rewards and actions up to this moment
- agent and environment act depend on history

What is state?

- state is a function (representation) of history
- it is more efficient to use state than history for running models
- there are 3 popular representations of history (state)

P10 1.10 what is environment state

What should we know about environment state?

- env state is env's private representation of the world it perceives
- such representation (number or data) will determine its emission of observations and rewards
- but env state is usually invisible to agent, and probably much irrelevant if visible
- not really useful to help us build agent model

P11 1.11 what is agent state

What to understand about agent state?

- agent state is what agent take in from history for algorithms to produce an action
- agent state, again, is a function or representation of history, a particular kind history filtered by agent

P12 1.12 what is information or Markov state

What to know about information or Markov state?

- Markov state is a state or representation of history which captures the key aspect of history which help determine the future
- Markov state has the equivalent effect of the whole history in predicting the future

P13 1.13 example of agent state for prediction

How does the rate predict the future with state?

- it depends on how rate define the state it receives
- if state = last three items, the rat predict electricity hit
- if state = count of items, the rat predict cheese
- if state = entire history, the rat won't know

State, is what data agent select for algorithm processing

P14 1.14 fully observable environment or Markov decision process

What is fully observable env or Markov decision process?

- env state = all useful info produced by env to determine the observations and rewards
- agent can see all env state, therefore, agent state = env state = Markov state
- another name for this above = Markov decision process

P15 1.15 partial observability env

What to know about partial obs env or partial observable Markov Decision Process POMDP?

- env state is partially observable, so how to build agent state?
- version 1: take all history available to be agent state
- version 2: take a vector of different probability of env state to be agent state
- version 3: use recurrent neural net to construct agent state with previous agent state and current observations

P16 1.16 major components of RL agent

What are the three key components of RL agent?

- policy: agent use it to transform state to action
- value: agent evaluate how good is the state or how much reward expected to get
- model: agent's representation of the environment
- they are not always required for defining an agent

P17 1.17 what is policy

What to know about policy?

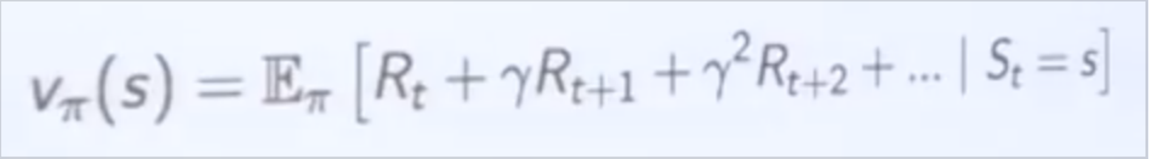
- policy : a map from state to action
- deterministic policy: a = F (s) , use a clearly define function to generate action from state
- stochastic policy: a = F (a | s) = P(A = a | S = s), action from the state does not have to be fixed, randomness can be introduced too

P18 1.18 value function and example

What to know about value function?

- it is to predict the total future reward at current state
- i.e., to evaluate how good is the current state
- according to different state and reward prediction, agent choose actions accordingly

What is the formula of value function?

- 

$$v_\pi(s) = \mathbb{E}_\pi\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots \mid S_t = s\right]$$

- 100 steps to anticipate into the future, e.g., t, t+1, … t + 99

What to note in Atari game?

- note how value changes and how different actions chosen

P19 1.19 what are model transition reward *

What does model, transitions do?

- a model predicts what environment would do next (observations, reward)
- transitions predict the next state given previous state and action
- then to predict next step reward as expected reward based on previous state and action

P20 1.20 maze example on policy value function model

What does policy look like in maze example?

What does value function look like in maze example?

What does model look like in maze example?

P21 1.21 value based agent categorization

What does it mean by categorizing agent with value?

- agent's action can be fully determined by value calculated
- policy, as actions, kind of implicitly contained by value function
- see this in Maze example

P22 1.22 policy based actor critic agent categorization

What is Agent based on policy?

- agent defined by its actions (maze example)

what is agent of actor critic

- agents use both policy and value

P23 1.23 model free and model based agents

What is model free agent?

- agent only uses policy and value
- agent does not attempt to figure out how env works

What is model based agent?

- agent use both policy and value
- agent also tries to figure out how env works

P24 1.24 differentiat learning and planning

How RL sees the world and solves world problems?

How planning sees the world solves world problems?

They are also linked

P25 1.25 atari example on planning and learning

How to see and solve atari problem with RL approach?

How to see and solve atari problem with planning?

P26 1.26 balance exploitation and exploration

RL is a sort of trial and error learning

Exploration is to sacrifice some reward for more information of env

Exploitation is exploit known info to maximize reward

many example of balancing exploitation and exploration

P27 1.27 prediction and control with course outline *

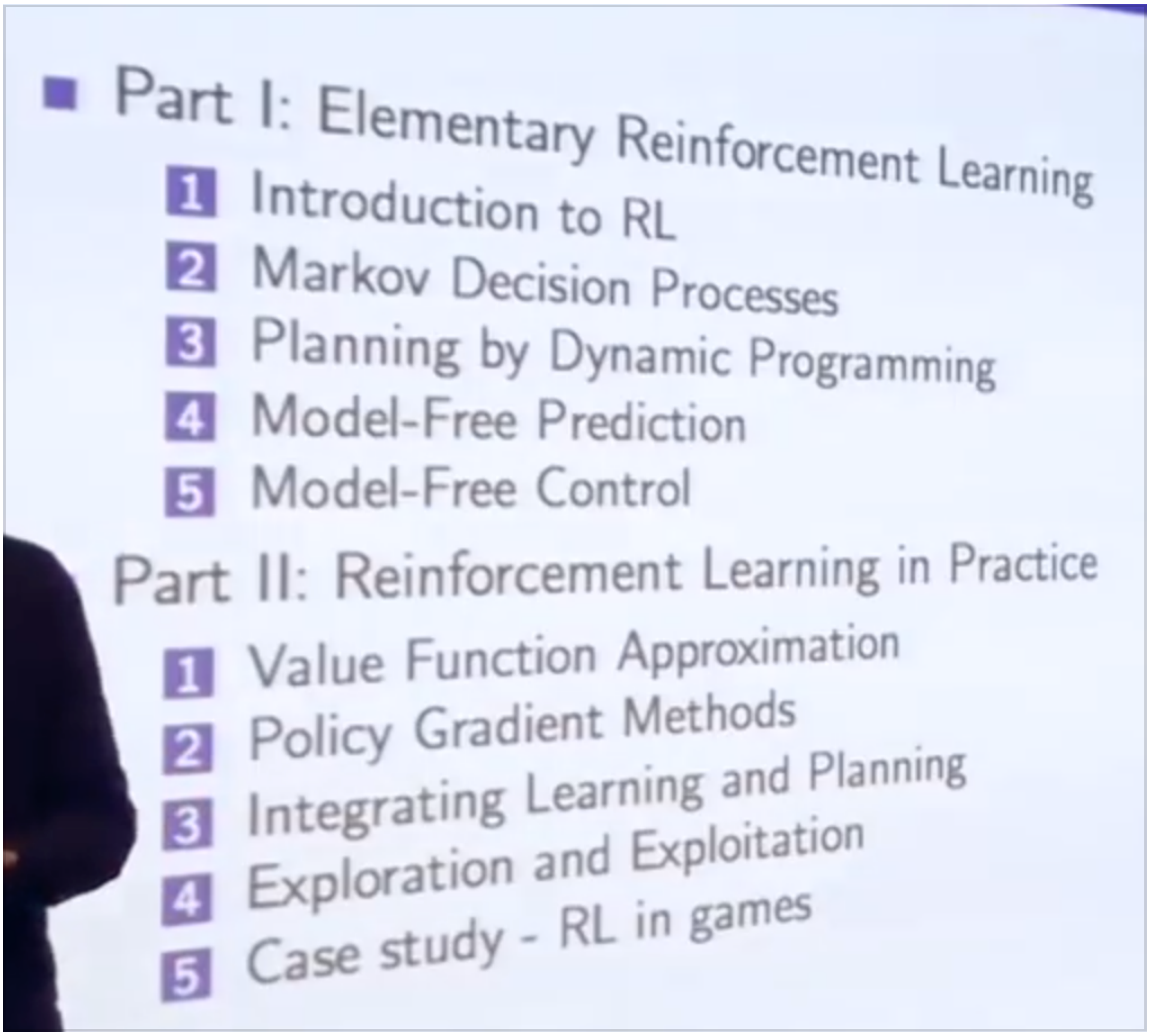
What is prediction?

- the expected reward for the given policy

What is control?

- to find the best policy by finding out all policies and their reward

Course Outlines



- Part I: Elementary Reinforcement Learning
  1. Introduction to RL
  2. Markov Decision Processes
  3. Planning by Dynamic Programming
  4. Model-Free Prediction
  5. Model-Free Control

Part II: Reinforcement Learning in Practice
  1. Value Function Approximation
  2. Policy Gradient Methods
  3. Integrating Learning and Planning
  4. Exploration and Exploitation
  5. Case study - RL in games