《深度强化学习》课程之第一讲（2021年春季研究生）

深度强化学习概述

苏州大学计算机科学与技术学院

主讲：刘 全

目 录

1. 课程要求（1）

➢对数学理论基础的要求。

强化学习以积分学、线性代数、统计学、最优化理论等数学理论为基础，并不断吸纳生物、机械、物理等方面的学科知识。

➢对计算机专业的要求。

- ◆ 强化学习课程的核心是培养学生分析问题和解决问题的能力；
- ◆ 即要兼顾理论知识（包括理论推导证明），同时培养学生实际编程能力（包括大数据的处理和模型算法的编写）。

➢具有将数学和计算机专业的内容紧密地结合在一起的能力 。

1. 课程要求（2）

- 教学要求：
  - 理论教学与实践教学紧密相连；
  - 组织教学内容，合理分配实验环节，激发学生的学习兴趣；
  - 加强学生实践动手能力的培养，达到知识传授和能力培养的有效结合。

1. 课程要求（3）

- 教学方法要求：
  - 强化学习很多算法理论性强、抽象、不易理解，单纯采用文字叙述和公式推导的教学手段，教学效果并不好。
  - 在理论教学中，可以结合实例讲解，注重理论联系实际；
  - 在强化学习教学中，以“扫地机器人”应用贯穿整个教学过程；
  - 通过实例，知道算法的应用场景和方法，学习兴趣和效率自然提高。

1. 课程要求（4）

- 实验要求：
  - ◆根据理论教学内容，结合学生的实际情况，按照由浅入深的原则安排实验；
  - ◆验证性实验，要求学生通过实现相关算法，验证教材实例的正确性。这对理解算法、掌握算法的技巧非常有益；
  - ◆综合性实验，运用图像处理、可视化编程、深度学习、强化学习等知识，解决实际问题；
  - ◆实验的难度由易到难，层层深入，有利于学生动手能力的培养。

1. 课程要求（5）

➢教学内容

◆研究生（54+36）：

（1）环境搭建及编程（4课时）；

（2）基于表格的DP、MC、TD方法 （10课时）；

（3）模型学习（10课时）；

（4）深度学习及PyTorch（10课时）；

（5）策略梯度（6课时）；

（6）深度强化学习算法（DQN、DDPG、A3C等）（14课时）。

1. 课程要求（6）

- 理论：实验=6：4
- 理论包括：

  平时作业+考试（小论文）

- 实验包括：本学期4个编程题目，每个题目10分。
  - ✓ Gym平台；
  - ✓ Python+Pytorch实现；
  - ✓ Jupyter Notebook编写实验报告。

目 录

2. 引言（1）

- 目前机器学习领域中较热门的两个分支

  深度学习（Deep Learning，DL）

  强化学习（Reinforcement Learning，RL）

- 深度学习的基本思想：通过堆叠多层的网络结构和非线性变换，组合低层特征以实现对输入数据的分级表达。

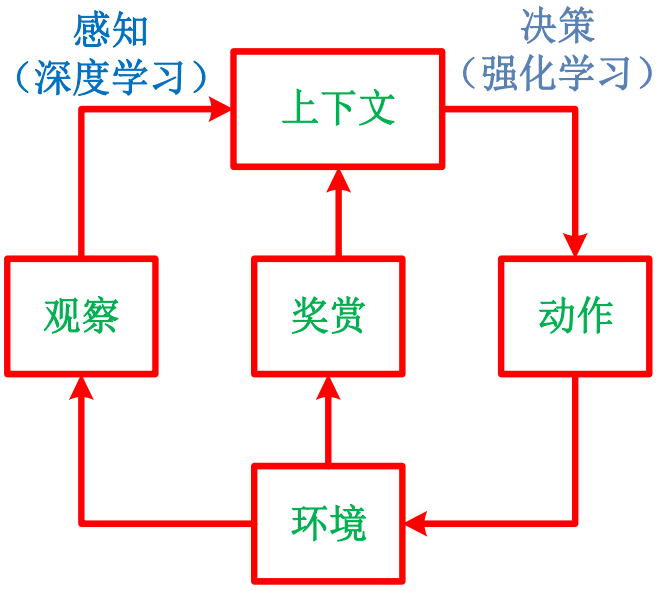
- 强化学习并没有提供直接的监督信号来指导智能体（agent）的行为。

2. 引言（2）

- 在强化学习中，**agent**是通过试错的机制与环境进行不断的交互，以最大化从环境中获得的累计奖赏。
- 深度强化学习（**Deep Reinforcement Learning，DRL**）将具有感知能力的深度学习和具有决策能力的强化学习相结合，初步形成从输入原始数据到输出动作控制的完整智能系统。

2. 引言（3）

- 深度强化学习是一种端对端（**end-to-end**）的感知与控制系统，具有很强的通用性。



2. 引言（4）

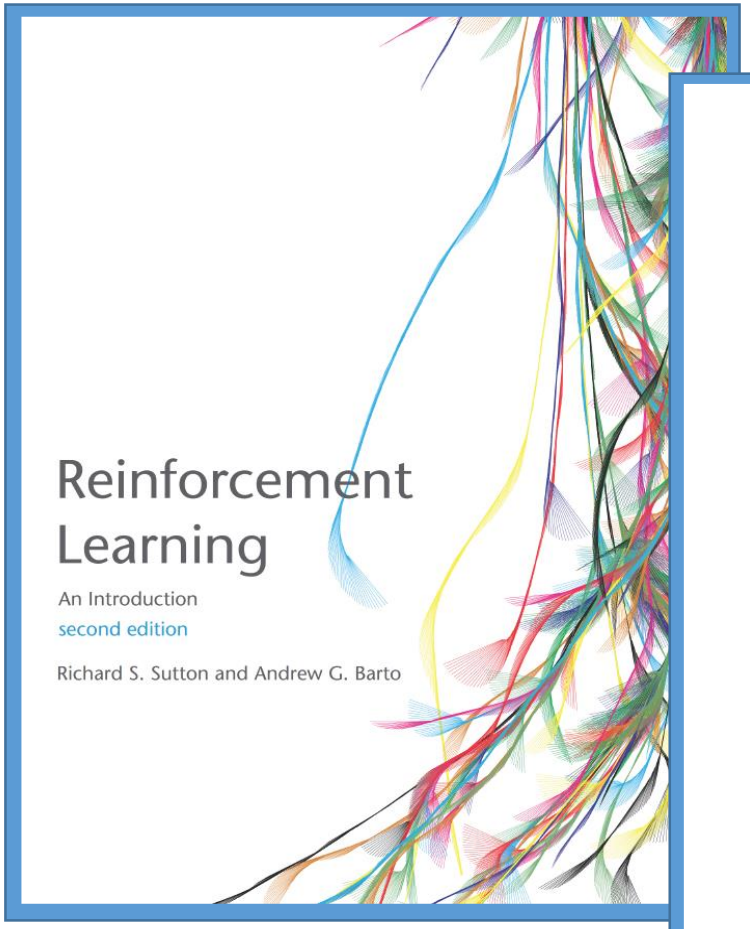
➢其学习过程可以描述为：

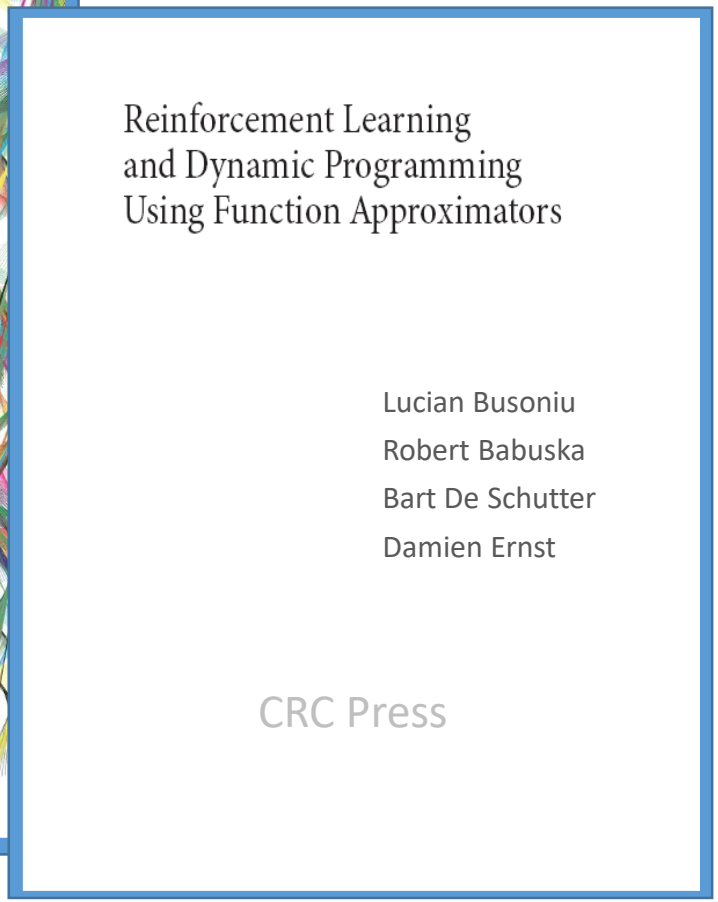
（1）在每个时刻，agent与环境交互得到一个高维度的观察，并利用深度学习方法来感知观察，以得到抽象、具体的状态特征表示；

（2）基于预期回报来评价各动作的价值函数，并通过某种策略将当前状态映射为相应的动作；

（3）环境对此动作做出反应，并得到下一个观察。通过不断循环以上过程，最终可以得到实现目标的最优策略。

2. 引言（5）


Reinforcement
Learning
An Introduction
second edition
Richard S. Sutton and Andrew G. Barto


Reinforcement Learning
and Dynamic Programming
Using Function Approximators
Lucian Busoniu
Robert Babuska
Bart De Schutter
Damien Ernst
CRC Press

2. 引言（6）

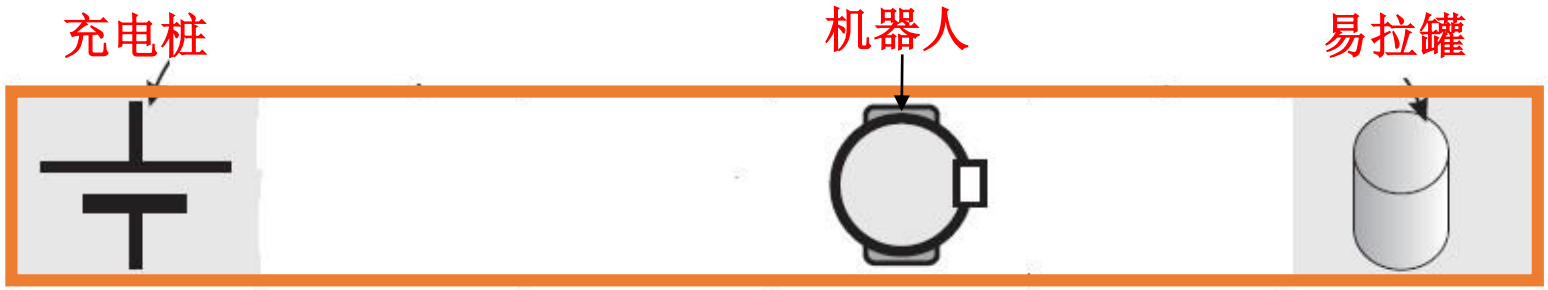

CRC Press
Reinforcement Learning and
Dynamic Programming
Using Function Approximators
基于函数逼近的
强化学习与动态规划
中国工信出版集团
人民邮电出版社


大规模强化学习
刘全 傅启明 钟珊 黄蔚 著

目 录

3. 强化学习实例（1）

任务（1）：清洁机器人问题



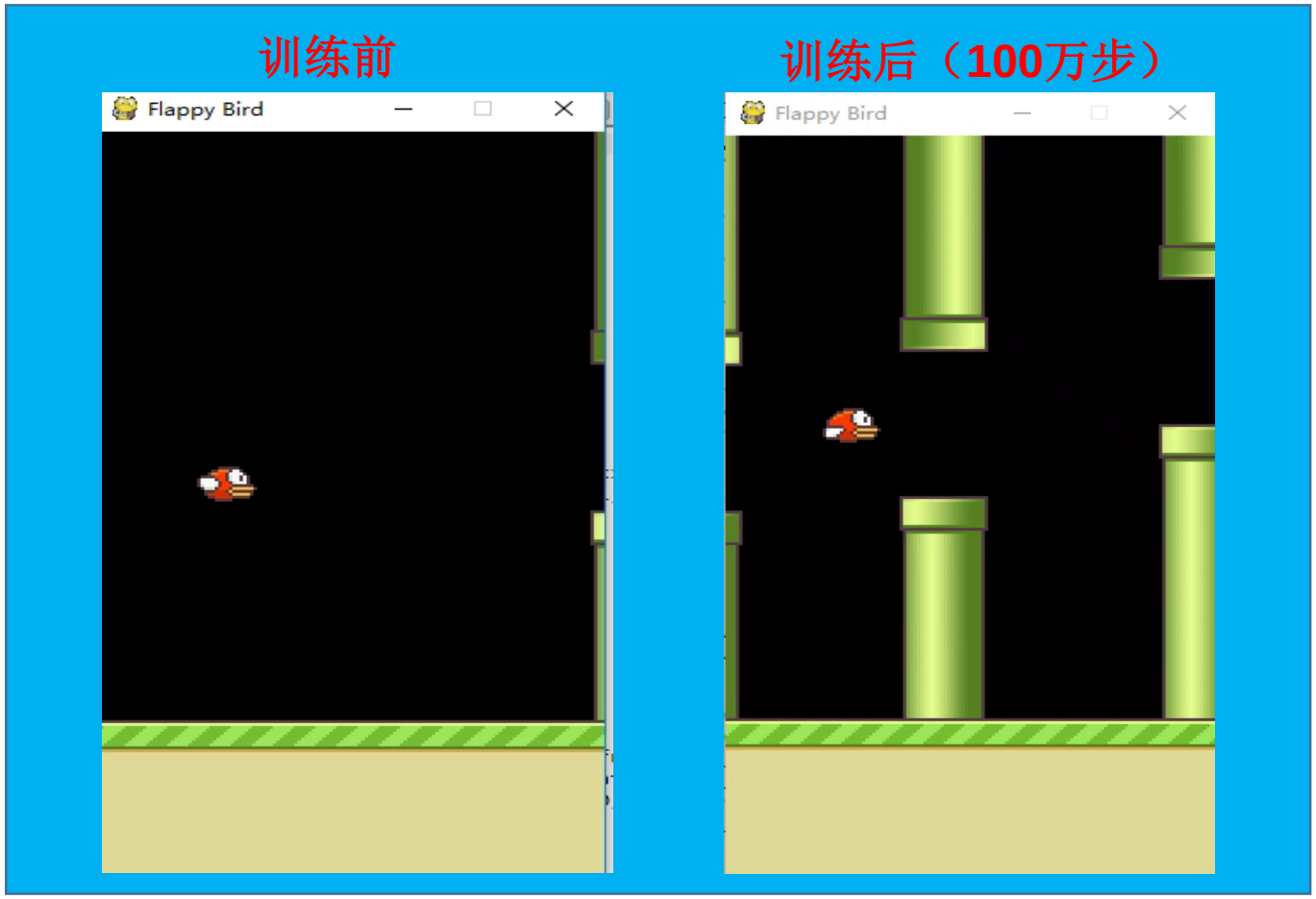
任务:

清洁机器人：收集易拉罐，充电。

机器人采取怎样的行动，才能达到预期的目标。

3. 强化学习实例（2）

任务（2）：笨鸟先飞



3. 强化学习实例（3）

任务（3）： AlphaGo ＆ AlphaGo Zero

AlphaGo’ s game with Lee Sedol and Ke Jie

AlphaGo：DL、RL、MC Tree

AlphaGo Zero：DRL

目 录

4. 强化学习概念（1）

- 所谓强化学习，是指从环境状态到行为映射的学习，以使系统行为从环境中获得的累积奖赏（reward）最大。
- 在强化学习中，算法来把外界环境转化为最大化奖励量的方式做动作，算法并没有告诉Agent要做什么或者采取哪个动作。
- Agent的动作的影响不只是立即得到的奖励，而且还影响接下来的动作和最终的累积奖赏。

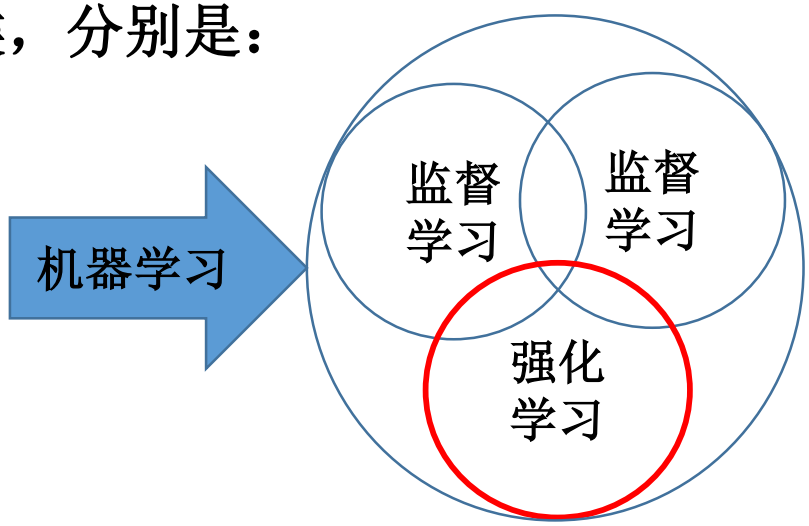
4. 强化学习概念（2）

- 1954年，Minsky：提出强化学习的概念和术语。
- 1956年，Bellman：MDP的动态规划方法。
- 1977年，Werbos：自适应动态规划算法。
- 1988年，Sutton：时序差分算法。
- 1992年，Watkins：Q-Learning算法。
- 1994年，Rummery：Sarsa算法。
- 2006年，Kocsis：置信上界树算法。
- 2009年，Kewis：反馈控制自适应动态规划算法。
- 2014年，Silver：确定性策略梯度算法。
- 2015年，Google deep mind：DQN算法。

4. 强化学习概念（3）

强化学习与机器学习

- 强化学习是智能体（**Agent**）以“试错”的方式进行学习，通过与环境进行交互获得奖励指导行为，目标是使智能体获得最大的累积奖赏（回报）。
- 机器学习可以分为三类，分别是：
  - ✓ 监督学习
  - ✓ 无监督学习
  - ✓ 强化学习



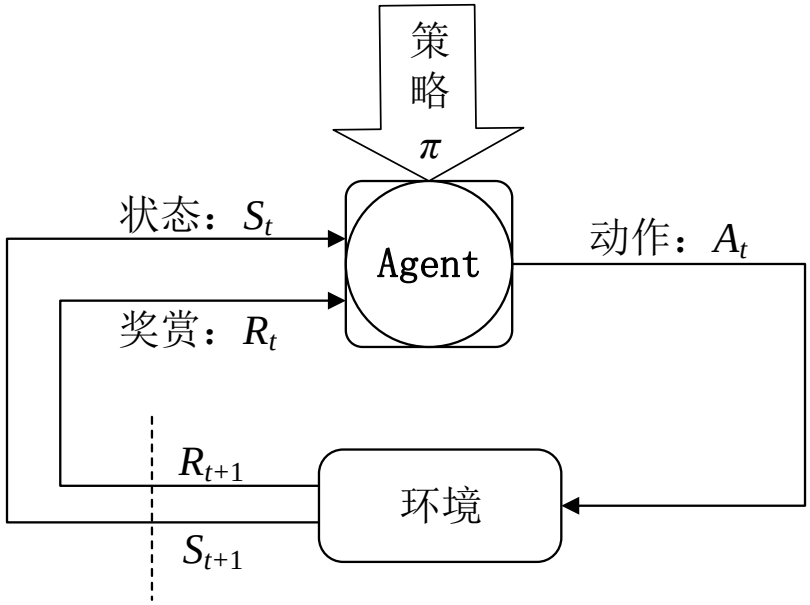
4. 强化学习概念（4）

强化学习与其他机器学习不同：

- 没有教师信号，也没有label，只有reward；
- 反馈有延时，不是立即返回；
- 数据是序列化的，数据与数据之间是有关联的，而不是i.i.d的；
- Agent执行的动作会影响之后的数据。

4. 强化学习概念（5）

强化学习的模型图：



4. 强化学习概念（6）

强化学习的关键要素：

- 强化学习的关键要素有：环境、奖赏、动作和状态。

  有了这些要素，就可以建立一个强化学习模型；

- 强化学习解决的问题是：针对一个具体问题，得到一个最优策略，使得在该策略下获得的长期回报最大；

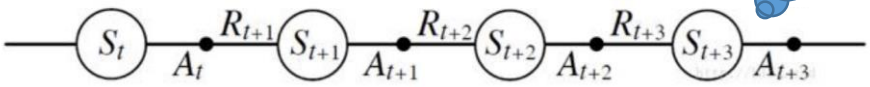
- 策略：在系列状态下，采取的动作或动作概率。

4. 强化学习概念（7）

Agent与环境的交互：

注意：在状态S_i时，执行A_i动作，然后获得R_{i+1}立即奖赏，到达S_{i+1}状态。

- 交互过程更准确地表述：



- 每一步：Agent根据策略选择一个动作执行，然后感知下一步状态和立即奖赏，通过经验再修改自己的策略；
- Agent的目标：找到最优策略，最大化长期回报。

4. 强化学习概念（8）

状态与策略：

- **状态（state）：**

  就是指当前agent所处的状态。

- **策略（policy）：就是指agent在特定状态下的动作依据，是从state到action的映射。**

  - 确定策略：某一状态下的确定动作 $a = \pi(s)$；
  - 随机策略：以概率来描述，即某一状态下执行这一动作的概率 $\pi(a \mid s)=\mathrm{P}[A_t = a \mid S_t = s]$ 。

4. 强化学习概念（9）

动作与奖赏：

- **动作（action）：**
  - 来自于动作空间，每个状态通过采取动作进行状态转移；
  - 执行动作的目的是达到最大化期望奖赏，直到最终算法收敛，所得到的策略就是一系列action的序列数据。
- **奖赏（reward）：**
  - 奖赏通常被记作R_t，表示第t个时间步的返回奖励值。所有强化学习都是基于奖赏假设的。
  - 奖赏通常为一个标量。
  - 注意：回报（return）是奖赏（reward）的累积。

4. 强化学习概念（10）

策略的种类：

- 行为策略（$b(s)$）：
  - 用来指导个体产生与环境进行实际交互行为的策略；
  - 实际采样的策略。
- 目标策略（$\pi(s)$）：
  - 用来评价状态或行为价值的策略（或待优化的策略）。

4. 强化学习概念（11）

预测与控制：

- **预测：**

  给定某个策略，估计该策略下，每个状态或状态动作对的价值。

- **控制：**

  找到一个最优的策略。

- 在RL算法中，通常都是迭代地进行先预测，再控制的过程，直到收敛。

目 录

5. 强化学习方法分类（1）

- 环境模型：理解环境或感知环境
- 更新方式：回合更新或单步更新
- 求解方式：基于价值或基于策略
- 策略使用：同策略或异策略

5. 强化学习方法分类（2）

- 环境模型：理解环境或感知环境

✓ **Model-based:** 先理解真实世界是怎样的，并通过实验，建立一个模型来模拟现实世界的反应，通过想象来预判断下来将要发生的所有情况，并且通过计算来选择下一步采取的策略。整个过程只需要计算即可，而不需要实际去“经历”。

◆ 例如：DP

✓ **Model-free:** 不依赖环境，不尝试去理解环境，Agent会根据现实环境的反馈采取下一步的动作，一步一步等待真实世界的反馈，再根据反馈采取下一步的动作。需要实际去“经历”。

◆ 例如：Q-learning, Sarsa, 策略梯度

5. 强化学习方法分类（3）

- 更新方式：回合更新或单步更新

✓ **MC-更新:** 在情节式任务中，一个情节完成后才进行更新。即 **episode by episode**。

  - 例如：**REINFORCE, MC**

✓ **TD-更新:** 在情节式任务或连续任务中，不需要等到情节结束，而是每一步都在更新。即**step by step**。

  - 例如：**Q-learning, Sarsa, 策略梯度**

5. 强化学习方法分类（4）

- 求解方式：基于价值或基于策略

✓ **Value-based:** 目标是找到状态或状态动作对的价值，通过价值来选择动作，这类方法对连续动作不适用。

  - 例如： **Q-learning, Sarsa**

✓ **Policy-based:** 目标是找到最优策略，通过感知分析所处的环境，直接输出下一步要采取的各种动作的概率，然后根据概率采取动作。这类方法对连续动作适用。

  - 例如：策略梯度，AC

5. 强化学习方法分类（5）

- 策略使用：同策略或异策略

✓ **on-policy:** 目标策略和行为策略相同。

- 例如： Sarsa, Sarsa(), TRPO

✓ **off-policy:** 目标策略和行为策略不同。

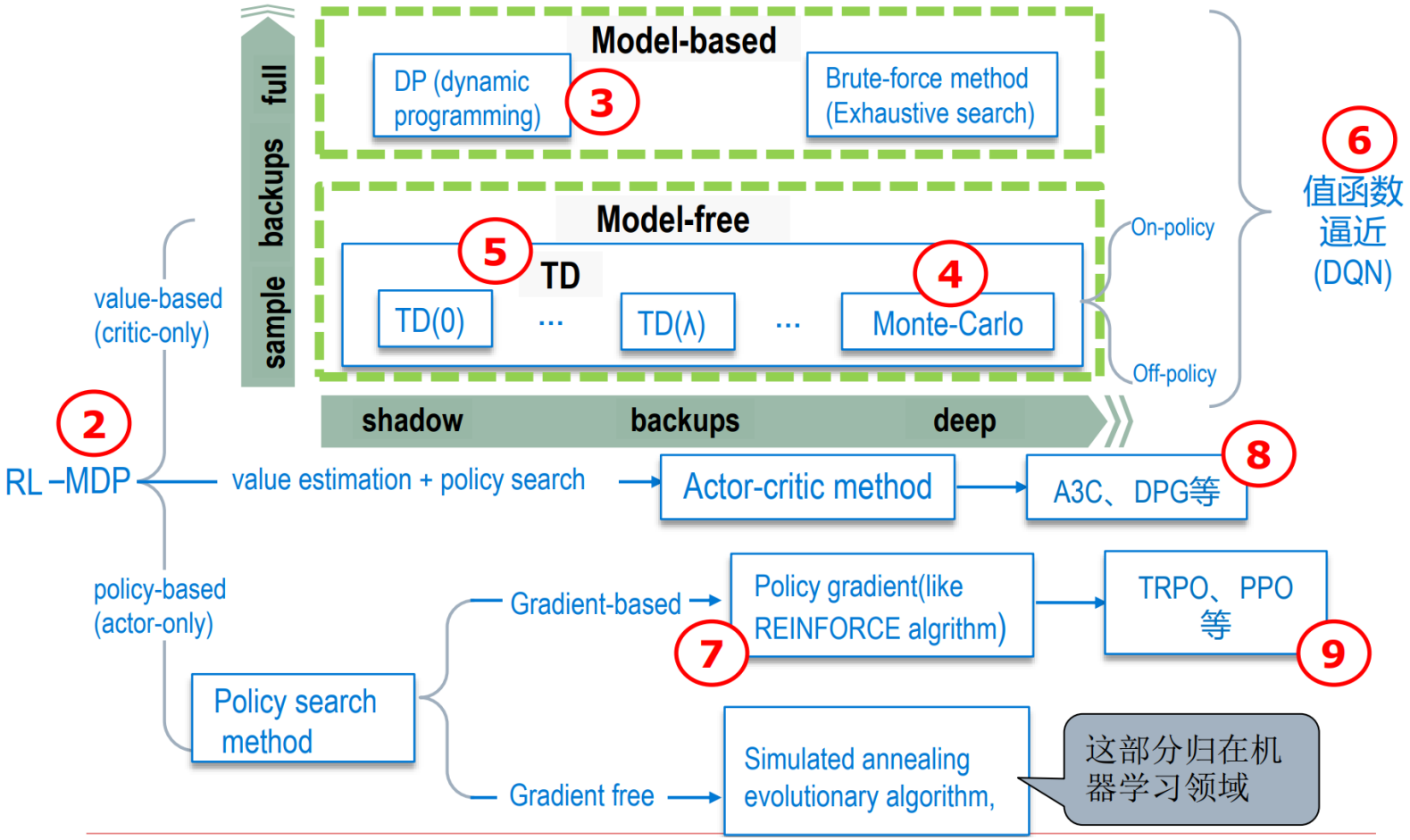
- 例如：Q-learning, DQN, 确定策略梯度

两者的区别：更新Q值时是使用既定策略还是新的策略。

5. 强化学习方法分类（6）

- 异策略的特点
  - ✓ 可以从人类给出的示教样本或其他智能体给出的引导样本中学习；
  - ✓ 可以重用由旧策略生成的经验；
  - ✓ 可以在使用一个探索性策略的同时，学习一个确定性策略；
  - ✓ 可以用一个策略进行采样，然后同时学习多个策略。

5. 强化学习学习线路图（7）



目 录

6. 常用的实验环境

- 与其他机器学习方向一样，强化学习也有一些经典的实验场景，如**Mountain-Car**，**Cart-Pole**等；
- 由于近年来深度强化学习（**DRL**）的兴起，各种新的更复杂的实验场景也在不断涌现，出现一系列优秀的平台。
- 常见的强化学习实验平台：
  - ✓ **OpenAI Gym**，**OpenAI Baselines**
  - ✓ **MuJoCo**，**rllab**，**TORCS**，**PySC2**

目 录

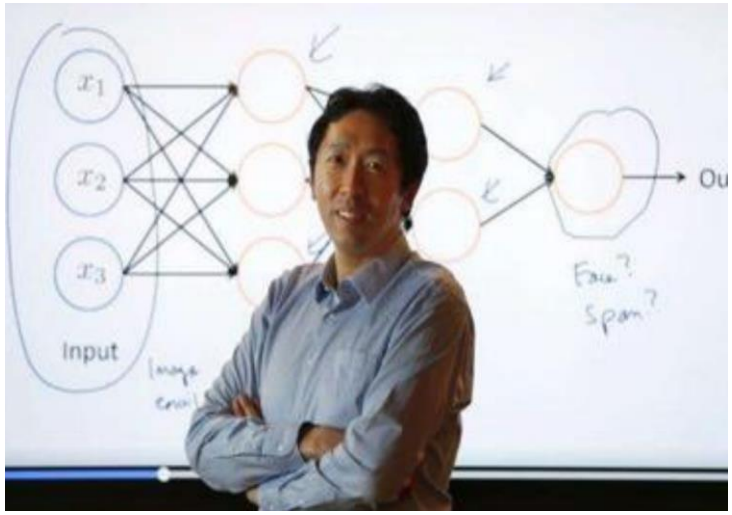
7. 著名学者（1）

- **Richard S. Sutton**
  - ◆ 现代强化学习理论的创始人之一。
  - ◆ 贡献：
    - 时序差分学习
    - 策略梯度方法
    - Dyna架构
  - ◆ 《Reinforcement Learning: An Introduction》

个人主页： http://richsutton.com

7. 著名学者（2）

- 吴恩达



◆ 是国际上人工智能和机器学习领域最权威的学者之一。

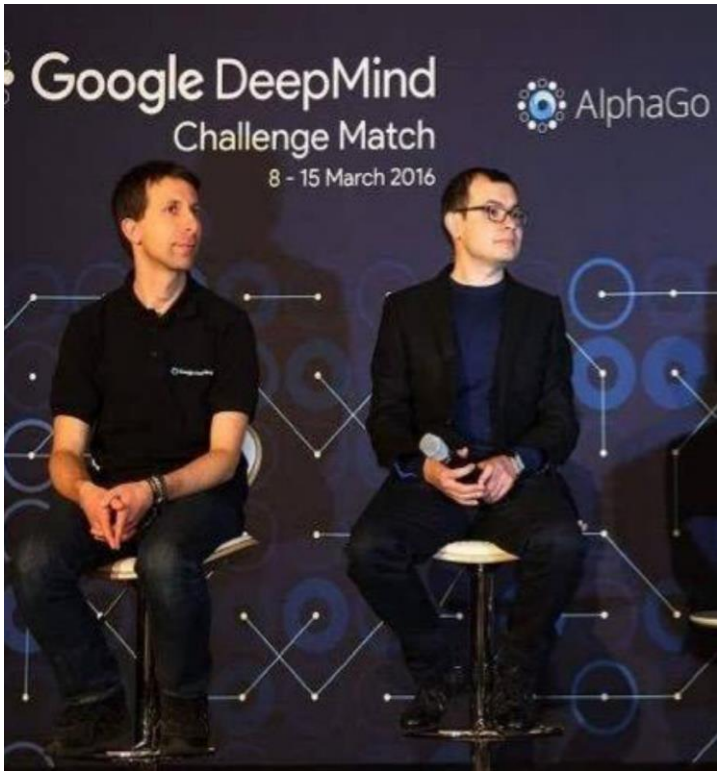
◆ 在线教育平台Coursera的联合创始人。

◆ 在DL和RL两个领域都有突出贡献。

◆ 2014年，加入百度，负责Baidu Brain计划。

7. 著名学者（3）

- **David Silver**
  - **DeepMind AlphaGo项目首席研究员；**
  - **DPG的提出者；**
  - **强化学习公开课。**
- **Demis Hassabis**
  - **DeepMind联合创始人兼CEO；**
  - **DQN的提出者。**



推荐参考资料

- 陈仲铭，何明.深度强化学习原理与实践. 北京：人民邮电出版社，2019.
- 廖星宇. 深度学习入门之PyTorch. 北京：电子工业出版社，2018.
- 冯超. 强化学习精要：核心算法与TensorFlow实现. 北京：电子工业出版社，2018.
- 郭宪，方勇纯. 深入浅出强化学习：原理入门. 北京：电子工业出版社，2018.
- 邹伟，鬲玲，刘昱杓. 强化学习. 北京：清华大学出版社，2020.
- 刘全，傅启明，钟珊，黄蔚. 大规模强化学习. 北京：科学出版社，2016.
- 刘全，傅启明，章宗长译. 基于函数逼近的强化学习与动态规划. 北京：人民邮电出版社，2018.
- 俞凯等译. 强化学习. 北京：电子工业出版社，2019.
- Sutton R S, Barto A G. Reinforcement learning: An introduction. Cambridge: MIT press, 2018.

习题

1. 机器学习主要分为哪几个类别？根据强化学习的基本原理，简述强化学习与其他机器学习方法的异同点。

2. 深度强化学习主要有哪些类别？

3. 阐述深度学习、强化学习及深度强化学习三者之间的关系。

4. 举例说明深度强化学习的未来发展方向。

The End