

2026 年秋季学期《强化学习》课程大纲

任课教师: 王宏宁 (hw-ai@tsinghua.edu.cn)
助教: 李谷蝉 (li-gc22@mails.tsinghua.edu.cn)
清华大学计算机科学与技术系

1 课程概述

知其然，知其所以然。

– 《朱子语类》卷九《论知行》

强化学习是人工智能和机器学习领域的一个重要研究分支，其主要研究目的是解决如何使得智能体在与环境的交互中学会做出最优的决策与规划。强化学习方法已广泛地应用于我们的日常生活中，从机器人技术、自动驾驶和游戏，到智能信息系统，例如搜索引擎和推荐系统，都可以找到它成功应用的案例。同时，这一方法在计算机科学之外的不同学科领域中也获得了广泛的关注与深入研究，包括心理学、神经科学、优化和运筹学等领域均对强化学习方法从不同角度进行相关研究。多年前，随着深度学习和表征学习方法的兴起，强化学习方法再次唤起人们的广泛关注。尤其是 DeepMind 的 AlphaGo 取得的巨大成功使得强化学习再次成为人工智能领域的焦点，并继续多年在工业应用与学术研究中获得日益增长的关注。

本课程内容面向清华大学本科生设置。我们将讨论强化学习的基础知识，从多臂老虎机问题、马尔可夫决策过程入手，深入介绍基于模型 (model-based) 和非基于模型 (model-free) 的强化学习方法，探讨不同策略评价 (policy evaluation) 和控制方法 (control)，并深入学习最新的深度强化学习方法。通过本课程的学习，同学们将能够：

- 了解并掌握强化学习领域内最基本与重要的概念与方法。学习如何使用强化学习的思想思考实际问题，尤其是交互式序列决策、不确定环境下的规划与学习等重要问题。
- 强化学习对于数学工具（如微积分、线性代数等）和编程实践有较强的要求，因此可以很好的培养同学们在实践中运用数学、科学和工程知识的能力。
- 由于强化学习已在众多不同的实践领域中获得成功应用，如自动驾驶、机器人控制等，课程教学将帮助同学们更好地理解这些相关领域的技术挑战和解决方案，从而提高大家发现、提出和解决工程问题的能力。
- 同时强化学习涉及计算机科学、数学、心理学等多个学科的知识，同学们通过学习可以实现跨学科知识的融合，综合运用技术、技能和现代工程工具来进行工程实践的能力，为自己的知识体系打下更加坚实的基础。

2 先修要求

首先，扎实的数学功底将帮助选课的同学深入理解课程中讨论的概念和算法，并为大家开创新的解决方案提供灵感和基础。选课的同学需要熟悉概率论的基本概念（例如，概率分布、贝叶斯定理、期望和方差等）、线性代数（例如，矩阵求逆和矩阵分解）以及微积分（例如，Hessian 矩阵和二阶最优性条件）。其次，丰富的编程经验将帮助大家能够将精力更多地集中在探索强化学习算法上，而不是编程语言的具体语法上。强烈推荐大家能够熟练掌握 Python 语言，因为当前几乎所有的开放强化学习代码库都是基于 Python 实现的。

本门课程虽然未将机器学习和深度学习设定为先修课程，但对于这两门课程的充分了解将帮助大家快速适应本课程的内容讨论，从而更加专注地参与关于强化学习方法的讨论，并从机器学习和深度学习的相关知识中更好地体会强化学习方法的独特。因此建议选课同学已拥有机器学习和深度学习的相关了解和知识储备。如果有同学不确定是否已满足这些先修条件，请在补退选结束前与任何教师联系，以确保本学期能够顺利完成课程要求。

3 课程内容与教学计划

为了帮助同学们广泛地接触强化学习方法，我们将尽量多得涵盖强化学习中的基本要素、重要方法和最新进展。但由于强化学习领域本身内容的深度与广度，一学期的课程注定无法面面俱到。因此期望同学们可以积极主动地阅读任课教师课后推荐的阅读材料，开阔自己的视野。我们将通过课程讲座、课堂讨论、作业和课程项目等方式共同学习。本课程将涵盖的主题包括：

1. 导论 (1 周)：我们将重点介绍本课程的基本结构和主要内容，并讨论一些课程的时间安排及考核要求。这部分的讨论将为整门课程奠定基础，使同学们对一学期的学习计划有一个全面的了解。
2. 强化学习中的基本要素 (1 周)：我们将介绍强化学习中的基本概念，例如奖励 (reward) 与回报 (return)、价值函数 (value function) 与策略 (policy) 等，并讲解强化学习方法中的常见模块。此外，还将介绍强化学习问题的分类方法，例如计划 (planning) / 控制 (control) 与学习 (learning) 等概念之间的区别，以及相应的解决方案，例如基于模型的强化学习 (model-based RL) 与无模型的强化学习 (model-free RL)。
3. 多臂老虎机问题 (约 1.5 周)：该问题是理解更为复杂的强化学习问题的起点。它可以被视为单状态 (上下文无关, context-free) 或无状态转移 (给定上下文, contextual) 的强化学习问题。我们将重点讨论多臂赌博机问题中的关键挑战，即探索 (exploration) 与利用 (exploitation) 的权衡，并讲解一系列经典的多臂老虎机算法，例如置信区间上界方法 (Upper Confidence Bound)，汤普森采样方法 (Thompson Sampling) 等，帮助有效解决这一权衡并取得理论最优解。
4. 马尔可夫决策过程 (约 1 周)：这是强化学习领域中研究最为深入的问题设置之一，因此它经常性地被错误地认为是强化学习问题的代名词。马尔可夫决策过程提供了在部分随机、部分可控的环境中进行决策的有效数学框架。我们将深入讨论其结构、潜在假设以及局限性。马尔可夫决策过程将成为我们在本课程中讨论现有强化学习算法的主要问题设定。
5. 动态规划 (约 1 周)：动态规划是计算机科学中、尤其是计算机算法设计中，众所周知的求解原则。它同时也是在环境模型完备的情况 (model-based) 下计算最优策略的基础。我们将介绍源于动态规划技术的重要概念，包括贝尔曼 (Bellman) 最优性方程、价值迭代 (value iteration) 和策略迭代 (policy iteration) 等经典方法。
6. 蒙特卡洛方法 (约 1 周)：这一方法放宽了对环境完全可见的假设，允许智能体从与环境交互中获得的经验中学习 (model-free)。我们将介绍蒙特卡洛方法在策略评价 (policy evaluation) 和控制 (control) 中的应用。同时我们还将讨论强化学习中的另一个重要方向，即利用蒙特卡洛方法进行非策略评估和学习的问题 (off-policy policy evaluation and learning)。
7. 时序差分学习 (约 2 周)：时序差分学习 (temporal difference) 是一类重要的无模型 (model-free) 强化学习方法，它结合了蒙特卡洛方法 (从经验中学习) 与动态规划 (基于模型的规划) 的思想，通过自举的方法 (bootstrapping) 进行学习。时序差分学习始终被认为是强化学习中的最基本思想之一。在课程中，我们将讲解 n 步 TD 学习方法、TD(λ) 方法以及 Q 学习 (Q-learning) 方法。
8. 策略梯度方法 (约 2 周)：策略梯度方法是另一类重要的无模型 (model-free) 强化学习方法，它通过在参数化策略空间中直接进行梯度优化，尤其适用于连续状态空间的问题。我们将介绍经典的 REINFORCE 算法、actor-critic 方法、off-policy 策略梯度方法等。
9. 深度强化学习 (约 2 周)：深度学习方法带来的强大表征学习能力为强化学习在更多实际问题中的应用带来了巨大的可能性，使其不再拘泥于表格类的有限状态空间、或简单的线性空间假设。我们将介绍几款经典的深度强化学习方法，例如 Deep Q-Learning (DQN)、Asynchronous Advantage Actor Critic (A3C)、Deep Deterministic Policy Gradient (DDPG) 等，同时建立他们与传统强化学习方法的联系。

由于强化学习方法中概念繁杂，很多专有名词的中文翻译无法简明扼要得体现其内涵，例如 off-policy learning 与 offline learning。本课程将采取**中文授课、结合英文讲义和英文教材**的方式。在课程讲解过程中也将尽量使用方法和概念的通用英文表述，以保证沟通过程无歧义。但对于随堂小测、作业、课程项目等均**不要求**英文完成。如果有同学对此有担忧或疑问，请随时联系任课教师。

4 课程考核

本课程将采用以任课教师课堂讲授为主的方式完成教学。课程设置四次轻量级的作业，帮助同学们练习经典强化学习方法。最后的课程项目将引导大家深入探索这一研究领域。通过这种教学安排帮助同学们全面地理解课程内容。

课程作业 (40%) 我们精心设计了四次以程序实践为主的课程作业，旨在引导同学们更加深入地理解多臂老虎机算法 (MP1)、使用动态规划方法求解马尔可夫决策过程 (MP2)、时序差分算法和策略梯度方法 (MP3)、以及深度强化学习方法 (MP4) 的核心算法细节。针对强化学习的特性，我们将针对具体问题使用不同的模拟环境和评价方法，例如累积收益、收敛速率等，来评价大家的算法实现优劣。与此同时，由于这些算法通常需要在在线环境中应用，算法实现的效率也会成为一个重要的评价标准。

随堂测验 (20%) 随着课程进行，我们将组织四次随堂闭卷测验，主要考察大家对于课程基础知识和核心算法的掌握。考察形式以客观题为主（选择、判断、和简答题），每次测验 25 分，总共 100 分。

AI 辅助自学 (5%) 本课程将借助课题组自主研发的多智能体辅助教学平台 AI Cosmos (aicosmos.ai) 补齐课堂授课中无法覆盖的算法细节与最新发展。AI Cosmos 平台提供全面与多样的学习进度反馈，辅助同学们对于自学进度的把控。

课程项目 (35%) 鉴于强化学习的内在复杂性，只有将课本中学到的知识应用于解决实际问题大家才会对这些概念和方法产生更加深入和独到的理解。每位同学需要应用课堂中及课外所学的关于强化学习的知识，并与另外 1-2 名学生组队、合作解决问题。我们强烈建议同学们发挥团队协作精神，群策群力共同解决问题。同时我们更希望每个小组提出的解决方案具备学术发表的潜力，例如针对某些重要的或新的问题提出独到且有效的解决方案，或者是开发出可以独立展示的原型系统。另外与任课教师和助教保持通畅的沟通，讨论你们的项目想法和进展是确保最终成功的重要途径。在学期末每个小组都需要向全班汇报所提出的解决方案，并提交一份书面报告解释各小组的工作成果。

5 课程资源

目前在互联网上已经有大量关于强化学习的视频讲座、关于具体算法的详细技术文档和博客文章、技术报告、研究论文以及开源代码实现。利用好这些在线资源来提升自己的知识和研究水平是非常重要的。

本课程的选用教材为《*Reinforcement Learning: An Introduction*》，第二版，由 Richard S. Sutton 和 Andrew G. Barto 编著，MIT 出版社，2018 年 11 月出版。我们的课程讲授将主要基于这本教材进行讨论。此外，在我们的课程主页上任课教师还列出了一些优质的在线资源，以帮助大家更好地掌握课程内容。这些资源包括在线教程、其他高校开设的类似课程、公共工具包和库，以及相关的研究论文和报告。同学们可以在我们课程的网站上找到这些资源，同时我们也欢迎大家在课程论坛中分享任何你认为对学习有帮助的材料。

6 课堂准则

如何参与到课程讨论中？ 课程讨论是实现高效教学非常重要的一环，任课教师非常享受与同学们在课堂上的讨论与争辩。没有“傻”问题，只有没有想到的问题。虽然我们不会将课堂参与计入大家的成绩，但积极的课堂参与总会帮助到每一位同学更加融洽的融入课堂学习。我们将利用不同的组织形式（如分组讨论、辩论等）调动大家的课堂参与。

什么时候开始做作业？ 每位同学将有一周的时间来完成每次作业。我们的截止日延期政策很简单：1) 所有同学在四次作业中总共有 4 天的自由延期时间；2) 大家可以将这些延期时间以一天为最小单位用于任意一次作业，每延期一天截止日期将延长 24 小时；3) 一旦用完所有的延期时间，每额外延期一天将扣除 10% 的分数（直到分数降为 0 为止）。因此，我们强烈建议大家尽早开始：由于我们的作业都需要编程，而计算机编程的特殊性往往导致问题和错误只会在最后一步出现。

每次作业如何批改？ 作业、项目展示和报告的详细评估标准将在相应的作业说明中进行详细讨论，大家可以在我们的课程网站上找到这些标准。请注意，学期末不会进行任何人为的调分，最终成绩将根据评估部分中定义的每项考核的权重进行计算。

在这门课程中不要做什么？在计算机科学领域，无论在工业界还是学术界，剽窃都是一种严重的学术不当行为：它不仅损害你的信誉，还可能涉及版权和知识产权的法律问题。因此，在我们的课程作业中，允许大家与其他同学或老师讨论，但严禁抄袭他人的代码或实现（包括在互联网上能够找到的公开实现）。我们的编程作业是为每位同学独立完成而设计的，因此**不允许共享代码/实现或合作**。一旦查实抄袭等学术不端行为，当次作业记零分。在**课程项目**中鼓励使用第三方公共库，但请在最终的项目报告中清楚地注明并解释其使用。

7 课程交流

上课时间 我们的课程安排在六教 A207 教室，每周五下午 3 点 20 分至 4 点 55 分。

答疑安排 任课教师的答疑时间为每周一下午 4 点至 5 点，FIT 3-520 办公室。助教的答疑时间为每周一晚 7 点到 8 点，FIT 4-504 实验室。为了保证大家时间利用的有效性，希望大家能够提前 24 小时向任课教师或助教预约答疑时间。

课程网页 我们的课程网页设置在https://coai.cs.tsinghua.edu.cn/Courses/RL2026/_site/index.html。课程的教学公告、幻灯片、作业指导与要求等都会在此网站同步更新。同学们将使用网络学堂提交作业和最终的项目报告。

线上讨论 同学们可以随时在网络学堂的讨论区中发帖讨论与课程相关的内容。我们平时也会使用以下微信群进行课程通知的发布，欢迎同学们扫码加入微信群，请将你的群内名称设为“班号-姓名”，如“计33-王宏宁”。



8 写在最后

非常感谢你耐心地读完我们的教学大纲！如果在这之前你还对强化学习这门课程的教授内容和方式有些许担心或不确定性，希望现在你已经有了一个非常明确的定位和目标。让我们一起更多地关注于课程本身，共同探索和学习这精彩而丰富的领域！