强化学习：理论、算法与前沿

Reinforcement Learning: Theory, Algorithms and Frontiers

MarkZZZ WeChat: MarkZZZ20XX

课程简介

强化学习（Reinforcement Learning, RL）是机器学习的核心分支，研究智能体如何在与环境的序贯交互中通过奖励信号学习最优决策策略。与监督学习不同，强化学习无需标注数据，而是通过试错（Trial and Error）与延迟反馈（Delayed Reward）驱动学习。自 Sutton & Barto 于 1988 年奠定现代强化学习理论框架以来，该领域经历了从经典动态规划到深度强化学习的深刻变革，并在 Atari 游戏、围棋、蛋白质结构预测、大语言模型对齐（RLHF）等领域取得了历史性突破。

本课程以马尔可夫决策过程（Markov Decision Process, MDP）为核心形式体系，系统讲授强化学习的数学基础与核心算法。课程内容涵盖 MDP 理论（Bellman 方程、压缩映射）、动态规划（值迭代、策略迭代）、蒙特卡洛方法（首次访问/每次访问 MC，重要性采样）、时序差分学习（TD(0)、Q-learning、Sarsa）、资格迹与 TD(λ)（前向-后向等价性完整证明）、函数逼近理论（线性 TD 收敛性、致命三角、梯度 TD）、深度 Q 网络（DQN 及其变体）、策略梯度（策略梯度定理完整证明、REINFORCE、Actor-Critic）、近端策略优化（TRPO 单调改进定理、PPO、GAE）、模型基强化学习（Dyna、World Models、MBPO）以及多智能体 RL 与逆向 RL（IRL、RLHF）。

课程定位为博士研究生水平，强调严格的数学推导与完整的收敛性证明，同时结合现代深度强化学习前沿方法，帮助学员建立从经典 MDP 理论到现代深度 RL 实践的完整知识体系。

适合人群

- 人工智能、机器学习、运筹学、统计学等方向的博士研究生
- 从事强化学习、最优控制、序贯决策研究的科研人员
- 希望深入理解强化学习数学基础（MDP、策略梯度定理、收敛性证明）的高年级本科生
- 对 RLHF、多智能体 RL、模型基 RL 等前沿方向有研究兴趣的从业者

前置知识

- **概率论**：概率空间、随机变量、条件期望、鞅、马尔可夫链、大数定律
- **实分析**：σ-代数、Lebesgue 积分、收敛定理（单调收敛、控制收敛）
- **泛函分析基础**：Banach 空间、压缩映射定理（Banach 不动点定理）、算子范数
- **线性代数与矩阵论**：特征值、谱范数、正定矩阵、矩阵收敛
- **优化理论**：凸函数、梯度下降、随机梯度下降（SGD）、Lagrange 对偶
- **随机逼近**（有则更佳）：Robbins–Monro 算法、ODE 方法、Borkar–Meyn 定理

1 课程内容

讲次	主题	内容概要
1	MDP 形式化	马尔可夫决策过程（MDP）的测度论框架；状态/动作/转移/奖励的严格定义；策略类层次结构；Bellman 期望方程的推导；折扣/平均报酬准则；有限 MDP 的矩阵形式
2	值函数理论	Bellman 最优方程；Bellman 最优算子的压缩性（γ-压缩，ℓ^∞ 范数）；Banach 不动点定理；最优值函数的存在性与唯一性完整证明；平稳最优策略的存在性定理
3	动态规划算法	值迭代的收敛速度分析与误差界（a posteriori 与 a priori 估计）；策略评估方程的解析解；策略迭代的单调改进与有限步收敛定理；修正策略迭代；线性规划方法
4	蒙特卡洛方法	首次访问 MC 与每次访问 MC 的估计量及收敛性；MC 控制（GPI 框架）；重要性采样（Ordinary IS 与 Weighted IS）及方差分析；off-policy MC 评估
5	时序差分学习	TD(0) 的随机逼近视角与 Robbins–Monro 收敛条件；Sarsa 的收敛性定理；Q-learning（Watkins, 1989）的收敛性完整证明；on-policy vs off-policy 的统一理解
6	n 步方法与资格迹	n 步 TD 与 n 步 Sarsa；资格迹（Eligibility Traces）的向量化表示；TD(λ) 前向观点（λ-return）与后向观点（TD(λ)）等价性的完整数学证明；True Online TD(λ)
7	函数逼近理论	线性函数逼近下的半梯度 TD 收敛性（Tsitsiklis–Van Roy 定理）；致命三角（Deadly Triad）分析；梯度 TD（GTD2, TDC）的收敛性；非线性函数逼近的挑战
8	深度 Q 网络	DQN（Mnih et al., 2015）的经验回放与目标网络稳定化原理；Double DQN（van Hasselt et al.）；Prioritized Experience Replay；Dueling Network；Rainbow 集成
9	策略梯度方法	策略梯度定理的完整证明（Sutton et al., 1999）；REINFORCE 算法的方差分析；基线（Baseline）的最优选取；单步 Actor-Critic；兼容函数逼近定理

讲次	主题	内容概要
10	近端策略优化	TRPO 单调改进定理（Schulman et al., 2015）的完整证明；KL 约束与 Clip 版本；广义优势估计（GAE，Schulman et al., 2016）；PPO 的实现技巧与稳定性
11	模型基强化学习	Dyna-Q 框架；采样效率分析；World Models（Ha & Schmidhuber, 2018）；MBPO（Janner et al., 2019）中模型误差的传播界；规划与学习的统一视角
12	多智能体 RL 与逆向 RL	Nash 均衡下的多智能体 RL（minimax Q-learning）；逆向强化学习（Abbeel & Ng, 2004）的最大边界方法；RLHF（Christiano et al., 2017）的理论框架；RLHF 与奖励学习

2 参考资料

1. **Sutton, R. S. & Barto, A. G.** *Reinforcement Learning: An Introduction.* 2nd ed., MIT Press, 2018.
   强化学习领域最权威的入门教材，覆盖从 MDP、TD 学习到函数逼近的完整体系，是本课程的核心参考。
2. **Szepesvári, Cs.** *Algorithms for Reinforcement Learning.* Morgan & Claypool, 2010.
   简洁而严格的数学处理，对 TD 学习和函数逼近的收敛性证明尤为深刻。
3. **Bertsekas, D. P.** *Reinforcement Learning and Optimal Control.* Athena Scientific, 2019.
   从动态规划角度系统阐述强化学习，与最优控制理论紧密联系。
4. **Bertsekas, D. P. & Tsitsiklis, J. N.** *Neuro-Dynamic Programming.* Athena Scientific, 1996.
   神经动态规划（近似 DP）的奠基性著作，对函数逼近的理论分析极为深刻。
5. **Puterman, M. L.** *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, 2005.
   MDP 理论的标准参考，对策略迭代、线性规划方法的数学处理极为完整。
6. **Lattimore, T. & Szepesvári, Cs.** *Bandit Algorithms.* Cambridge University Press, 2020.
   多臂赌博机理论的现代教材，对探索-利用权衡的遗憾分析有深刻的数学处理。
7. **Agarwal, A., Jiang, N., Kakade, S. M. & Sun, W.** *Reinforcement Learning: Theory and Algorithms.* 2022. (Online)
   现代理论 RL 的最新教材，涵盖样本复杂度、探索与函数逼近的最新理论成果。

3 评分标准

- **课后作业**（Lesson 习题集）：60%——每讲含 7–8 道习题（ / / 难度分级），着重考察定理证明与算法分析能力。

- **期末项目**：40%——选择一个前沿方向（深度 RL 算法实现与理论分析，逆向 RL，多智能体 RL，RLHF 理论等），完成一份 15–20 页的研究报告。