REINFORCEMENT LEARNING FOR DECISION-MAKING AND CONTROL

Full Chapters

Shengbo Eben Li
(李升波)

Email: lishbo@tsinghua.edu.cn
Intelligent Driving Laboratory (iDLab)
www.idlab-tsinghua.com

TSINGHUA UNIVERSITY
Beijing, 100084, China

DECEMBER 25, 2021

About the Author

Dr. Li is the full professor at Tsinghua University. He is now leading Intelligent Driving Laboratory (iDLab) at School of Vehicle and Mobility. His active research interests include intelligent vehicles and driver assistance, reinforcement learning, optimal control and estimation, etc.

He is the author of over 130 peer-reviewed journal/conference papers, and the co-inventor of over 30 patents. Dr. Li is the recipient of best paper (student) awards of IEEE ITSC 2020/2021, ICCAS 2020 IEEE ICUS 2020, CCCC 2018/2019, ITSAPF 2015, IEEE ITSC 2014, etc. His academic services include Member of Board of Governors of IEEE ITS Society, Senior Editor of IEEE OJ ITS, AEs of IEEE ITSM, IEEE Trans ITS, JICV, and Automotive Innovation, etc.

Contact information:

Shengbo Eben Li（李升波）

Ph.D., Tenured Professor

Rm 643, Lee Shau Kee Sci&Tech Building,

Tsinghua University, 100084, Beijing

Office: 86-10-62796150

Email: lishbo@tsinghua.edu.cn

Website: www.idlab-tsinghua.com

Preface/前言

自 2000 年以来，人工智能的快速崛起正重塑人类社会的各个角落，有望引导工业文明进入第四次革命浪潮。以道路交通为例，汽车的智能化变革促使整个行业发生了翻天覆地的变化，包括驾驶辅助、自动驾驶、云控协同等一系列新技术如雨后春笋般涌现，它们在提升地面车辆行驶性能的同时，也为解决交通事故、排放污染、城市拥堵等问题提供了一条可行的途径。近年随着机器学习和自动控制的融合发展，以模仿人类大脑学习机制为原理的强化学习（RL，Reinforcement Learning）技术迅速进入人们的视野，它为大规模复杂动态系统的高性能决策与高实时控制提供了一套极具前景的解决方案。一个引人注目的成功案例是以 Alpha Go 为代表的围棋智能：它利用深度强化学习算法实现围棋智能的自我进化，自我超越，以超乎想象的速度进化出打败人类专业棋手的能力，引发学术界和工业界的热切关注。

尽管强化学习具有优异的潜在优势，但是该方法的工程应用尚属于起步阶段。一个重要的原因是该方法既具有理论学习的复杂度，又具有工程实践的挑战性。该方法隶属于统计学习、最优控制、最优化三者的交叉结合部，涉及的数理基础较深，知识涵盖面较广，难以学习更难工程应用。入门者不易短期内掌握关联的理论体系，若对原理不够熟悉，难以对代码进行针对性调整，不能发挥算法应有的性能。为了应对上述挑战，依托笔者在清华大学开设的研究生课程《强化学习与控制》，撰写了这一本教学参考书，主要面向工程应用领域的科研工作者和技术人员，按照原理剖析、主流算法、典型示例的架构，介绍用于动态系统的决策与控制的强化学习方法。所涉及的知识点包括马尔科夫决策、蒙特卡洛学习、时序差分学习、函数近似学习、策略梯度学习、近似动态规划、深度强化学习等。希望本书为该领域的行业同仁，包括研究生和本科生，提供一本体系较为全面，且适合研究者学习和应用的参考书籍。

全书总共包括 11 章。第 1 章介绍 RL 概况，包括发展历史、知名学者、典型应用以及主要挑战等。第 2 章介绍 RL 的基础知识，包括定义概念、自洽条件、最优性原理与问题架构等。第 3 章介绍免模型学习的蒙特卡洛法，包括 Monte Carlo 估计，On-policy/Off-policy，重要性采样等。第 4 章介绍免模型学习的时序差分法，包括它衍生的 Sarsa，Q-learning，Expected Sarsa 等算法。第 5 章介绍带模型学习的动态规划法，包括策略迭代、值迭代、通用迭代架构与收敛性证明等。第 6 章介绍间接型 RL 的函数近似方法，包括常用近似函数，值函数近似，策略函数近似以及所衍生的 Actor-critic 架构等。第 7 章介绍直接型 RL 的策略梯度法，包括 On-policy gradient，Off-policy gradient，以及它们的代价函数、优化算法等。第 8 章介绍带模型的强化学习，即近似动态规划（ADP），包括离散时间系统的 ADP，连续时间系统的 ADP，以及 ADP 与 MPC 的联系与区别等。第 9 章介绍有限时域的近似动态规划，同时探讨了状态约束的处理手段以及它与求解可行性之间的关系。第 10 章介绍深度强化学习，即以神经网络为载体的 RL，包括神经网络的原理与训练，深度化挑战以及 DQN、DDPG、TD3、TRPO、DSAC 等典型深度化算法。第 11 章介绍 RL 的各类拾遗，包括鲁棒性、POMDP、多智能体、元学习、逆强化学习、离线强化学习、训练框架与平台等。

借此机会，诚挚感谢清华大学智能驾驶课题组的老师和同学们，他们为本书的撰写付出了极大的心血与努力，包括文献查阅、文字编排、公式推导、示例代码等

一系列工作，贡献良多。还有很多师长、同仁和朋友，也为本书的撰写和完善提供了大量宝贵地建议，受篇幅所限不能一一赘述，借此机会深表感谢。

受笔者的水平所限，书中难免存在疏漏和不足，包括图片版权冲突、公式推导错误，文字写作粗陋等不一而足。借此机会，恳请国内外同仁给予批评和指正。如有发现任何错误，烦请将意见发送给笔者邮箱：lishbo@tsinghua.edu.cn。每一位读者的反馈将是本书进一步完善的重要保障，提前谢谢大家！

Preface/前言

From the beginning of the 21st century, artificial intelligence (AI) is reshaping almost all areas of human society, which is promising to spark the fourth industrial revolution. Noticeable examples can be found in the sector of road transportation, where AI has drastically changed automobile design and traffic management. Lots of new technologies, such as driver assistance, autonomous driving, and cloud-based cooperation, are rising in an unbelievable speed. These technologies have the potential to significantly improve driving ability, reduce traffic accidents, and relieve urban congestion.

As one of the most important AI branches, reinforcement learning (RL) is attracting increasing attention in the past decades. RL is an interdisciplinary field of trail-and-error learning and optimal control, which provides a promising solution for decision-making and control of large-scale and complex dynamic processes. One of its most eye-catching success is AlphaZero from Google DeepMind, which beats the most professional human player. The key technology behind is called deep reinforcement learning, which equips AlphaGo with an amazing self-evolution ability.

Despite a few success, the application of RL is still in its infancy stage. This is because most RL algorithms are difficult to comprehend. In one hand, RL deeply connects with statistic learning and convex optimization, and involves a wide range of concepts and theories. On the other hand, it is a tedious and long learning process for a beginner to become an RL master. Without fully understanding those principles, it is very difficult for users to make necessary adjustments to achieve the best performance. This book aims to provide a systematic introduction of fundamental RL theories, mainstream RL algorithms and typical RL applications to fellow researchers and engineers. The topics mainly include Markov Decision Processes, Monte Carlo learner, Temporal Difference learner, RL with function approximation, policy gradient method, approximate dynamic programming, deep reinforcement learning, etc.

The book contains 11 chapters. Chapter 1 provides an overview of RL, including its history, famous scholars, successful examples and up-to-date challenges. Chapter 2 briefs the basis of RL, including its concepts, optimality conditions, and problem formulation. Chapter 3 introduces Monte Carlo methods for model-free RL, including on-policy/off-policy and importance sampling technique. Chapter 4 introduces temporal-difference methods for model-free RL, including Sarsa, Q-learning, expected Sarsa, etc. Chapter 5 introduces stochastic dynamic programming, i.e., model-free RL with tabular representation, including value iteration, policy iteration and their convergence mechanism. Chapter 6 introduces how to approximate policy and value function in indirect RLs, as well as its associated actor-critic architecture. Chapter 7 derives different kinds of direct policy gradients, including likelihood ratio gradient, natural policy gradient and a few variants. Chapter 8 introduces infinite horizon ADP, and its connection with model predictive control. Chapter 9 introduces finite-horizon ADP, and puts great emphasis on how to handle state constraints. Chapter 10 devotes to deep reinforcement learning, including artificial neural networks and typical deep RL algorithms like DQN, DDPG, TD3, TRPO, DSAC, etc. Finally, Chapter 11 provides various RL relics, including robust RL, POMDP, multi-agent RL, meta-RL, inverse RL, offline RL, major RL libraries and platforms, etc.

In closing, I wish to offer my sincere gratitude to all the faculties and students in Intelligent Driving Laboratory (iDLab) for their great contribution to this book. I also express my deep appreciation to those friends and colleagues, who support writing and

polishing this book. They have provides numerous priceless suggestions. Any comments and corrections from readers would be much appreciated. I look forward to seeing your email at lishbo@tsinghua.edu.cn. Thanks a lot in advance!

Contents