Deep Reinforcement Learning: Principles and Practices

深度强化学习

原理与实践

陈仲铭
何　明　著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

Deep Reinforcement Learning: Principles and Practices

深度强化学习

原理与实践

陈仲铭
何　明　　著

人民邮电出版社
北　京

图书在版编目（CIP）数据

深度强化学习原理与实践 / 陈仲铭，何明著. -- 北京 : 人民邮电出版社，2019.5
ISBN 978-7-115-50532-3

Ⅰ. ①深… Ⅱ. ①陈… ②何… Ⅲ. ①机器学习－研究 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2019)第000679号

内 容 提 要

本书构建了一个完整的深度强化学习理论和实践体系：从马尔可夫决策过程开始，根据价值函数、策略函数求解贝尔曼方程，到利用深度学习模拟价值网络和策略网络。书中详细介绍了深度强化学习相关最新算法，如 Rainbow、Ape-X 算法等，并阐述了相关算法的具体实现方式和代表性应用（如 AlphaGo）。此外，本书还深度剖析了强化学习各算法之间的联系，有助于读者举一反三。

本书分为四篇：初探强化学习、求解强化学习、求解强化学习进阶和深度强化学习。涉及基础理论到深度强化学习算法框架的各方面内容，反映了深度强化学习领域过去的发展历程和最新的研究进展，有助于读者发现该领域中新的研究问题和方向。

本书适用于计算机视觉、计算机自然语言的相关从业人员，以及对人工智能、机器学习和深度学习感兴趣的人员，还可作为高等院校计算机等相关专业本科生及研究生的参考用书。

◆ 著　　陈仲铭　何　明
责任编辑　张　爽
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：22.5　　彩插：4
字数：508 千字　　2019 年 5 月第 1 版
印数：1－3 000 册　　2019 年 5 月北京第 1 次印刷

定价：99.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

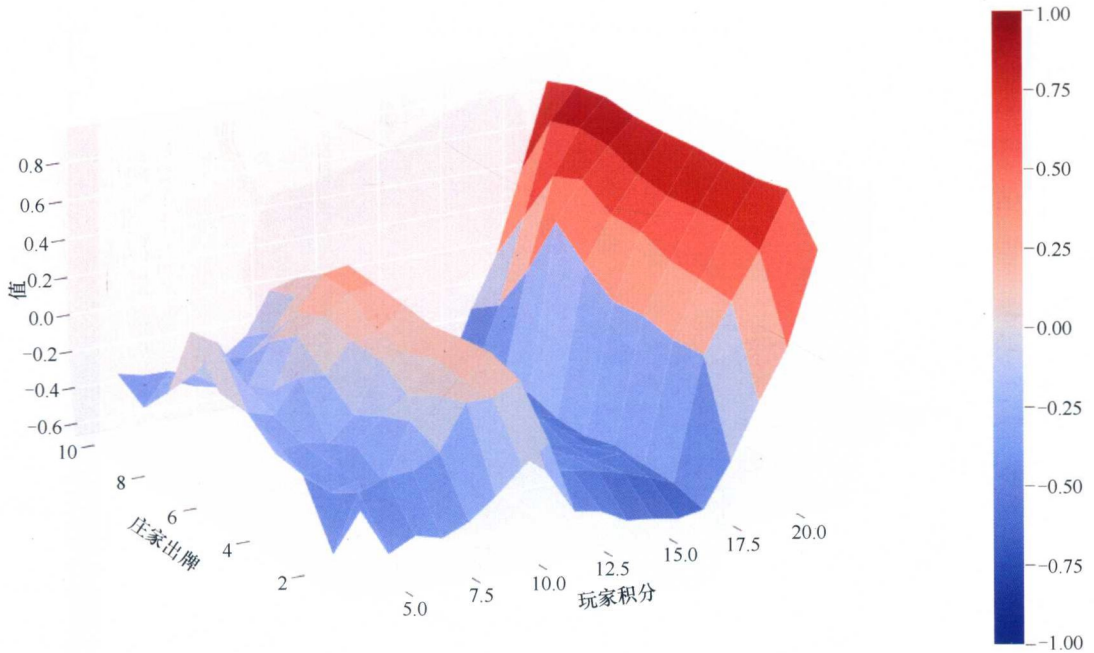


图4.3　首次访问蒙特卡洛预测算法在21点游戏中状态值的三维示例图

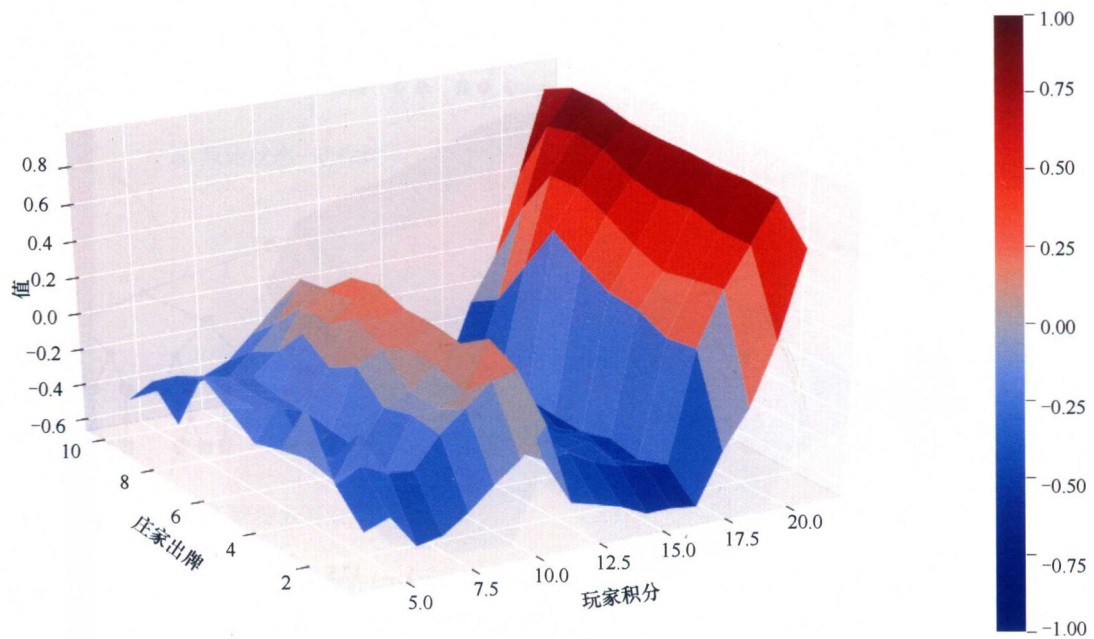


图4.4　每次访问蒙特卡洛预测算法在21点游戏中的状态值三维示例图

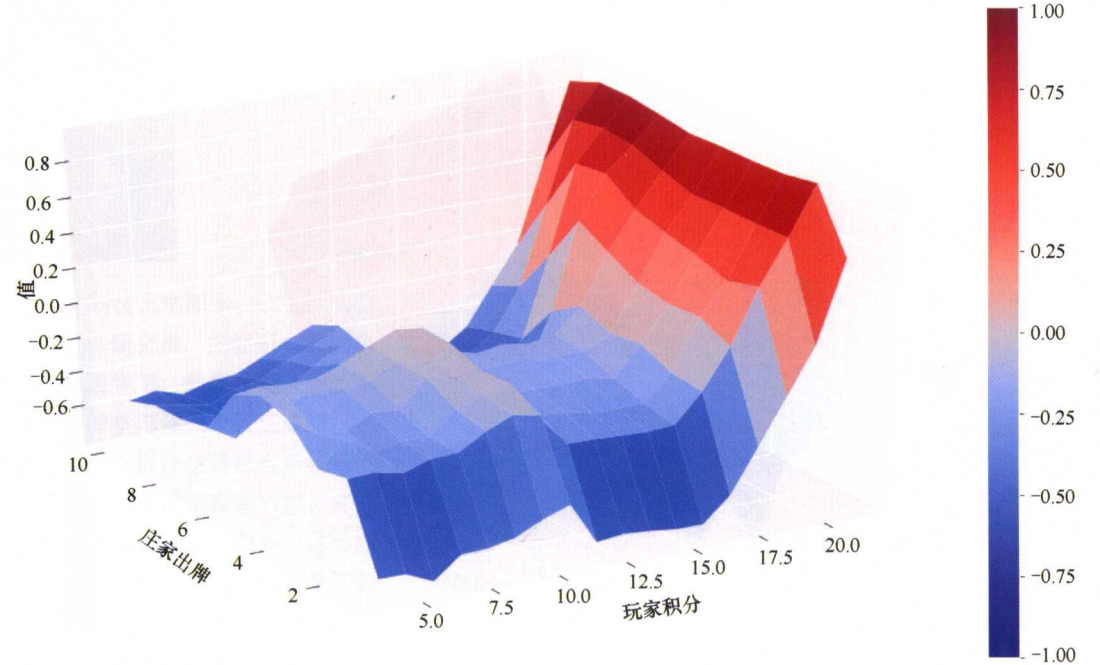


图4.7　固定策略的非起始点探索的蒙特卡洛控制21点游戏的对应价值

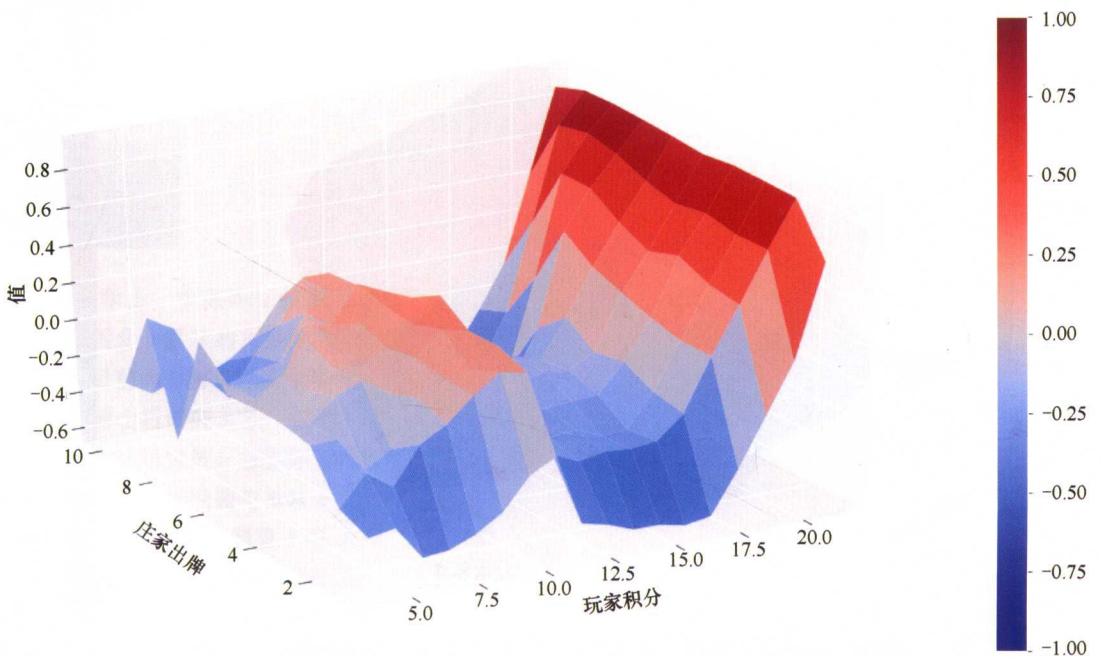


图4.8　非固定策略的非起始点探索的蒙特卡洛控制21点游戏对应的状态值

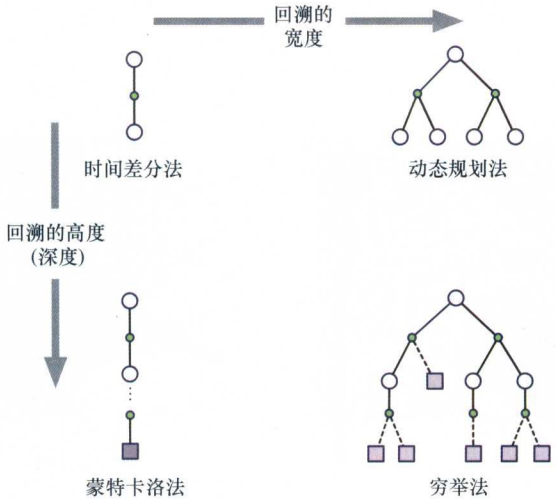
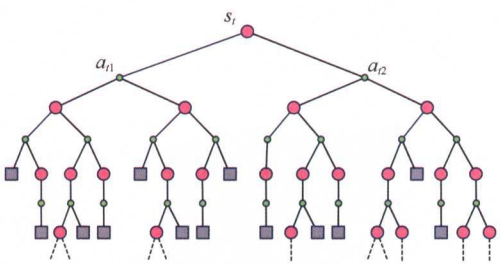
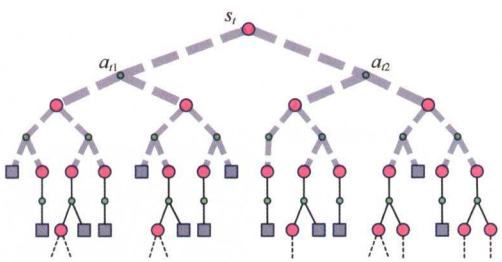


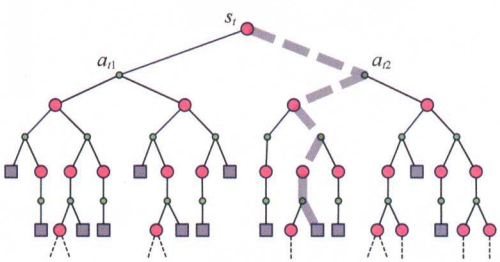
图5.7　强化学习求解方法的差异



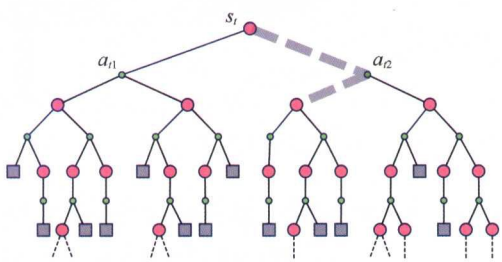
a) 原始状态–动作树



b) 动态规划法

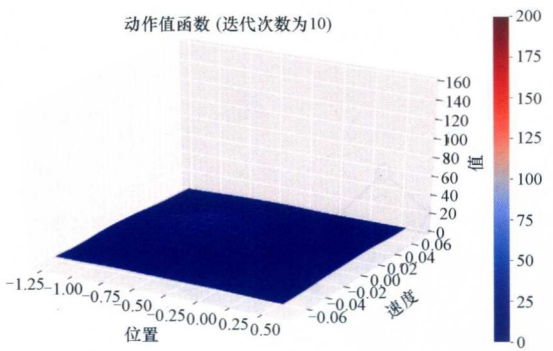


c) 蒙特卡洛法

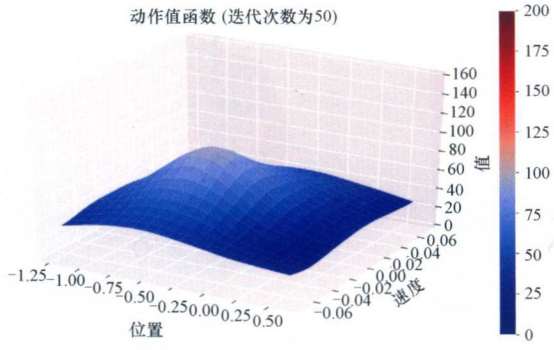


d) 时间差分法

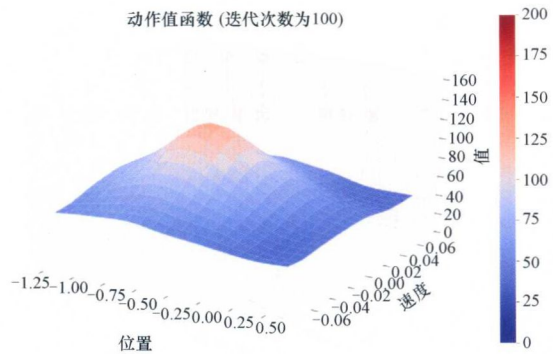
图5.8　使用树状结构模拟马尔可夫决策过程，综合对比动态规划法、蒙特卡洛法、时间差分法的差异[Sutton el al. 1998]



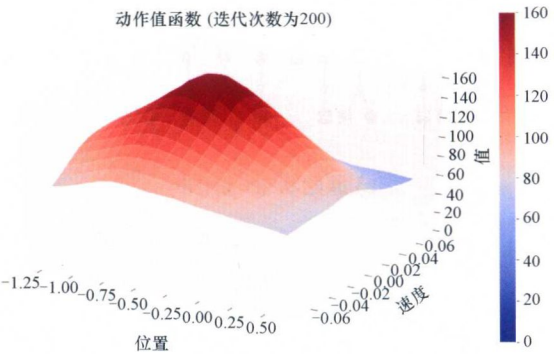
a) 迭代10次



b) 迭代50次

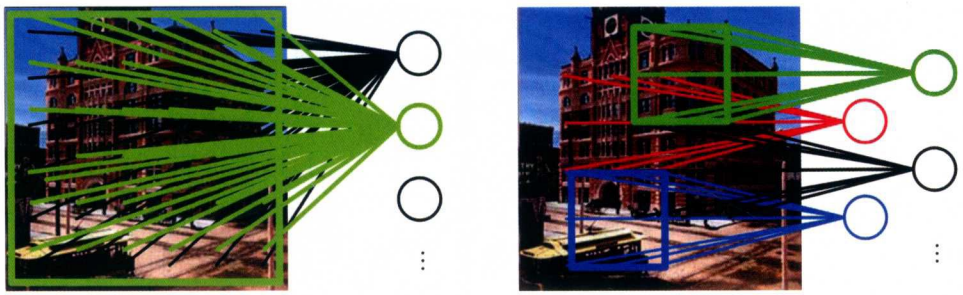


c) 迭代100次



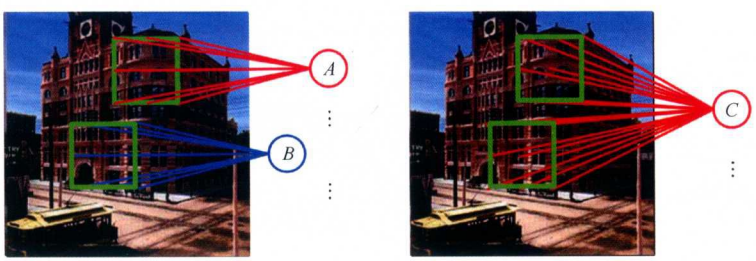
d) 迭代200次

图 6.9　爬山车游戏中，基于值函数近似法的不同迭代次数的动作值函数对比图。x轴为小车当前所在位置，y轴为小车当前速度，z轴为小车位置和速度对应的状态值



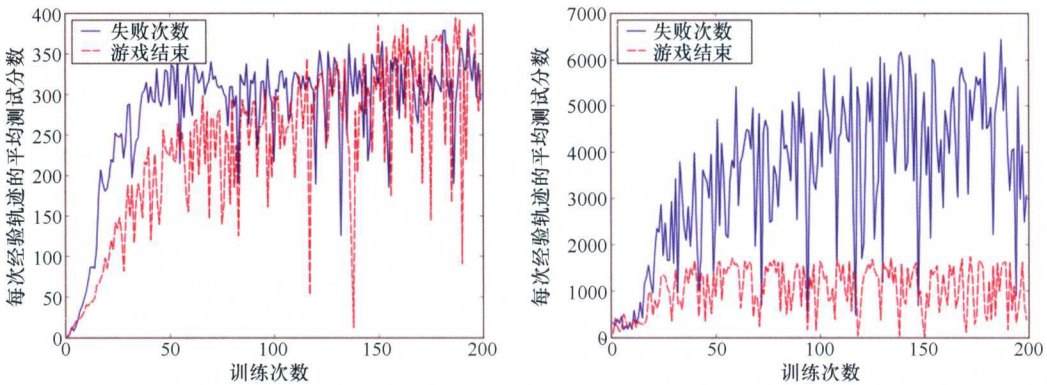
a) 全连接神经网络　　b) 局部连接神经网络

图9.12　全连接与局部连接的对比示例



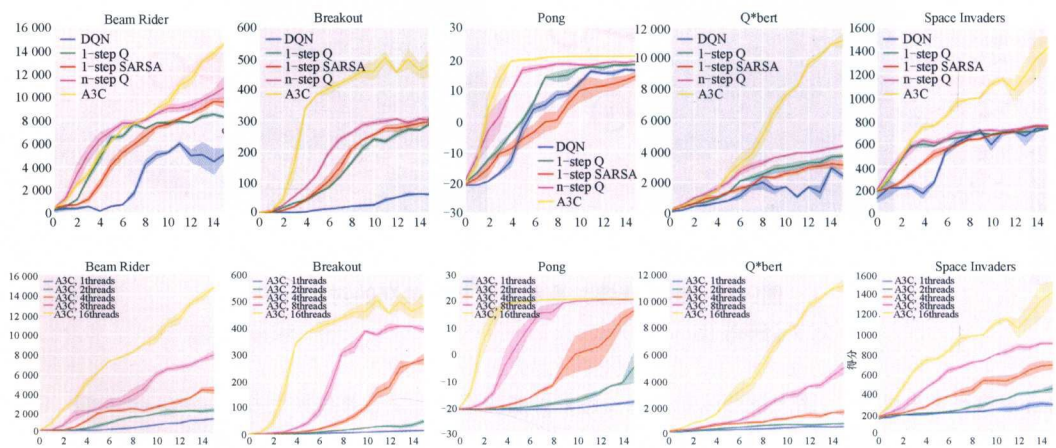
a) 没有权值共享　　b) 权值共享

图9.13　带有权值共享和没有权值共享的对比示例



a) Breakout游戏　　b) Seaquest游戏

图10.7　在Breakout 和 Seaquest 游戏中，基于失败次数和游戏结束的结果对比[Roderick et al. 2017]



注：各图横坐标为训练时间（单位：小时），纵坐标为得分。

图11.4　A3C算法实验对比图[Mnih et al. 2016]

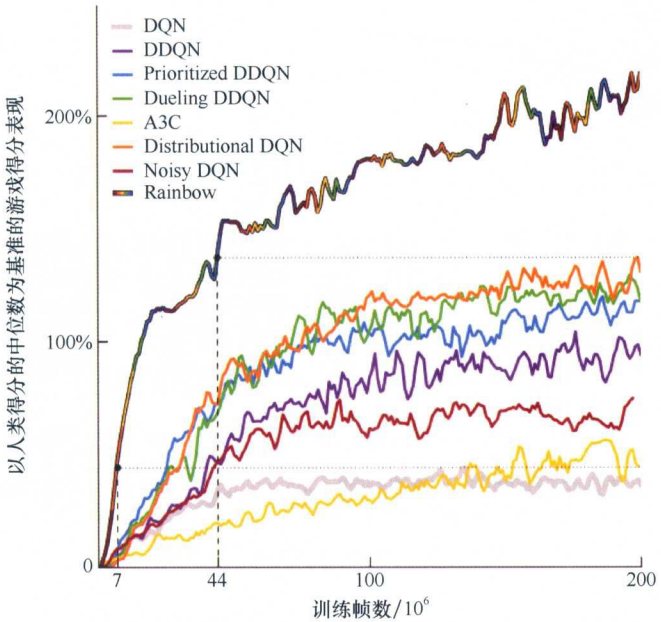


图11.5　Rainbow模型测试结果[Matteo et al. 2017]

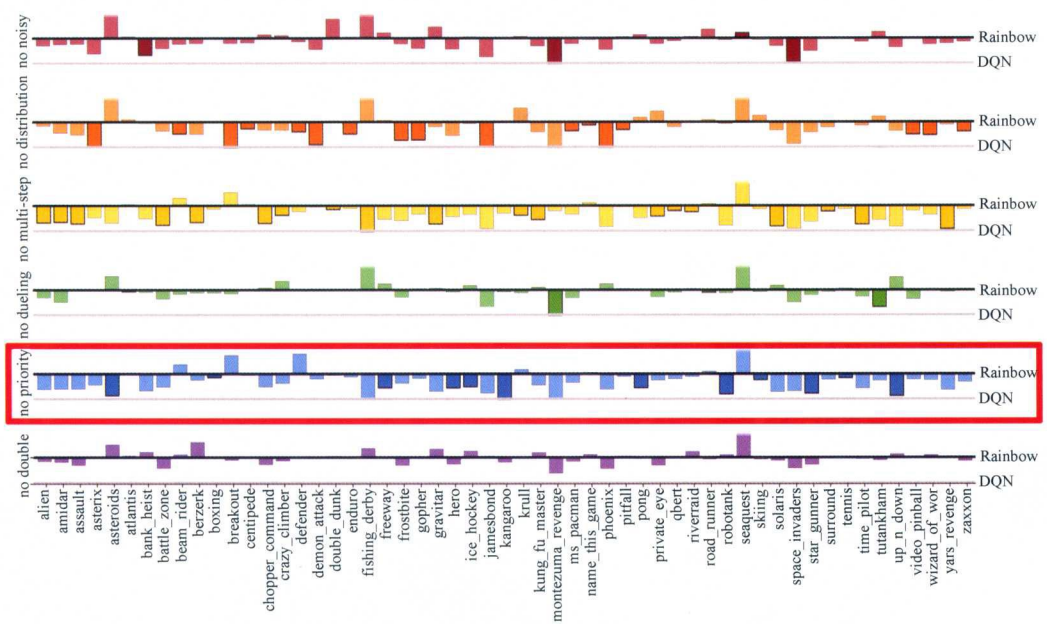
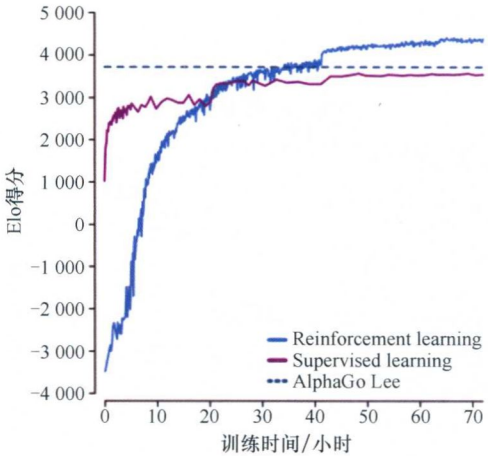
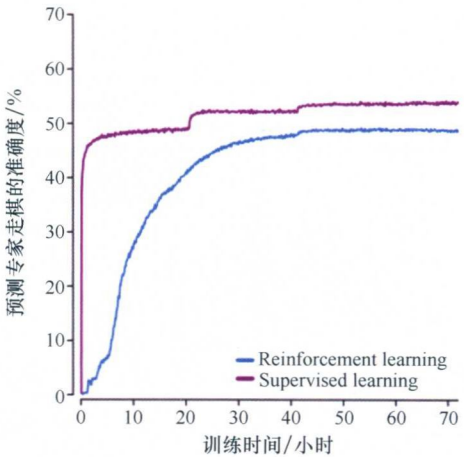


图11.6　Rainbow元素对比图[Matteo et al. 2017]



a）AlphaGo Zero随着训练时间的棋力增长趋势



b）预测专家走棋的准确度变化趋势

图12.13　AlphaGo Zero训练结果

序　一

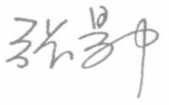
强化学习是一门具有50多年历史的学科。该学科从生物学的试错方式和数学的最优控制问题开始萌芽。直到20世纪90年代，受马尔可夫决策过程理论的影响，强化学习的现代形式才逐渐兴起和趋于完善，并于20世纪90年代后期在Sutton和Barto的努力下，建立了完整的学科体系。

近年来，DeepMind团队提出了第一个深度强化学习算法（DQN），开发出了首次战胜人类职业棋手的围棋程序（AlphaGo）。受到DeepMind团队关于深度强化学习研究的影响，深度强化学习领域得到空前关注。据统计，在国际机器学习大会（ICML 2018）提交的论文中，强化学习相关的论文提交数量仅次于深度学习，成为ICML 2018第二大研究主题。

深度强化学习是深度学习和强化学习的结合，这两种学习方式在很大程度上是正交的，其数学结合方式非常优美。强化学习需要通过数据逼近函数的方法来部署价值函数、策略、环境模型和更新状态，而深度学习则是近年来最热、最成功的函数逼近器，两者的结合能够显著提升深度强化学习的应用范围。另外，在人工智能算法中，理想的智能系统能够在不接受持续监督的情况下自主学习、自主判断对错，而深度强化学习正是其中的最佳代表之一。

《深度强化学习原理与实践》一书对深度强化学习的基本概念、原理和应用技术做了深入浅出的讲解。相信本书的出版会对从事人工智能相关研究的工作者和研究人员大有裨益，能够在一定程度上促进国内深度强化学习的研究和应用。希望我国能够有更多的研究者参与到科研工作中，同时我也很高兴可以看到我国新一代人工智能创新活动的蓬勃发展。

中国科学院院士



序　二

AlphaGo围棋程序在人机大战中所取得的成绩，比专家预测的人工智能在围棋领域战胜人类职业选手的时间提前了10年，这极大地突破了人类已有的认知。同时，这一成绩使得AlphaGo围棋程序背后的核心技术——深度强化学习，第一次大范围地进入大众视野。

理实交融，相辅相成。深度强化学习是一种理论与实践结合较为紧密的技术，缺一不可。一方面，在深度强化学习中，涉及大量的基础知识，如动态优化、贝尔曼方程、蒙特卡洛采样等；另一方面，深度强化学习的核心组成（即强化学习和深度学习）均为实践性较强的技术。从业者想要系统性地掌握深度强化学习，需要很好地兼顾理论与实践两个方面。

目前，关于深度强化学习的参考书还较为缺乏，能够兼顾理论与实践两个方面的参考书更是少之又少，而《深度强化学习原理与实践》这本书就较好地兼顾了理论与实践两个部分。该书在对深度强化学习原理进行系统性梳理和介绍的基础上，给出了众多重要算法的代码实现，内容丰富且翔实。

相信《深度强化学习原理与实践》的出版，不仅能够为从业者带来更多的理论参考与指导，而且能够为深度强化学习算法的落地提供更好的实践参考与指导，使得深度强化学习在更多领域开花、结果。

中国科学技术大学大数据学院常务副院长

序　三

近年来，深度强化学习在工业界呈现“星火燎原”之势。从围棋对弈到自动工业机器人，从个性化电商到自动驾驶，背后都依靠基于深度强化学习的智能决策系统。智能手机，作为与用户联系最为紧密的终端设备，尤为需要基于深度强化学习所提供的自学习能力。

事实上，在深度强化学习中，智能体通过与环境进行交互并动态地完善自身动作策略的学习模式，与用户使用终端设备的行为习惯极为类似，这使得在智能终端落地深度强化学习拥有天然的优势。另外，通过充分利用深度强化学习所具备的感知优势和决策优势，能够更好地捕捉与刻画终端用户的兴趣漂移和行为变化。基于此，OPPO研究院针对深度强化学习技术与终端的有机融合做了大量的研究与实践，进而为用户提供更个性、更智能的使用体验。

“工欲善其事，必先利其器”。将深度强化学习落地到实际应用场景，既需要系统性地了解其背后的算法原理，又需要具备良好的工程实践能力。在《深度强化学习原理与实践》一书中，两位作者既给出了详细的理论介绍，同时也给出了大量的算法代码实现。相信通过对此书的学习，读者能够较好地理解和掌握这两方面的内容。

OPPO作为全球性的智能终端制造商和移动互联网服务提供商，将持续探索更多前沿技术（如生物科技、量子通信等）与手机融合的可能性，以持续提升OPPO在终端和互联网服务的智能化程度。

OPPO研究院院长

刘畅

前　言

编写背景

2014年1月，Google斥巨资收购了位于英国伦敦的人工智能公司——DeepMind。DeepMind在深度强化学习领域中，设计出第一个深度强化学习算法DQN，并开发出战胜了人类最为顶尖的围棋职业选手李世石的AlphaGo围棋程序，震惊了世人。

随着AlphaGo的成名，深度强化学习开始吸引众多研究者的关注和研究。大量与深度强化学习相关的技术论文开始出现在人工智能领域的学术会议上，如IJCAI（国际人工智能联合会议）、AAAI（美国人工智能协会年会）、ICML（国际机器学习大会）和NIPS（神经信息处理系统大会）等。此外，越来越多的企业也开始加码对深度强化学习的布局和研究，致力于降低深度强化学习的准入门槛。如Google于2018年开源的深度强化学习框架——多巴胺（Dopamine），旨在为入门或资深的深度强化学习研究人员提供具备灵活性、稳定性和可重复性的研究平台。

不可否认的是，深度强化学习在实际应用中依然存在着一定的约束和弊端，如面临维数灾难、奖励稀疏等挑战。但基于深度强化学习所拥有的强大表征优势和决策优势，能够为人工智能领域的发展带来更多的可能：医疗领域，通过深度强化学习能够对恶性肿瘤进行精确检测，其检测准确率比普通医生提高了20%；自动驾驶领域，通过深度强化学习能够进一步提升出行和驾驶体验；智能终端领域，通过深度强化学习能够让数字设备更加人性化。

回顾过去十年，云计算的兴起和数据的爆炸式增长，极大地推动了深度强化学习的发展。尤其是随着越来越多从业者的加入和研究，相信深度强化学习能够在更多领域取得如AlphaGo一样的成就。

“数风流人物，还看今朝！”

本书结构

本书包含12个章节和5个附录，其中第1~8章围绕强化学习领域，第9~12章围绕深度强化学习领域，附录A~附录E主要介绍深度学习相关的基础知识。基于章节之间的逻辑关系，本书将12个章节分成四篇（核心为第二~四篇），接下来对这四篇内容分别进行简要介绍。

第一篇（第1~2章）

这部分主要围绕强化学习的概念和基础框架，包括其基本概念和数学原理。该部分介绍的基础知识将贯穿全书，尽管涉及的数学公式和推导方程稍显复杂，但有助于深度理解强化学习的基础概念。

第1章按顺序依次介绍强化学习的发展历史、基础理论、应用案例、特点与未来。从强化学习的发展历史中可以了解强化学习与机器学习之间的关系；基础理论可以帮助读者对强化学习有一个整体的认识与了解，通过具体的应用案例可以了解如何对强化学习进行落地应用。最后，从宏观角度对强化学习的特点与未来进行了讨论。第2章则集中介绍强化学习涉及的数学概念，从马尔可夫决策过程对强化学习任务的表示开始，到介绍价值函数和策略。其中，价值函数是强化学习的核心，后续章节的大部分求解方法都集中在价值函数的逼近上。

第二篇（第3~5章）

这部分主要探讨如何通过数学求解获得强化学习的最优策略。对于基于模型的强化学习任务可以使用动态规划法，对于免模型的强化学习任务可以使用蒙特卡洛法和时间差分法。值得注意的是，本部分对于强化学习任务的求解使用的是基于表格的求解方法。

第3章介绍使用动态规划法求解强化学习任务，通过策略评估和策略改进的迭代交互计算方式，提出了用以求解价值函数和策略的策略迭代算法。然而策略迭代算法存在效率低、初始化随机性等问题，研究者又提出了值迭代算法。由于实际情况中不一定能够获得完备的环境知识，因此出现了第4章的针对免模型任务的强化学习求解方法。其中，蒙特卡洛求解法基于采样的经验轨迹，从真实/仿真的环境中进行采样学习，并分别从蒙特卡洛预测、蒙特卡洛评估到蒙特卡洛控制进行了详细介绍。事实上，蒙特卡洛法同样存在一些不足，如使用离线学习方式、数据方差大、收敛速度慢等，这会导致在真实环境中的运行效果并不理想。第5章中引入了在线学习的时间差分法，主要分为固定策略的Sarsa算法和非固定策略的Q-learning算法。需要注意的是，Q-learning算法将作为深度强化学习（即第四篇）中的基础算法之一。

第三篇（第6~8章）

动态规划法、蒙特卡洛法、时间差分法都属于基于表格的求解方法。近似求解法通过寻找目标函数的近似函数，大大降低了表格求解法所需的计算规模和复杂度。近似求解方法主要分为3种：基于价值的强化学习求解法——值函数近似法；基于策略的强化学习求解法——策略梯度法；基于模型的强化学习求解法——学习与规划。

第6章详细介绍了基于价值的强化学习任务求解方法，即对价值函数进行近似求解。通过对函数近似进行数学解释，来引入值函数近似的数学概念和值函数近似法。然而，基于值函数近似的方法难以处理连续动作空间的任务，因此有了第7章介绍的策略梯度法。其将策略的学习从概率集合变换成策略函数，通过求解策略目标函数的极大值，得到最优策略。第8章为基于模型的强化学习，智能体从真实的经验数据中学习环境模型，并基于该环境模型产生的虚拟经验轨迹进行规划，从而获得价值函数或者策略函数。

第四篇（第9~12章）

此部分主要围绕深度强化学习展开，该技术通过结合深度学习的表征能力和强化学习的决策

能力，使得智能体具备了更好的学习能力，能够解决更为复杂的感知决策问题。

第9章首先概述深度学习中较为经典的3种网络结构模型：深度神经网络、卷积神经网络和循环神经网络。随后介绍深度强化学习相关概念，并对深度强化学习当前具有代表性的应用进行简单介绍。第10章介绍了第一个深度强化学习算法：DQN算法。该方法通过结合Q-learning算法、经验回放机制以及卷积神经网络生成目标Q值等技术，有效地解决了深度学习和强化学习融合过程中所面临的问题和挑战，实现了深度学习与强化学习的深层次融合。第11章介绍了DQN算法所存在的不足，以及后续研究者所提出的具有代表性的深度强化学习算法：DDPG算法、A3C算法、Rainbow算法和Ape-X算法。第12章全面而细致地介绍了AlphaGo程序的设计思想与原理，并给出了AlphaGo和AlphaGo Zero程序的算法细节。

本书的最后提供了附录A~附录E，内容涵盖深度学习方面相关函数、算法及技巧，供读者学习使用。

建议和反馈

为了帮助广大读者更好地理解和使用书中的案例代码，本书的实验代码托管在GitHub上，地址为：https://github.com/chenzomi12/Deep-Reinforcement-Learning。

写一本书是一项极其琐碎、繁重的工作，尽管两位作者已经竭力使本书趋于完美，但仍然可能存在很多漏洞和瑕疵。如果读者对本书有任何评论和建议，可提交到异步社区中，我们将不胜感激。

致谢

编写一本书不是一件一蹴而就的事情。两位作者在工作之余为此付出了巨大的努力；感谢张爽编辑的支持和配合；同时，感谢两位作者的亲人长时间的支持、理解、陪伴和帮助。

感谢本书参考文献中的作者，在深度强化学习领域，正是有了他们的辛勤付出以及对各自子领域的深入研究，才有了本书所介绍到的专业知识。

感谢张景中院士、陈恩红教授、OPPO研究院院长刘畅为本书作序。他们无论在学术界还是工业界都有着深厚的积累，更是机器证明、距离几何及动力系统、机器学习与数据挖掘方面的领军人物。同时，感谢他们为计算机和数学领域所做的巨大贡献，也感谢他们对两位作者的信任和支持。

感谢购买本书的读者对两位作者专业水平的认可。我们希望本书能够为读者提供专业的指导，并希望本书介绍的强化学习技术能够对你们的实际工作有所帮助。

陈仲铭
何　明
2019年1月于上海