



# Reinforcement Learning: An Introduction

强化学习导论第二版翻译

作者：Sutton & Barto

组织：UESTC

时间：March 2, 2020

译者：吕昀璿



*Victory won't come to us unless we go to it. — M. Moore*

# 目录

第二版前言	iv
第一版前言	viii
符号总结	x
翻译说明	xiv
<b>1 介绍</b>	<b>1</b>
1.1 强化学习 . . . . .	1
1.2 例子 . . . . .	3
1.3 强化学习的要素 . . . . .	4
1.4 局限和范围 . . . . .	5
1.5 拓展例子：井字游戏 . . . . .	6
1.6 总结 . . . . .	10
1.7 强化学习的早期历史 . . . . .	10
 <b>第一部分 表格解决方法</b>	 <b>18</b>
<b>2 多臂赌博机</b>	<b>19</b>
2.1 $k$ 臂赌博机问题 . . . . .	19
2.2 动作值方法 . . . . .	20
2.3 10 臂试验 . . . . .	21
2.4 渐增实现 . . . . .	23
2.5 非平稳问题 . . . . .	25
2.6 乐观初始值 . . . . .	26
2.7 置信上限动作选择 . . . . .	27
2.8 梯度赌博机算法 . . . . .	28
2.9 关联搜索（上下文赌博机） . . . . .	32
2.10 总结 . . . . .	32
2.11 书目与历史评论 . . . . .	34
 <b>3 有限马尔可夫决策过程</b>	 <b>36</b>

<b>4</b>	<b>Elegant<math>\text{\LaTeX}</math> 系列模板介绍</b>	<b>38</b>
4.1	ElegantBook 更新说明	38
4.2	模板安装与更新	39
4.2.1	在线使用模板	39
4.2.2	本地免安装使用	39
4.2.3	发行版安装使用	39
4.2.4	更新问题	39
4.2.5	其他发行版本	40
4.3	用户作品计划	40
4.4	关于提交	41
4.5	协作人员招募	41
4.6	致谢	41
4.7	捐赠	41
<b>5</b>	<b>ElegantBook 设置说明</b>	<b>43</b>
5.1	语言模式	43
5.2	设备选项	43
5.3	颜色主题	43
5.4	封面	44
5.4.1	封面个性化	44
5.4.2	封面图	45
5.4.3	徽标	45
5.4.4	自定义封面	45
5.5	章标标题	45
5.6	数学环境简介	45
5.6.1	定理类环境的使用	46
5.6.2	其他环境的使用	46
5.7	装饰物	47
5.8	列表环境	47
5.9	参考文献	47
5.10	添加序章	48
5.11	目录选项与深度	48
5.12	章节摘要	48
5.13	章后习题	49
	第五章习题	49
5.14	旁注	49



<b>6</b>	<b>字体选项</b>	<b>51</b>
6.1	数学字体选项 . . . . .	51
6.2	使用 newtx 系列字体 . . . . .	51
6.2.1	连字符 . . . . .	51
6.2.2	宏包冲突 . . . . .	51
6.3	中文字体选项 . . . . .	52
6.3.1	方正字体选项 . . . . .	52
6.3.2	其他中文字体 . . . . .	52
<b>7</b>	<b>ElegantBook 写作示例</b>	<b>54</b>
7.1	Lebesgue 积分 . . . . .	54
7.1.1	积分的定义 . . . . .	54
	第七章习题 . . . . .	56
<b>8</b>	<b>常见问题集</b>	<b>57</b>
<b>9</b>	<b>版本更新历史</b>	<b>59</b>
<b>A</b>	<b>基本数学工具</b>	<b>61</b>
A.1	求和算子与描述统计量 . . . . .	61



## 第二版前言

自本书第一版出版以来的 20 年里，人工智能取得了巨大的进步，这在很大程度上是由机器学习的进步推动的，包括强化学习的进步。虽然令人印象深刻的可用的计算能力是一部分这些进步中的原因，但理论和算法方面的新发展也一直是推动力量。面对这一进展，我们 1998 年出版的书籍的第二版早就应该出版了，我们终于在 2012 年开始了这个项目。我们第二版的目标与第一版的目标相同：提供清晰而简单的强化学习的关键思想和算法的描述，所有相关学科的读者都可以接触到这些关键思想和算法。这个版本仍然是一个介绍性的版本，我们将重点放在核心的在线学习算法上。这一版包括了一些新的话题，这些话题在过去的几年里变得越来越重要，我们扩大了对现在更好理解的话题的覆盖范围。但我们没有试图提供该领域的全面覆盖，它已经在许多不同的方向爆发。很抱歉，我们不得不删除了少数几个贡献之外的所有内容。

与第一版一样，我们选择不强化学习进行严格的正式处理，或者用最一般的术语来表示它。然而，自从第一版以来，我们对一些主题的深入理解需要更多的数学知识来解释；我们在阴影框内设置了更多数学部分，不喜欢数学的人可以选择跳过。我们还使用了与第一版略有不同的符号。在教学中，我们发现新的记法有助于解决一些常见的混乱之处。它强调随机变量（用大写字母表示）和它们的实例（用小写表示）之间的差异。例如，时间步  $t$  处的状态、动作和奖励分别表示为  $S_t$ 、 $A_t$  和  $R_t$ ，而它们可能的值可以表示为  $s$ 、 $a$  和  $r$ 。与此同时，对于值函数（例如  $v_\pi$ ）使用小写字母并将大写字母限制为它们的表格估计值（例如  $Q_t(s, a)$ ）是很自然的。近似值函数是随机参数的确定性函数，因此也是小写形式（例如  $\hat{v}(s, \mathbf{w}_t) \approx v_\pi(s)$ ）。诸如权重向量  $\mathbf{w}_t$ （以前记为  $\boldsymbol{\theta}_t$ ）和特征向量  $\mathbf{x}_t$ （以前记为  $\boldsymbol{\phi}_t$ ）等向量即使是随机变量，也会以粗体并以小写形式书写。矩阵保留使用大写粗体。在第一版中，我们使用了特殊的符号  $\mathcal{P}_{s,s'}^a$  和  $R_{s,s'}^a$  来表示转移概率和期望奖励。这种符号的一个缺点是，它仍然没有完全描述奖励的动态性，只给出他们的期望，这对于动态规划来说是足够的，但是对于强化学习来说却不是。另一个缺点是下标和上标过多。在这个版本中，我们使用显式符号  $p(s', r | s, a)$  来表示给定当前状态和动作的下一个状态和奖励的联合概率。第 x 页的表格总结了符号的所有变化。

第二版有了很大的扩充，顶层组织也发生了变化。在第一章绪论之后，第二版分为三个新的部分。第一部分（第 2-8 章）尽可能多地讨论强化学习，但不超出可以找到精确解的表格情况。我们涵盖了表格情况的学习和规划方法，以及它们在  $n$  步方法和 Dyna 中的统一。这一部分中介绍的许多算法对第二版来说都是新的，包括 UCB、期望 Sarsa、双重学习、树备份、 $Q(\sigma)$ 、RTDP 和 MCTS。首先彻底地完成表格情况，可以在最简单的背景中开发出核心思想。然后，本书的第二部分（第 9-13 章）致力于将思想扩展到函数逼近。它在人工神经网络，傅立叶基础，LSTD，基于核的方法，Gradient-TD 和 Emphatic-TD 方法，平均奖励方法，真实在线 TD( $\lambda$ ) 和策略梯度方法方面有新的章节。第二版大大扩展了对离线策略学习的处理方式，首先是在第 5-7 章中讨论表格情况，然后在第 11 章和



第 12 章中讨论函数逼近。另一个变化是第二版将  $n$  步自举的前向观点（现在在第 7 章中更全面地讨论）与资格迹的后向观点（现在在第 12 章中独立讨论）分开。这本书的第三部分有大量关于强化学习与心理学（第 14 章）和神经科学（第 15 章）的关系的新章节，以及一个更新的案例研究章节，包括 Atari 游戏，Watson 的下注策略，以及围棋程序 AlphaGo 和 AlphaGo Zero（第 16 章）。尽管如此，出于必要我们只包括了该领域已完成的所有工作中的一小部分。我们的选择反映了我们对廉价的免模型方法的长期兴趣，这些方法应该能够很好地扩展到大型应用中。最后一章现在包括关于强化学习的未来社会影响的讨论。不管好坏，第二版篇幅大约是第一版的两倍。

这本书被设计用来作为一个或两个学期的强化学习课程的主要文本。对于一个学期的课程，前十章应该按顺序覆盖并形成一个好的核心，根据口味，可以从其他章节、其他书籍（如 Bertsekas 和 Tsitsiklis（1996）、Wiering 和 van Otterlo（2012）和 Szepesvari（2010））或文献中添加材料。根据学生的背景，一些关于在线监督学习的额外材料可能会有所帮助。Option 和 option 模型的思想是很自然的补充（Sutton, Precup 和 Singh, 1999）。一门为期两学期的课程可以涵盖所有章节以及补充材料。这本书也可以作为机器学习、人工智能或神经网络等更广泛课程的一部分。在这种情况下，可能只希望涵盖材料的一部分。建议您简要介绍第 1 章，第 2 章至第 2.4 节，第 3 章，然后根据时间和兴趣从其余各章中选择各节。第 6 章对于本主题和本书的其余部分来说是最重要的。侧重于机器学习或神经网络的课程应涵盖第 9 章和第 10 章，而侧重于人工智能或规划的课程则应涵盖第 8 章。在整本书中，对于本书其余部分而言那些较为困难和不重要的部分和章节都标有 \* 号。这些可以在第一次阅读时省略，而不会在以后造成问题。有些练习也标有 \* 号，表示它们更高级，对理解本章的基本材料不是必需的。

大多数章节的结尾都有一个标题为“书目和历史评论”的章节，在这一章节中，我们对这一章中提出的观点的来源给予了信任，为进一步阅读和正在进行的研究提供了指导，并描述了相关的历史背景。尽管我们试图使这些部分具有权威性和完整性，但我们无疑遗漏了一些重要的前期工作。为此，我们再次表示歉意，我们欢迎将更正和扩展纳入该书的电子版。

像第一版一样，这本书的这个版本是为了纪念 A. Harry Klopf。是 Harry 把我们介绍给彼此的，也是他关于大脑和人工智能的想法让我们开始了长时间的强化学习之旅。Harry 受过神经生理学训练，长期对机器智能感兴趣，是俄亥俄州莱特帕特森空军基地空军科学研究中心（AFOSR）航空电子董事会的一名资深科学家。他不满意平衡寻求过程，包括稳态和纠错模式分类方法在解释自然智能和为机器智能提供基础方面的重要性。他指出，试图最大化某些东西（无论是什么）的系统与寻求均衡的系统在性质上是不同的，他认为最大化系统是理解自然智能的重要方面和构建人工智能的关键。Harry 在从 AFOSR 获得资助的一个项目中发挥了重要作用，该项目旨在评估这些想法和相关想法的科学价值。该项目于 20 世纪 70 年代末在马萨诸塞大学阿默斯特分校（UMass Amherst）进行，最初的指导是 Michael Arbib、William Kilmer 和 Nico Spinelli，他们是阿默斯特大学计算机和信息科学系的教授，也是该大学系统神经科学控制论中心的创始成员，这是一个专注于神经科学和人工智能交叉的有远见的组织。最近从密歇根大学获得博士学位的 Barto

被聘为该项目的博士后研究员。与此同时，在斯坦福大学攻读计算机科学和心理学的本科生萨顿 Sutton 一直在与 Harry 通信，讨论他们对刺激计时在经典条件反射中的作用的共同兴趣。Harry 向马萨诸塞州大学的小组建议，Sutton 将对这个项目起到很好的补充作用。就这样，Sutton 成为了麻省大学的研究生，他的博士学位是由 Barto 指导的，Barto 已经成为一名副教授。本书中提出的强化学习研究理所当然地是该项目的成果，该项目由 Harry 发起，并受到他的想法的启发。此外，Harry 负责使我们，作者们之间进行了长期而愉快的互动。通过将本书献给 Harry，我们不仅在强化学习领域，而且在我们的合作中都对他的重要贡献表示敬意。我们也感谢 Arbib, Kilmer 和 Spinelli 教授为我们提供了开始探索这些想法的机会。最后，我们感谢 AFOSR 在我们的研究初期给予的大力支持，并感谢 NSF 在接下来的许多年中给予的大力支持。

我们要感谢很多人对第二版的启发和帮助。我们感谢每一个为第一版提供灵感和帮助的人，也应该为这一版得到我们最深切的感谢，如果没有他们对第一版的贡献，这一版就不会存在。在这份长长的名单中，我们必须加上许多其他专门为第二版做出贡献的人。我们的学生在我们教授这些材料的许多年里在很多方面做出了贡献：揭露错误，修正错误，以及尤其是在感到困惑的地方我们可以更好地解释这些事情。我们特别感谢 Martha Steenstrup 通篇阅读并提供详细的评论。如果没有许多心理学和神经科学领域专家的帮助，心理学和神经科学的章节是不可能写出来的。我们感谢 John Moore 多年来对动物学习实验、理论和神经科学的耐心指导，感谢他仔细阅读了第 14 章和第 15 章的多个草稿。我们也感谢 Matt Botvinick、Nathaniel Daw、Peter Dayan 和 Yael Niv 对这些章节草稿的深刻评论，他们在大量文献中的重要指导，以及他们拦截了我们早期草稿中的许多错误。当然，这些章节中的其余错误，而且肯定还有一些，完全是我们自己的错误。我们感谢 Phil Thomas 帮助我们让非心理学家和非神经学家能够接触到这些章节，我们感谢 Peter Sterling 帮助我们改进了论述。我们感谢 Jim Houk 向我们介绍基底神经节的信息处理这一主题，并提醒我们注意神经科学的其他相关方面。Jose Martinez、Terry Sejnowski、David Silver、Gerry Tesauro、Georgios Theodoridis 和 Phil Thomas 慷慨地帮助我们了解他们的强化学习应用的详细信息，并将其纳入案例研究一章，他们对这些部分的草稿提供了有用的意见。特别感谢 David Silver 帮助我们更好地理解 Monte Carlo Tree search 和 DeepMind 围棋程序。我们感谢 George Konidaris 在傅立叶基础上的章节帮助。Emilio Cartoni、Thomas Cederborg、Stefan Dornbach、Clemens Rosenbaum、Patrick Taylor、Thomas Colin 和 Pierre-Luc Bacon 在一些重要的方面帮助了我们，对此我们非常感激。

Sutton 还想感谢艾伯塔大学强化学习和人工智能实验室的成员为第二版做出的贡献。他欠 Rupam Mahmood 一笔特别的债，因为他在第 5 章中对 off-policy Monte Carlo 方法的处理做出了重要贡献。他特别感谢 Rupam Mahmood 在第 5 章中对 off-policy 蒙特卡罗方法的处理所做的重要贡献，感谢 Hamid Maei 在第 11 章中提出的 off-policy 学习观点，感谢 Eric Graves 在第 13 章中进行的实验，感谢张尚同 (Shangtong Zhang) 对几乎所有实验结果的复制和验证，感谢 Kris De Asis 改进了第 7 章和第 12 章的新技术内容，感谢 Harm van Seijen 的深刻见解，这导致了  $n$  步方法与资格迹的分离，以及（与 Hadovan Hasselt 一起）第 12 章中提出的资格迹的前向和后向视图的精确等价的思想。Sutton 还感谢艾伯

塔省政府和加拿大国家科学与工程研究委员会在第二版构思和撰写期间给予他的支持和自由。他特别要感谢 **Randy Goebe** 为艾伯塔省的研究创造了一个支持性和有远见的环境。他还想感谢 **DeepMind** 在写这本书的最后六个月里对他的支持。

最后，我们要感谢在互联网上发布的第二版草稿的许多细心的读者。他们发现了我们遗漏的许多错误，并提醒了我们一些可能的困惑之处。





# 第一版前言

---

我们最初在 1979 年末开始关注于现在所谓的强化学习。我们俩都在马萨诸塞大学工作，致力于最早的项目之一，以复兴像适应性元素这样的神经网络可能被证明是有前途的人工适应性智能的方法的想法。该项目探索了由 A. Harry Klopff 开发的“自适应系统的静态理论”。Harry 的工作是丰富的思想源泉，我们被允许对其进行批判性的探索，并将其与自适应系统中以前的悠久历史进行比较。我们的任务变成了将这些想法分开并理解它们之间的关系和相对重要性。这种情况一直持续到今天，但是在 1979 年，我们开始意识到，也许很久以来就被认为是最简单的想法，从计算的角度来看却很少受到关注。这仅仅是一个学习系统的思想，它需要某种东西，它可以调整其行为，以便最大化来自其环境的特殊信号。这就是“享乐主义”学习系统的思想，或者正如我们现在所说的，强化学习的思想。

和其他人一样，我们感觉强化学习在控制论和人工智能的早期就已经进行了彻底的探索。然而，经过仔细检查，我们发现它只被略微探索过。尽管强化学习显然推动了一些最早的学习计算研究，但这些研究人员中的大多数已经转向其他方面，如模式分类、监督学习和自适应控制，或者他们完全放弃了学习研究。结果，涉及学习如何从环境中获取东西的特殊问题受到相对较少的关注。回过头来看，关注这个思想是启动这一研究分支的关键一步。强化学习的计算研究进展甚微，直到人们意识到这一基本思想还没有被彻底探索。

从那时起，该领域已经走了很长一段路，并在多个方向上发展和成熟。强化学习已逐渐成为机器学习，人工智能和神经网络研究中最活跃的研究领域之一。该领域已经发展出坚实的数学基础和令人印象深刻的应用。强化学习的计算研究现已成为一个广阔的领域，在世界各地有数百名活跃的研究人员，他们来自不同的学科，如心理学、控制理论、人工智能和神经科学。特别重要的是建立和发展了与最优控制和动态规划理论的关系。从互动中学习以实现目标的整体问题仍远未解决，但我们对它的理解已经有了显著提高。我们现在可以将组成部分的思想，如时间差分学习、动态规划和函数逼近，放在一个关于整体问题的连贯视角中。

我们写这本书的目的是对强化学习的关键思想和算法提供一个清晰而简单的描述。我们想让所有相关学科的读者都能了解我们的论述方法，但我们不能详细介绍所有这些观点。在很大程度上，我们的论述采取了人工智能和工程学的观点。覆盖与其他领域的联系，我们将其留给其他人或其他时间。我们还选择不对强化学习进行严格的形式化处理，我们没有达到数学抽象的最高水平，也没有依赖于定理证明格式。我们试图选择一定程度的数学细节，以使数学朝正确的方向倾斜，而又不会分散基础思想的简单性和潜在的普遍性。

从某种意义上说，我们致力于这本书已有三十年，我们有很多人要感谢。首先，我们感谢那些亲自帮助我们发展本书中提出的总体观点的人：Harry Klopff 帮助我们认识到

需要复兴强化学习；Chris Watkins, Dimitri Bertsekas, John Tsitsiklis 和 Paul Werbos 帮助我们看到关系对动态规划的价值；John Moore 和 Jim Kehoe 来自动物学习理论的见解和启发；Oliver Selfridge 强调适应的广度和重要性；更一般地说，我们的同事和学生做出了无数的贡献：Ron Williams, Charles Anderson, Satinder Singh, Sridhar Mahadevan, Steve Bradtke, Bob Crites, Peter Dayan 和 Leemon Baird。通过与 Paul Cohen, Paul Utgoff, Martha Steenstrup, Gerry Tesauro, Mike Jordan, Leslie Kaelbling, Andrew Moore, Chris Atkeson, Tom Mitchell, Nils Nilsson, Stuart Russell, Tom Dietterich, Tom Dean 的讨论，我们的强化学习观得到了极大的丰富。我们感谢 Michael Littman, Gerry Tesauro, Bob Crites, Satinder Singh 和 Wei Zhang 分别提供了第 4.7、15.1、15.4、15.4 和 15.6 节的详细信息。我们感谢空军科学研究所，美国国家科学基金会和 GTE 实验室的长期和有远见的支持。

我们还要感谢阅读本书草稿并提供宝贵意见的许多人，包括 Tom Kalt, John Tsitsiklis, Pawel Cichosz, Olle G. Allmo, Chuck Anderson, Stuart Russell, Ben Van Roy, Paul Steenstrup, Paul Cohen, Sridhar Mahadevan, Jette Randlov, Brian Sheppard, Thomas O'Connell, Richard Coggins, Cristina Versino, John H. Hiatt, Andreas Badelt, Jay Ponte, Joe Beck, Justus Piater, Martha Steenstrup, Satinder Singh, Tommi Jaakkola, Dimitri Bertsekas, Torbjörn Ekman, Christina Björkman, Jakob Carlström 和 Olle Palmgren。最后，我们感谢 Gwyn Mitchell 在许多方面提供的帮助，并感谢 Harry Stanton 和 Bob Prior 成为麻省理工学院出版社的冠军。

## 符号总结

大写字母用于随机变量，而小写字母用于随机变量的值和标量函数。要求是实值向量的量以粗体和大小写形式书写（即使是随机变量）。矩阵是粗体大写字母。

$\doteq$ : 定义为真的等价关系

$\approx$ : 近似相等

$\propto$ : 成正比

$\Pr\{X = x\}$ : 随机变量  $X$  取值  $x$  的概率

$X \sim p$ : 随机变量  $X$  选自分布  $p(x) \doteq \Pr\{X = x\}$

$\mathbb{E}[X]$ : 随机变量  $X$  的期望，即  $\mathbb{E}[X] \doteq \sum_x p(x)x$

$\arg \max_a f(a)$ :  $f(a)$  取最大值时的  $a$  值

$\ln x$ :  $x$  的自然对数

$e^x, \exp(x)$ : 自然对数的底取  $x$  次方， $e \approx 2.71828$ ， $e^{\ln x} = x$

$\mathbb{R}$ : 实数集

$f: \mathcal{X} \rightarrow \mathcal{Y}$ : 从集合  $\mathcal{X}$  的元素到集合  $\mathcal{Y}$  的元素的函数  $f$

$\leftarrow$ : 赋值

$(a, b]$ :  $a$  和  $b$  之间的实际间隔，包括  $b$  但不包括  $a$

$\varepsilon$ : 在  $\varepsilon$  贪婪策略中采取随机动作的概率

$\alpha, \beta$ : 步长参数

$\gamma$ : 折扣率参数

$\lambda$ : 资格迹的衰减率参数

$\mathbb{I}_{predicate}$ : 指示函数（如果  $predicate$  为真，则  $predicate \doteq 1$ ，否则为 0）

多臂赌博机问题:

$k$ : 动作 (臂) 数

$t$ : 离散时间步或者玩耍次数

$q_*(a)$ : 动作  $a$  的真实值 (期望奖励)

$Q_t(a)$ :  $q_*(a)$  在时间  $t$  处的估计

$N_t(a)$ : 在时间  $t$  之前选择动作  $a$  的次数

$H_t(a)$ : 在时间  $t$  选择动作  $a$  的习得偏好

$\pi_t(a)$ : 在时间  $t$  选择动作  $a$  的概率

$\bar{R}_t$ : 给定  $\pi_t$ ，在时间  $t$  期望奖励的估计

马尔可夫决策过程:

$s, s'$ : 状态

$a$ : 动作

$r$ : 奖励

$\mathcal{S}$ : 所有非终结状态的集合

$\mathcal{S}^+$ : 所有状态的集合, 包括终结状态

$\mathcal{A}(s)$ : 状态  $s$  中所有可用动作的集合

$\mathcal{R}$ : 所有可能的奖励的集合,  $\mathbb{R}$  的有限子集

$\subset$ : 子集, 例如,  $\mathcal{R} \subset \mathbb{R}$

$\in$ : 元素, 例如,  $s \in \mathcal{S}$ ,  $r \in \mathcal{R}$

$|\mathcal{S}|$ : 集合  $\mathcal{S}$  中的元素数

$t$ : 离散时间步

$T, T(t)$ : 一个回合的最后时间步, 或包括时间步  $t$  的回合的最后时间步

$A_t$ :  $t$  时刻的行动

$S_t$ :  $t$  时刻的状态, 通常是随机的, 归因于  $S_{t-1}$  和  $A - t - 1$

$R_t$ :  $t$  时刻的奖励, 通常是随机的, 归因于  $S_{t-1}$  和  $A - t - 1$

$\pi$ : 策略 (决策规则)

$\pi(s)$ : 在确定性策略  $\pi$  下在状态  $s$  中采取的动作

$\pi(a|s)$ : 在随机策略  $\pi$  下在状态  $s$  采取动作  $a$  的概率

$G_t$ : 时间  $t$  后的回报

$h$ : 水平线, 向前看的时间步长

$G_{t:t+n}, G_{t:h}$ : 从  $t+1$  到  $t+n$  或到  $h$  的  $n$  步回报 (折扣和修正)

$\bar{G}_{t:h}$ :  $t+1$  至  $h$  的平滑回报 (未折扣和未修正) (第 5.8 节)

$G_t^\lambda$ :  $\lambda$  回报 (第 12.1 节)

$G_{t:h}^\lambda$ : 截断, 修正的  $\lambda$  回报 (第 12.3 节)

$G_t^{\lambda s}, G_t^{\lambda a}$ : 按估计状态或作用, 值修正的  $\lambda$  回报 (第 12.8 节)

$p(s', r|s, a)$ : 从状态  $s$  和动作  $a$  转移到具有奖励  $r$  的状态  $s'$  的概率

$p(s'|s, a)$ : 从状态  $s$  采取动作  $a$  转换到状态  $s'$  的概率

$r(s, a)$ : 采取动作  $a$  后状态  $s$  期望立即回报

$r(s, a, s')$ : 在动作  $a$  下从  $s$  转移到  $s'$  的期望立即回报

$v_\pi(s)$ : 策略  $\pi$  下状态  $s$  的值 (期望回报)

$v_*(s)$ : 最优策略  $\pi$  下状态  $s$  的值

$q_\pi(s, a)$ : 策略  $\pi$  下状态  $s$  中采取动作  $a$  的值

$q_*(s, a)$ : 最优策略  $\pi$  下状态  $s$  中采取动作  $a$  的值

$V, V_t$ : 状态值函数  $v_\pi$  或  $v_*$  的数组估计

$Q, Q_t$ : 动作值函数  $q_\pi$  或  $q_*$  的数组估计

$\bar{V}_t(s)$ : 近似动作值的期望, 比如,  $\bar{V}_t(s) \doteq \sum_a \pi(a|s) Q_t(s, a)$

$U_t$ : 在时间  $t$  估计的目标

$\delta_t$ :  $t$  (随机变量) 处的时间差分 (TD) 误差 (第 6.1 节)

$\delta_t^s, \delta_t^a$ : TD 误差的状态和动作特定形式 (第 12.9 节)

$n$ : 在  $n$  步方法中,  $n$  是自举的步数

- $d$ : 维度—— $\mathbf{w}$  的组成数量  
 $d'$ : 替代维度—— $\boldsymbol{\theta}$  的组成数量  
 $\mathbf{w}, \mathbf{w}_t$ : 近似值函数下的权重的  $d$  维向量  
 $w_i, w_{t,i}$ : 习得的权重向量的第  $i$  个分量  
 $\hat{v}(s, \mathbf{w})$ : 给定权重向量  $\mathbf{w}$  的状态  $s$  的近似值  
 $v_{\mathbf{w}}(s)$ :  $\hat{v}(s, \mathbf{w})$  的替代符号  
 $\hat{q}(s, a, \mathbf{w})$ : 给定权重向量  $\mathbf{w}$  的状态动作对  $s, a$  的近似值  
 $\nabla \hat{v}(s, \mathbf{w})$ :  $\hat{v}(s, \mathbf{w})$  相对于  $\mathbf{w}$  的偏导数的列向量  
 $\nabla \hat{q}(s, a, \mathbf{w})$ :  $\hat{q}(s, a, \mathbf{w})$  相对于  $\mathbf{w}$  的偏导数的列向量  
 $\mathbf{x}(s)$ : 处于状态  $s$  时可见的特征向量  
 $\mathbf{x}(s, a)$ : 在状态  $s$  采取动作  $a$  时可见的特征向量  
 $x_i(s), x_i(s, a)$ : 向量  $\mathbf{x}(s)$  或  $\mathbf{x}(s, a)$  的第  $i$  个分量  
 $\mathbf{x}_t$ :  $\mathbf{x}(S_t)$  或  $\mathbf{x}(S_t, A_t)$  的简写  
 $\mathbf{w}^\top \mathbf{x}$ : 向量的内积,  $\mathbf{w}^\top \mathbf{x} \doteq \sum_i w_i x_i$ , 例如,  $\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^\top \mathbf{x}(s)$   
 $\mathbf{v}, \mathbf{v}_t$ : 用于学习  $\mathbf{w}$  的权重的次  $d$  维向量 (第 11 章)  
 $\mathbf{z}_t$ : 时间  $t$  时资格迹的  $d$  维向量 (第 12 章)  
 $\boldsymbol{\theta}, \boldsymbol{\theta}_t$ : 目标策略的参数向量 (第 13 章)  
 $\pi(a|s, \boldsymbol{\theta})$ : 给定参数向量  $\boldsymbol{\theta}$ , 在状态  $s$  下采取动作  $a$  的概率  
 $\pi_{\boldsymbol{\theta}}$ : 对应于参数  $\boldsymbol{\theta}$  的策略  
 $\nabla \pi(a|s, \boldsymbol{\theta})$ :  $\pi(a|s, \boldsymbol{\theta})$  相对于  $\boldsymbol{\theta}$  的偏导数的列向量  
 $J(\boldsymbol{\theta})$ : 策略  $\pi_{\boldsymbol{\theta}}$  的性能度量  
 $\nabla J(\boldsymbol{\theta})$ :  $J(\boldsymbol{\theta})$  相对于  $\boldsymbol{\theta}$  的偏导数的列向量  
 $h(s, a, \boldsymbol{\theta})$ : 基于  $\boldsymbol{\theta}$  选择状态  $s$  中的动作  $a$  的偏好  
 $b(a|s)$ : 学习目标策略  $\pi$  时用于选择动作的行为策略  
 $b(s)$ : 策略梯度方法的基线函数  $b: \mathcal{S} \mapsto \mathbb{R}$   
 $b$ : MDP 或搜索树的分支因子  
 $\rho_{t:h}$ : 从时间  $t$  到时间  $h$  的重要性采样率 (第 5.5 节)  
 $\rho_t$ : 仅时间  $t$  的重要性采样率,  $\rho_{t:t}$   
 $r(\pi)$ : 策略  $\pi$  的平均奖励 (奖励率) (第 10.3 节)  
 $\bar{R}_t$ :  $t$  时刻  $r(\pi)$  的估计  
 $\mu(s)$ : 状态  $s$  上的在线策略分布 (第 9.2 节)  
 $\boldsymbol{\mu}$ : 所有  $s \in \mathcal{S}$  的  $\mu(s)$  的  $|\mathcal{S}|$  维向量  
 $\|v\|_{\boldsymbol{\mu}}^2$ : 值函数  $v$  的  $\boldsymbol{\mu}$  加权平方范数, 即  $\|v\|_{\boldsymbol{\mu}}^2 \doteq \sum_{s \in \mathcal{S}} \mu(s) v(s)^2$   
 $\eta(s)$ : 每回合访问状态  $s$  的期望次数 (第 199 页)  
 $\Pi$ : 值函数的投影算子 (第 268 页)  
 $B_{\pi}$ : 值函数的贝尔曼算子 (第 11.4 节)  
 $\mathbf{A}$ :  $d \times d$  矩阵  $\mathbf{A} \doteq \mathbb{E}[\mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^\top]$   
 $\mathbf{b}$ :  $d$  维向量  $\mathbf{b} \doteq \mathbb{E}[R_{t+1} \mathbf{x}_t]$



$\mathbf{w}_{\text{TD}}$ : TD 不动点  $\mathbf{w}_{\text{TD}} \doteq \mathbf{A}^{-1}\mathbf{b}$  (一个  $d$  维向量, 第 9.4 节)

$\mathbf{I}$ : 单位矩阵

$\mathbf{P}$ :  $\pi$  下的状态转移概率的  $|\mathcal{S}| \times |\mathcal{S}|$  维矩阵

$\mathbf{D}$ : 对角线上有  $\mu$  的  $|\mathcal{S}| \times |\mathcal{S}|$  维对角矩阵

$\mathbf{X}$ : 以  $x(s)$  为行的  $|\mathcal{S}| \times |d|$  维矩阵

$\bar{\delta}_{\mathbf{w}}(s)$ : 状态  $s$  的  $v_{\mathbf{w}}$  的贝尔曼误差 (期望 TD 误差) (第 11.4 节)

$\bar{\delta}_{\mathbf{w}}, \text{BE}$ : 贝尔曼误差向量, 分量为  $\bar{\delta}_{\mathbf{w}}(s)$

$\overline{\text{VE}}(\mathbf{w})$ : 均方值误差  $\overline{\text{VE}}(\mathbf{w}) \doteq \|v_{\mathbf{w}} - v_{\pi}\|_{\mu}^2$

$\overline{\text{BE}}(\mathbf{w})$ : 均方贝尔曼误差  $\overline{\text{BE}}(\mathbf{w}) \doteq \|\bar{\delta}_{\mathbf{w}}\|_{\mu}^2$

$\overline{\text{PBE}}(\mathbf{w})$ : 均方投影贝尔曼误差  $\overline{\text{PBE}}(\mathbf{w}) \doteq \|\Pi \bar{\delta}_{\mathbf{w}}\|_{\mu}^2$

$\overline{\text{TDE}}(\mathbf{w})$ : 均方时间差分误差  $\overline{\text{TDE}}(\mathbf{w}) \doteq \mathbb{E}_b[\rho_t \delta_t^2]$  (第 11.5 节)

$\overline{\text{RE}}(\mathbf{w})$ : 均方回报误差 (第 11.6 节)

## 翻译说明

---

强化学习领域目前国外研究比较多，所以我们常常需要看英文资料才能了解到最新的前沿动态。为了兼顾英文资料而不引起混淆，本人不对强化学习中的非熟知人名、特殊词汇以及强化学习中的方法和理论词汇进行翻译。

由于英文和中文各方面的差异，每个人有不同的翻译习惯，一个单词或者词语可以被翻译成各个形式。因此，很有必要列出翻译过程中遇到的特殊单词或词语，这种良好的方式可以帮助读者看其他文献或者书籍时迅速进行转换。下面列出本人在翻译过程中意译的特殊单词或词汇：

action: 动作  
associations: 关联  
agent: 译成 agent, 可译成智能体, 代理  
artificial neural network: 人工神经网络  
actor-critic: 译成 actor-critic, 可译成演员-评论家  
action-value: 动作值  
Alan Turing: 阿兰·图灵 (人名)  
arrays: 数组  
active exploration: 主动探索  
approximation: 近似, 逼近  
behavior/behaving/behavioral/behave: 行为, 表现  
backgammon: 双陆棋  
Bellman: 贝尔曼 (人名)  
bellman equation: 贝尔曼方程  
backwards: 向后  
bootstrap/bootstrapping: 自举  
blackjack: 21 点  
backpropagation neural networks: 反向神经网络  
backward-view : 后向观点  
bias: 有偏  
baseline: 基线, 基准线  
curse of dimensionality: 维度灾难  
credit: 信用  
control: 控制  
controller: 控制器  
context: 上下文  
conditioned reflexes: 条件反射

Claude Shannon: 克劳德·香农  
critic: 评论家  
dynamical: 动态  
dynamic programming: 动态规划  
dopamine: 多巴胺  
Double learning: 双重学习  
exploration/explore/exploring: 探索  
exploitation/exploit/exploiting: 利用  
environment: 环境  
events: 事件  
evolutionary: 进化  
episodes: 回合  
 $\varepsilon$ -greedy:  $\varepsilon$ -greedy  
exponential recency-weighted average: 指数近期加权平均  
evaluates: 评价, 评估  
function: 函数  
forward: 向前  
forward-view: 前向观点  
goal: 目标  
goal-directed: 目标导向  
genetic algorithms: 遗传算法  
genetic programming: 遗传规划  
generalize: 泛化  
game theory: 博弈论  
greedy: 贪婪  
hierarchical: 层级  
interactive: 互动, 交互  
immediate: 即时  
k-armed bandit: k 臂赌博机  
label: 标签  
look ahead: 预见  
Law of Effect: 效果定律  
Least-Mean-Square: 最小均方  
map: 映射  
markov decision processe: 马尔可夫决策过程  
model: 模型  
model-based: 基于模型  
model-free: 免模型, 无模型



Monte Carlo: 蒙特卡罗  
Monte Carlo Tree search: 蒙特卡罗树搜索  
memory: 内存  
organisms: 有机体  
optimal: 最优  
optimal control: 最优控制  
offline: 离线  
online: 在线  
one-armed bandit: 单臂赌博机  
option: 译成 option  
planning: 规划  
psychology: 心理学  
policy: 策略  
perceived: 感知  
prior knowledge: 先验知识  
player: 玩家  
Pavlov: 巴甫洛夫 (人名)  
pseudocode: 伪代码  
Q-learning: 译成 Q-learning, 可译成 Q-学习  
reinforcement learning: 强化学习  
rules: 规则  
reward: 奖励  
return: 回报  
state: 状态  
supervised learning: 监督学习  
search: 搜索  
signal: 信号  
simulated annealing: 模拟退火  
static: 静态  
space: 空间  
step: 步  
step-size: 步长  
slot machine: 老虎机  
trial-and-error: 试错  
tic-tac-toe: 井字游戏  
temporal-difference: 时间差分  
tables/tabular: 表格  
time step: 时间步



Thompson: 汤普森

unsupervised learning: 非监督学习

uncertainty: 不确定性

value function: 值函数

value: 值, 价值





# 第一章 介绍

---

当我们思考学习的本质时，我们通过与环境互动来学习的想法很可能首先出现在我们的脑海中。当婴儿玩耍、挥舞手臂或环顾四周时，他没有明确的老师，但他确实与周围的环境有直接的感觉运动联系。运用这种联系会产生大量关于因果关系，动作后果以及为了实现目标应该做些什么等方面的信息。在我们的一生中，这样的互动无疑是关于我们的环境和我们自身的主要知识来源。无论我们是在学习开车还是在进行对话，我们都敏锐地意识到我们的环境对我们所做的事情如何反应，我们试图通过我们的行为来影响所发生的事情。从互动中学习是几乎所有学习和智能理论的基本理念。

在这本书中，我们探索了一种从互动中学习的计算方法。我们不是直接对人或动物如何学习进行理论推导，而是主要探索理想化的学习情境，并评估各种学习方法的有效性<sup>1</sup>。也就是说，我们采用人工智能研究人员或工程师的观点。我们探索能有效解决科学或经济利益的学习问题的机器设计，通过数学分析或计算实验来评估设计。我们探索的方法被称为强化学习，与其他机器学习方法相比，它更注重从互动中进行目标导向学习。

## 1.1 强化学习

强化学习是学习做什么——如何将情境映射到动作——以便最大化数字奖励信号。学习者不会被告知要采取哪些动作，而是必须通过尝试发现哪些动作产生的奖励最大。在最有趣和最具挑战性的情况下，动作不仅会影响到眼前的奖励，还会影响到下一情境，并通过这种情况影响到所有随后的奖励。试错搜索和延迟奖励这两个特征是强化学习的两个最重要的区别性特征。

强化学习，就像许多以“ing”结尾的主题一样，比如机器学习和登山，同时也是一个问题，一类在这个问题上很有效的解决方法，以及研究这个问题及其解决方法的领域。对这三件事使用一个名字是很方便的，但同时保持三者概念上的分离也很重要。特别是，问题和解决方法的区分在强化学习中非常重要；未能做出这种区分是许多困惑的根源。

我们利用动态系统理论的思想将强化学习问题形式化，特别是作为不完全已知的马尔可夫决策过程的最优控制。这种形式化的细节必须等到第3章，但基本思想只是捕捉学习 agent 随着时间的推移与环境互动以实现目标所面临的实际问题的最重要方面。学习 agent 必须能够在一定程度上感知其环境的状态，并且必须能够采取影响该状态的动作。agent 还必须有一个或多个与环境状态相关的目标。马尔可夫决策过程的目的是只包括这三个方面——感觉、行动和目标——以其最简单的可能形式，而不会使它们中的任何一个变得琐碎。任何适合于解决这类问题的方法都被认为是强化学习方法。

强化学习不同于监督学习，监督学习是目前机器学习领域研究最多的一种学习方式。监督学习是从知识渊博的外部监督者提供的带标签示例的训练集中学习。每个示例都是

---

<sup>1</sup>第14章和第15章总结了与心理学和神经科学的关系。

对情况的描述以及系统应针对该情况采取的正确动作的规范（标签），该规范通常用于标识该情况所属的类别。这种学习的目的是使系统推断或概括其对策，以使其在训练集中不存在的情况下正确行动。这是一种重要的学习方式，但单靠它并不足以从互动中学习。在互动问题中，要获得既正确又代表 **agent** 必须采取行动的所有情况的期望行为的示例通常是不切实际的。在未知领域——人们期望学习是最有益的——**agent** 必须能够从自己的经验中学习。

强化学习也不同于机器学习研究人员所说的无监督学习，无监督学习通常是关于发现隐藏在未标记数据集中的结构。监督学习和无监督学习这两个术语似乎对机器学习范式进行了详尽的分类，但事实并非如此。虽然有人可能会认为强化学习是一种无监督的学习，因为它不依赖于正确行为的示例，但强化学习试图最大化奖励信号，而不是试图发现隐藏的结构。揭示 **agent** 经验中的结构在强化学习中当然是有用的，但本身并不能解决最大化奖励信号的强化学习问题，因此我们认为强化学习是第三种机器学习范式，与监督学习、无监督学习以及其他范式并驾齐驱。

强化学习而不是其他类型学习中会出现的挑战之一是探索与利用之间的权衡。为了获得更多的奖励，强化学习 **agent** 必须偏爱过去尝试过并且发现能有效产生奖励的行为。但要发现这样的动作，它必须尝试以前没有选择过的动作。**Agent** 必须利用它已经经历过的来获得奖励，但是它也必须探索以便在将来做出更好的行动选择。进退两难的境地是，无论是探索还是利用，都不可能在不失败的情况下专心致志地进行。**Agent** 必须尝试各种方法，并逐渐偏爱那些看起来最好的方法。对于随机任务，每个动作都必须多次尝试才能获得期望回报的可靠估计。探索-利用问题是数学家们几十年来一直在深入研究的问题，但至今仍未得到解决。目前，我们只是注意到，平衡探索和利用的整个问题甚至不会出现在监督和无监督的学习中，至少在这些范式的最纯粹形式中是不会出现的。

强化学习的另一个关键特征是它明确地考虑了目标导向的 **agent** 与不确定环境交互的整个问题。这与许多考虑子问题的方法形成鲜明对比，这些方法没有考虑它们可能如何适应更大的图景。例如，我们已经提到，许多机器学习研究都与监督学习有关，但没有明确说明这种能力最终将如何发挥作用。其他研究人员发展了有总体目标的规划理论，但没有考虑规划在实时决策中的作用，也没有考虑规划所需的预测模型从何而来的问题。虽然这些方法已经产生了许多有用的结果，但它们对孤立的子问题的关注是一个很大的限制。

强化学习采取相反的策略，从一个完整的、互动的、寻求目标的 **agent** 开始。所有的强化学习 **agent** 都有明确的目标，能够感知环境的各个方面，并能够选择动作来影响环境。此外，通常从一开始就假定 **agent** 必须运行，尽管它所面临的环境有很大的不确定性。当强化学习涉及规划时，它必须解决规划和实时动作选择之间的相互作用，以及如何获取和改进环境模型的问题；当强化学习涉及监督学习时，它这样做是出于特定的原因，这些原因决定了哪些能力是关键的，哪些能力不是关键的。为了使学习研究取得进展，重要的子问题必须被分离和研究，但它们应该是在完整的、互动的、寻求目标的 **agent** 中发挥明确作用的子问题，即使完整 **agent** 的所有细节还不能被填充。

我们所说的完整的、互动的、寻找目标的 **agent** 并不总是指像完整的有机体或机器人

这样的东西。这些都是很明显的例子，但是一个完整的、互动的、寻求目标的 agent 也可以是一个更大的行为系统的组成部分。在这种情况下，agent 直接与较大系统的其余部分交互，并间接与较大系统的环境交互。一个简单的例子是一个 agent，它监控机器人电池的电量，并向机器人的控制架构发送命令。这个 agent 理的环境是机器人的其余部分以及机器人的环境。人们必须超越最明显的 agent 及其环境的例子来理解强化学习框架的普遍性。

现代强化学习最令人兴奋的方面之一是它与其他工程和科学学科的大量和富有成效的互动。强化学习是人工智能和机器学习数十年来与统计学、最优化和其他数学学科更紧密结合的趋势的一部分。例如，一些强化学习方法使用参数化逼近器学习的能力解决了运筹学和控制理论中经典的“维数灾难”。更明显的是，强化学习还与心理学和神经科学产生了强烈的互动，具有双向的实质性好处。在机器学习的所有形式中，强化学习是最接近于人类和其他动物所做的那种学习，强化学习的许多核心算法最初都是受到生物学习系统的启发。强化学习也带来了回报，既通过动物学习的心理学模型更好地匹配了一些经验数据，也通过大脑奖励系统的部分有影响力的模型。本书的主体提出了与工程和人工智能有关的强化学习的思想，并在第 14 章和第 15 章中概述了与心理学和神经科学的联系。

最后，强化学习也是人工智能向简单通用原则回归的更大趋势的一部分。自 20 世纪 60 年代末以来，许多人工智能研究人员假设没有可以发现的一般原理，相反，智能是由于拥有大量特殊目的的技巧、程序和启发式方法。人们有时会说，如果我们能把足够多的相关事实输入机器，比方说一百万或十亿，那么它就会变得智能。基于一般原则的方法，如搜索或学习，被描述为“弱方法”，而那些基于特定知识的方法被称为“强方法”。这种观点在今天仍然很普遍，但并不占主导地位。从我们的观点来看，现在就下结论还为时过早：在寻找一般原则方面投入的努力太少，无法得出没有任何一般原则的结论。现代人工智能现在包括许多研究寻找学习、搜索和决策的一般原则。目前还不清楚钟摆会往回摆动多久，但强化学习研究肯定是朝着更简单、更少的人工智能一般原理摇摆的一部分。

## 1.2 例子

理解强化学习的一个好方法是考虑一些指导它发展的例子和可能的应用。

- 一个象棋大师出棋。这种选择是由规划——预期可能的回复和反回复——以及对特定位置和移动的期望的即时、直觉判断所决定。
- 自适应控制器实时调整炼油厂的操作参数。控制器根据指定的边际成本优化产量/成本/质量权衡，而不严格遵守工程师最初建议的设定点。
- 一只小瞪羚在出生几分钟后挣扎着站起来。半小时后，它以每小时 20 英里的速度运行。
- 移动机器人决定它是否应该进入一个新房间来收集更多的垃圾，还是开始试图找到回到电池充电站的路。它根据电池当前的充电水平以及过去找到充电器的速度和容

易程度做出决定。

- **Phil** 在准备他的早餐。仔细观察，即使是这种看似平凡的活动也揭示了一个复杂的条件行为和连锁的目标-子目标关系的网络：走到橱柜前，打开它，选择一个麦片盒，然后伸手去拿，抓住，再拿出盒子。要获得碗、勺子和牛奶盒，还需要其他复杂的、可调的、互动的行为序列。每一步都包括一系列的眼球运动，以获取信息并指导到达和运动。人们会不断地快速判断如何搬运这些物品，或者在拿到其他物品之前先把其中一些摆放到餐桌上是否更好。每一步都有目标的引导，比如抓住勺子或拿到冰箱，并服务于其他目标，比如一旦谷类食品准备好，就用勺子吃，并且最终获得营养。无论他是否意识到这一点，**Phil** 都在获取关于他身体状况的信息，这些信息决定了他的营养需求、饥饿程度和食物偏好。

这些例子都有一些非常基本的特点，因此很容易被忽略。所有这些都涉及到一个积极的决策 **agent** 和其环境之间的互动，在这种互动中，**agent** 寻求实现一个目标，尽管其环境是不确定的。**agent** 的动作被允许影响环境的未来状态（例如，下一个象棋位置、精炼厂的蓄水池水位、机器人的下一个位置和其电池的未来充电水平），从而影响 **agent** 在以后时间可用的动作和机会。正确的选择需要考虑到动作的间接、延迟后果，因此可能需要远见或规划。

与此同时，在所有这些例子中，动作的效果无法完全预测；因此，**agent** 必须频繁地监视其环境并做出适当的反应。例如，**Phil** 必须注意他倒进麦片碗里的牛奶，以防止牛奶溢出。所有这些例子都涉及明确的目标，即 **agent** 可以根据它能直接感觉到的东西来判断朝着目标的进展。棋手知道自己是否赢了，炼油厂的控制器知道正在生产多少石油，小羚羊知道它何时倒下，移动机器人知道它的电池什么时候耗尽，**Phil** 知道他是否正在享受他的早餐。

在所有这些例子中，随着时间的推移 **agent** 可以利用其经验来提高其性能。棋手提炼他用来评估位置的直觉，从而提高他的棋艺；瞪羚幼仔提高了它奔跑的能力；**Phil** 学会简化早餐的制作。**Agent** 从一开始就带给任务的知识（无论是从以前的相关任务经验还是通过设计或演化而来）会影响有用的或易于学习的内容，但是与环境的交互对于调整行为以利用任务的特定特性是至关重要的。

## 1.3 强化学习的要素

除了 **agent** 和环境之外，还可以确定强化学习系统的四个主要子要素：策略、奖励信号、值函数，以及可选的环境模型。

策略定义学习 **agent** 在给定时间的行为方式。粗略地说，策略是从感知的环境状态到处于这些状态时要采取的动作的映射。它对应于心理学上所说的一套刺激——反应规则或关联。在某些情况下，策略可以是简单的函数或查找表，而在其他情况下，它可能涉及诸如搜索过程之类的大量计算。策略是强化学习 **agent** 的核心，因为只有它才能决定行为。一般来说，策略可能是随机的，为每个动作指定概率。

奖励信号定义了强化学习问题的目标。在每个时间步，环境都会向强化学习 **agent** 发



送一个称为奖励的单个数值。**Agent** 的唯一目标是使其长期获得的总奖励最大化。因此，奖励信号定义了对 **agent** 来说什么是好的和不好的事件。在生物系统中，我们可能认为奖励是快乐或痛苦经历的衍生物。它们是 **agent** 面临的问题的直接和决定性特征。奖励信号是改变策略的主要依据；如果策略选择的动作之后是低奖励，那么策略可能会改变，以选择未来在这种情况下其他动作。一般来说，奖励信号可能是环境状态和所采取动作的随机函数。

奖励信号表明什么是即时好的，而值函数则指明什么是长期好的。粗略地说，状态的值是 **agent** 从该状态开始，在未来期望积累的奖励总额。奖励决定了环境状态的即时、内在的可取性，而价值则表明了考虑到可能跟随的状态和那些状态可获得的奖励之后，状态的长期可取性。例如，一个状态可能总是产生较低的即时奖励，但仍然有很高的价值，因为其他产生高奖励的状态经常紧随其后。反之亦然。打个比方，奖励有点像快乐（如果高的话）和痛苦（如果低的话），而价值观对应于一种更精致和更有远见的判断，即我们对环境处于特定状态的满意或不满程度。

奖励在某种意义上是首要的，而价值作为对奖励的预测是次要的。没有奖励就没有价值，估计价值的唯一目的就是获得更多的奖励。然而，我们在制定和评估决策时最关心的是价值。动作选择是基于价值判断做出的。我们寻求能带来最高价值而非最高奖励的动作，因为从长远来看，这些动作为我们获得了最大量的奖励。不幸的是，确定价值要比确定奖励困难得多。奖励基本上是由环境直接给出的，但价值必须根据 **agent** 在其整个生命周期内所做的观察序列来估计和重新估计。实际上，我们考虑的几乎所有强化学习算法中最重要的组成部分是一种科学估算价值的方法。价值评估的中心作用可以说是过去六十年来有关强化学习的最重要的知识。

一些强化学习系统的第四个也是最后一个要素是环境模型。这是一种模仿环境行为的方法，或者更笼统地说，它允许对环境将如何表现做出推断。例如，给定状态和动作，模型可以预测所得的下一个状态和下一个奖励。模型是用于规划，我们所说的规划是指在实际经历之前，通过考虑未来可能发生的情况来决定动作方针的任何方式。使用模型和规划来解决强化学习问题的方法被称为基于模型的方法，这与简单的免模型方法相反，后者明确地说是试错学习者的方法，被视为几乎与规划对立。在第 8 章中，我们探索了强化学习系统，它通过试错，学习环境的模型，并使用该模型进行规划，同时学习。现代强化学习涵盖了从低级的试错学习到高级的深思熟虑的规划。

## 1.4 局限和范围

强化学习在很大程度上依赖于状态的概念，其作为策略和值函数的输入，以及作为模型的输入和输出。非正式地，我们可以认为状态是向 **agent** 传达特定时间“环境如何”的某种意义的信号。我们在这里使用的状态的正式定义由第 3 章中给出的马尔可夫决策过程框架给出。然而，更一般地，我们鼓励读者遵循非正式的含义，并将状态视为 **agent** 可以获得的关于其环境的任何信息。实际上，我们假设状态信号是由一些预处理系统产生的，这些预处理系统名义上是 **agent** 环境的一部分。我们不会在本书中讨论构造、改变或



学习状态信号的问题 (除了第 17.3 节中的简要介绍)。我们采取这种方法不是因为我们认为状态表示不重要，而是为了充分关注决策问题。换句话说，我们在这本书中关注的不是设计状态信号，而是根据任何可用的状态信号来决定采取什么动作。

我们在这本书中考虑的大多数强化学习方法都是围绕估计值函数来构建的，但并不一定需要这样做来解决强化学习问题。例如，诸如遗传算法、遗传规划、模拟退火和其他优化方法的求解方法从不估计值函数。这些方法应用多个静态策略，每个静态策略都在较长时间内与单独的环境实例交互。获得最多奖励的策略及其随机变化被传递到下一代策略，并重复这一过程。我们称这些方法为进化方法，因为它们的操作类似于生物进化产生具有熟练行为的有机体的方式，即使它们在自己的一生中没有学习。如果策略空间非常小，或者可以被构造使得好的策略很常见或者很容易找到，或者如果有很多时间用于搜索，那么进化方法可能是有效的。此外，进化方法在学习 agent 不能感知其环境的完整状态的问题上具有优势。

我们的重点是强化学习方法，这种方法在与环境互动的同时进行学习，而进化方法则不然。在许多情况下，能够利用个体行为互动细节的方法比进化方法更有效。进化方法忽略了强化学习问题的许多有用结构：它们没有利用这样一个事实，即它们正在寻找的策略是从状态到动作的函数；他们没有注意到个体在一生中经历了哪些状态，或者选择了哪些动作。在某些情况下，这种信息可能会产生误导 (例如，当状态被错误感知时)，但更常见的情况是，它应该能够实现更有效的搜索。虽然进化和学习有许多共同的特点，并且可以自然地协同工作，但是我们并不认为进化方法本身特别适合强化学习问题，因此，我们在本书中不涉及它们。

## 1.5 拓展例子：井字游戏

为了说明强化学习的一般思想，并将其与其他方法进行比较，我们接下来将更详细地考虑一个例子。

想一想我们熟悉的孩子玩的井字游戏。两个玩家轮流在一块三乘三的棋盘上玩。一个玩家玩 X，另一个玩家玩 O，直到一个玩家通过水平、垂直或对角连续放置三个标记排成一行来获胜，就像 X 玩家在右边的游戏中所做的那样。如果棋盘填满了，没有一个玩家连续三个一行，那么这场比赛就是一局平局。因为一个技术娴熟的玩家可以玩得永远不会输，让我们假设我们的对手是一个不完美的玩家，他的

X	O	O
O	X	X
		X

玩法有时是不正确的，可以让我们获胜。就目前而言，事实上，让我们考虑平局和失败对我们同样不利。我们如何才能打造出一个能够发现对手游戏中的不完美之处，并学会最大化取胜机会的玩家呢？

虽然这是一个简单的问题，但无法通过经典的技术以令人满意的方式轻松解决。例如，博弈论中经典的“极小极大”解决方案在这里是不正确的，因为它假设对手玩的是一种特殊的方式。例如，极小极大玩家永远不会达到它可能会输的游戏状态，即使实际上它总是因为对手的不正确游戏而从这种状态中获胜。序列决策问题的经典优化方法，如

动态规划，可以为任何对手的计算出最优解，但需要输入对手的完整说明，包括对手在每个棋盘状态下的每一步棋的概率。让我们假设这个问题没有先验的信息，因为它不适用于绝大多数实际感兴趣的问题。另一方面，这样的信息可以从经验中估计出来，在这种情况下，可以通过与对手进行多次游戏来估计。在这个问题上，我们能做的最好的事情是首先学习对手的行为模型，达到一定的置信度，然后应用动态规划来计算给定近似对手模型的最优解。最后，这与我们在本书后面讨论的一些强化学习方法没有什么不同。

应用于这个问题的进化方法将直接搜索可能的策略空间，以寻找赢得对手的可能性很高的策略。这里，策略是一种规则，它告诉玩家在游戏的每一种状态要走什么棋，状态是指三乘三的棋盘上 X 和 O 的每一种可能的配置。对于所考虑的每种策略，其获胜概率的估计将通过与对手进行一定数量的博弈来获得。然后，该评估将指示下一步考虑哪种或者哪些策略。一种典型的进化方法是在策略空间中攀爬，连续生成和评估策略，试图获得渐进的改进。或者，也许可以使用遗传方式的算法来维护和评估一组政策。实际上可以应用数百种不同的优化方法。

下面是如何利用值函数来处理井字游戏问题的方法。首先，我们会建立一个数字表格，一个代表游戏的每个可能状态。每个数字都是我们在该状态获胜的可能性的最新估计。我们把这个估计值当作状态的值，整个表就是学习值函数。状态 A 的值比状态 B 高，或者被认为比状态 B “更好”，如果我们当前估计的获胜概率比状态 B 高。假设我们总是玩 X，那么对于所有有三个 X 的状态来说，获胜的概率是 1，因为我们已经赢了。类似地，对于连续有三个 O 的所有状态，或者已经填满的状态，正确的概率都是 0，因为我们不能从他们那里获胜。我们将所有其他状态的初始值设置为 0.5，表示我们有 50% 的获胜机会。

然后我们和对手玩很多场游戏。为了选择我们的移动，我们检查每个可能的移动产生的状态（棋盘的每个空格对应一个状态），并在表中查找它们的当前值。大多数情况下，我们都在贪婪地移动，选择通向价值最大的状态的移动，也就是说，往具有最高的估计获胜概率移动。然而，有时我们会从其他动作中随机选择。这些被称为探索性的移动，因为它们使我们体验到我们本来可能永远看不到的状态。游戏过程中做出和考虑的一系列动作如图 1.1 所示。

当我们在玩的时候，我们会改变我们在游戏中发现自己所处的状态的值。我们试图让他们更准确地估计获胜的可能性。为此，我们在每次贪婪地移动之后将状态值“备份”到移动之前的状态，如图 1.1 中的箭头所示。更准确地说，较早状态的当前值被更新为更接近较晚状态的值，这可以通过将较早状态的值向较晚状态的值移动一小部分来实现。如果我们让  $S_t$  表示贪婪移动之前的状态，让  $S_{t+1}$  表示移动后的状态，则对  $S_t$  的估计值（表示为  $V(S_t)$ ）的更新可以写为

$$V(S_t) \leftarrow V(S_t) + \alpha[V(S_{t+1}) - V(S_t)],$$

其中  $\alpha$  是一个小的正分数，称为步长参数，它影响学习速率。此更新规则是时间差分学习方法的一个例子，之所以这样命名是因为它的更改是基于两个连续时间的差异  $V(S_{t+1}) - V(S_t)$ 。

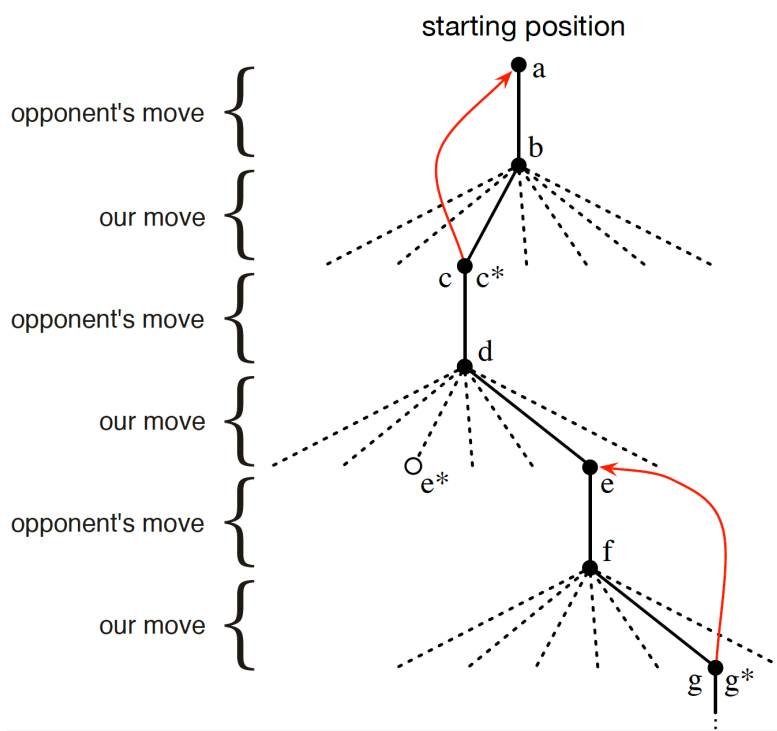


图 1.1: 井字游戏的一系列移动。实心的黑线表示在游戏中采取的移动；虚线表示我们（我们的强化学习玩家）考虑过但没有采取的移动。我们的第二个移动是探索性的移动，这意味着即使另一个兄弟移动，即导致  $e^*$  的那个，排名更高，我们还是采取了这个移动。探索性移动不会导致任何学习，但我们的每一次其他的移动可以，如红色箭头所示它们都会导致更新，其中估计值从树的后面节点移动到前面的节点，如正文中所详细描述的那样。

上面描述的方法在这项任务上执行得相当好。例如，如果步长参数随着时间的推移适当减小，那么对于任何固定的对手，这个方法都会收敛到我们的玩家在给定最优玩法的情况下从每个状态取胜的真实概率。此外，随后采取的移动（除探索移动外）实际上是针对这个（不完美的）对手的最优移动。换句话说，该方法收敛于与该对手进行博弈的最优策略。如果步长参数没有随着时间的推移一直减少到零，那么这个玩家也能很好地对付那些慢慢改变他们游戏方式的对手。

这个例子说明了进化方法和学习值函数的方法之间的区别。为了评估策略，进化方法保持策略不变，并与对手玩许多游戏，或者使用对手的模型来模拟许多游戏。获胜的频率给出了该策略获胜概率的无偏估计，并可用于指导下一个策略选择。但是每一个策略的改变都是在许多游戏之后才做出的，并且只使用每一个游戏的最终结果：游戏期间发生的事情被忽略。例如，如果玩家赢了，那么它在游戏中的所有行为都被给予信用，与对于获胜至关重要的特定移动无关。甚至对从未发生的移动也给予了肯定！相反，值函数方法允许评估各个状态。最后，进化方法和值函数方法都搜索策略空间，但是学习值函数会利用游戏过程中可用的信息。

这个简单的例子说明了强化学习方法的一些关键特性。首先，强调在与环境互动的同时学习，在这种情况下是与对手玩家互动。其次，有一个明确的目标，考虑到选择的延迟效应，正确的行为需要规划或远见。例如，简单的强化学习玩家将学会为短见的对手设置多步移动陷阱。强化学习解决方案的一个显著特点是不需要使用对手的模型，也不需要可能的未来状态和动作序列进行显式搜索，就可以实现规划和前瞻的效果。

虽然这个例子说明了强化学习的一些关键特性，但它非常简单，可能会给人一种强化学习比实际情况更有限的印象。虽然井字棋是一种双人游戏，但强化学习也适用于没有外部对手的情况，即“与自然对抗的游戏”。强化学习也不局限于行为被分成不同回合的问题，就像井字游戏那样，只有在每一回合结束时才能获得奖励。当行为无限期地持续和任何时候都能收到各种数量的奖励时，它也同样适用。强化学习也适用于那些甚至不能分解成离散时间步长的问题，比如井字游戏。一般原理也适用于连续时间问题，尽管理论变得更加复杂，我们在这个介绍性的讨论中省略了它。


井字游戏具有相对较小的有限状态集，然而强化学习可以在状态集非常大甚至无限时使用。例如，Gerry Tesauro (1992, 1995) 将上述算法与人工神经网络相结合，学习下具有大约 1020 个状态的双陆棋。拥有如此众多的状态，不可能只经历其中的一小部分。Tesauro 的程序学会了比以前的任何程序玩得更好，最终也比世界上最好的人类玩家打得更好（第 16.1 节）。人工神经网络为程序提供了从其经验中进行泛化的能力，因此在新的状态中，它根据其网络所确定的从过去面临的相似状态中保存的信息来选择移动。强化学习系统在具有如此大的状态集的问题中能够工作得多好，与它从过去的经验中进行概括的恰当程度密切相关。正是在这一角色中，我们最需要强化学习的监督学习方法。人工神经网络和深度学习（第 9.6 节）不是唯一的，也不一定是最好的方法。

在这个井字游戏的例子中，学习从没有超出游戏规则的先验知识开始，但强化学习绝不意味着对学习和智力的全盘看法。相反，先验信息可以以各种方式结合到强化学习中，这些方式对于特定的学习至关重要（例如参见第 9.5、17.4 和 13.1 节）。在井字游戏的例子中，我们也可以进入真实状态，而强化学习也可以应用于部分状态被隐藏的情况，或者不同的状态对学习者是相同的情况。

最后，井字游戏玩家能够预见，并知道它的每一步可能会导致的状态。要做到这一点，它必须有一个游戏模型，让它能够预见环境将如何变化，以应对它可能永远不会采取的移动。许多问题都是这样的，但在其他问题中，甚至缺乏动作效果的短期模型。强化学习可以在任何一种情况下应用。模型不是必需的，但是如果模型是可用的或可以学习，那么可以很容易地使用它们（第 8 章）。

另一方面，有一些强化学习方法根本不需要任何环境模型。免模型系统甚至无法考虑它们的环境将如何改变以响应单个动作。在这个意义上，井字游戏玩家对于它的对手来说是免模型的：它没有任何类型的手对手模型。因为模型必须相当精确才能有用，所以当解决问题的真正瓶颈是建立一个精确的环境模型时，免模型方法比更复杂的方法更有优势。免模型方法也是基于模型方法的重要构件。在这本书里，我们在讨论如何将免模型方法作为更复杂的基于模型的方法的组成部分之前，专门用了几章来讨论免模型方法。

在一个系统中，强化学习既可以在高级别上使用，也可以在低级别上使用。尽管井字游戏玩家只学到了游戏的基本动作，但没有什么能阻止强化学习在更高的水平上发挥作用，在那里，每个“动作”本身都可能是一种精心设计的解决问题的方法的应用。在层级学习系统中，强化学习可以在多个层次上同时工作。

 **练习 1.1：Self-Play** 假设上面描述的强化学习算法不是与随机对手对战，而是与自身对战，双方都在学习。你认为在这种情况下会发生什么？它会学习一个不同的策略来选择



动作吗? □

🔥 **练习 1.2: Symmetries** 许多井字游戏的位置看起来不同,但由于对称性实际上是一样的。我们如何修改上述学习过程来利用这一点?这种变化会在哪些方面改善学习过程?现在再想想。假设对手没有利用对称性。那样的话,我们应该吗?那么,对称相等的位置必然具有相同的值,这是真的吗? □

🔥 **练习 1.3: Greedy Play** 假设强化学习玩家是贪婪的,也就是说,他总是做出让他达到最佳位置的移动。它会比不贪婪的玩家学得更好或更差吗?可能会发生什么问题? □

🔥 **练习 1.4: Learning from Exploration** 假设学习更新发生在所有移动之后,包括探索移动。如果随时间逐步减小步长参数(而不是探索的趋势),则状态值将收敛到一组不同的概率。当我们从或者不从探索性动作中学习时对应的两组概率是什么(概念上)?假设我们确实在继续进行探索移动,那么哪一组概率可能更好学习?哪个会带来更多胜利? □

🔥 **练习 1.5: Other Improvements** 你还能想出其他方法来提高强化学习玩家吗?你能想出更好的办法来解决所提出的井字游戏问题吗? □

## 1.6 总结

强化学习是一种理解和自动化目标导向学习和决策的计算方法。它与其他计算方法的不同之处在于,它强调 **agent** 通过与其环境的直接交互进行学习,而不需要示例性监督或完整的环境模型。在我们看来,强化学习是第一个认真解决在与环境交互中学习以实现长期目标时出现的计算问题的领域。

强化学习使用马尔可夫决策过程的形式框架来定义学习 **agent** 与其环境之间的状态、动作和奖励方面的交互。该框架旨在成为一种简单的方式来表示人工智能问题的基本特征。这些特征包括因果,不确定性和非决定性,以及明确目标的存在。

值和值函数的概念是我们在本书中考虑的大多数强化学习方法的关键。我们的立场是,值函数对于在策略空间进行有效搜索非常重要。值函数的使用将强化学习方法与进化方法区分开来,进化方法在整个策略评估的指导下直接在策略空间进行搜索。

## 1.7 强化学习的早期历史

强化学习的早期历史有两条主线,既漫长又丰富,在现代强化学习交织在一起之前,这两条主线都是独立追求的。一条主线关注的是在试错中学习,它起源于动物学习的心理学。这条主线贯穿了人工智能的一些最早的工作,并导致了强化学习在 20 世纪 80 年代初的复兴。第二条主线关注的是最优控制问题及其利用值函数和动态规划的解决方案。在大多数情况下,这条主线并不涉及学习。这两条主线大部分是独立的,但在某种程度上,围绕着第三条不太明显的主线变得相互关联,这第三条主线关注的是时间差分方法,比如本章中井字游戏的例子。这三条主线在 20 世纪 80 年代末汇集在一起,产生了我们在本书中介绍的现代强化学习领域。

关注于试错学习的主线是我们最熟悉的,也是我们在这段短暂的历史中最有发言权的。然而,在此之前,我们先简要讨论一下最优控制主线。



术语“最优控制”在 20 世纪 50 年代末开始使用，用来描述设计控制器以最小化或最大化动态系统随时间的行为量度的问题。解决这一问题的方法之一是由理查德·贝尔曼 (Richard Bellman) 等人在 20 世纪 50 年代中期通过推广 19 世纪的哈密顿 (Hamilton) 和雅各比 (Jacobi) 理论而发展起来的。这种方法使用动力系统状态和值函数或“最优回报函数”的概念来定义函数方程，现在通常称为贝尔曼方程。通过求解这个方程来解决最优控制问题的一类方法被称为动态规划 (Bellman, 1957a)。贝尔曼 (1957b) 还引入了最优控制问题的离散随机版本，称为马尔可夫决策过程 (MDPs)。Ronald Howard (1960) 提出了 MDP 的策略迭代方法。所有这些都是现代强化学习理论和算法的基本要素。

动态规划被广泛认为是解决一般随机最优控制问题唯一可行的方法。它受到贝尔曼所谓的“维数灾难”的困扰，这意味着它的计算需求随着状态变量的数量呈指数增长，但它仍然比任何其他通用方法更精确、更广泛地适用。自 20 世纪 50 年代后期以来，动态规划得到了广泛的发展，包括对部分可观察的 MDPs 的扩展 (Lovejoy, 1991 年调查)、许多应用 (White, 1985 年, 1988 年, 1993 年调查)、近似方法 (Rust, 1996 年调查) 和异步方法 (Bertsekas, 1982, 1983)。有许多关于动态规划的出色的现代方法可供使用 (例如 Bertsekas, 2005, 2012; Puterman, 1994; Ross, 1983; 和 Whittle, 1982, 1983)。Bryson (1996) 提供了最优控制的权威历史。

一方面是最优控制和动态规划，另一方面是学习，这两者之间的联系很难识别。我们不能确定是什么导致了这种分离，但它的主要原因可能是所涉及的学科和它们不同的目标之间的分离。另一个贡献可能是流行的观点，即动态规划是一种离线计算，本质上依赖于精确的系统模型和贝尔曼方程的解析解。此外，动态规划最简单的形式是一种在时间上向后进行的计算，这使得很难理解它如何参与必须向前进行的学习过程。动态规划的一些最早的工作，如贝尔曼和 Dreyfus (1959) 的工作，现在可以归类为遵循学习的方法。Witten (1977) 的工作 (下面讨论) 无疑是学习和动态编程思想的结合。Werbos (1987) 明确地论证了动态规划和学习方法之间更大的相互关系，以及动态规划与理解神经和认知机制的相关性。对我们来说，直到 1989 年 Chris Watkins 的工作才将动态规划方法与在线学习完全结合起来，他使用 MDP 形式对强化学习的处理已经被广泛采用。从那时起，这些关系被许多研究人员广泛发展，特别是 Dimitri Bertsekas 和 John Tsitsiklis (1996)，他们创造了术语“神经动态规划”来指动态规划和人工神经网络的结合。目前使用的另一个术语是“近似动态规划”，这些不同的方法强调了该学科的不同方面，但它们都与强化学习有共同的兴趣，即避开动态规划的经典缺点。

我们认为最优控制的所有工作在某种意义上也是强化学习的工作。我们将强化学习方法定义为解决强化学习问题的任何一种有效的方法，现在可以清楚地看到，这些问题与最优控制问题密切相关，特别是随机最优控制问题，如 MDP 形式的随机最优控制问题。因此，我们必须考虑最优控制的求解方法，如动态规划，也是强化学习方法。因为几乎所有的传统方法都需要完整的系统知识才能被控制，所以说它们是强化学习的一部分有点不自然。另一方面，许多动态规划算法是增量和迭代的。就像学习方法一样，它们通过逐次逼近逐渐得到正确的答案。正如我们在这本书的其余部分所展示的那样，这些相似之处远远不是表面上的。完全知识和不完全知识的理论和解决方法是如此紧密地

联系在一起，以至于我们觉得必须把它们作为同一主题的一部分来考虑。

现在让我们回到通向现代强化学习领域的另一条主线，这条主线以试错学习的思想为中心。我们在这里只涉及主要的接触点，在第 14.3 节中会更详细地讨论这个主题。根据美国心理学家 R.S.Woodworth (1938) 的说法，试错学习的概念可以追溯到 19 世纪 50 年代 Alexander Bain 关于通过“摸索和实验”来学习的讨论，更明确地说，是英国行为学家和心理学家 Conway Lloyd Morgan 在 1894 年用这个术语来描述他对动物行为的观察。也许第一个简明扼要地将试错学习的本质表达为学习原则的人是 Edward Thorndike:

在对同一情况做出的几种反应中，伴随或紧随其后的是对动物满意的反应，在其他条件相同的情况下，它们与这种情况的联系更紧密，因此，当它再次出现时，它们更有可能再次出现；那些伴随或紧随其后的是动物的不适，在其他条件相同的情况下，它们与这种情况的联系将被削弱，因此，当它再次发生时，它们发生的可能性就更小。满足感或不适感越大，纽带越强或越弱。(Thorndike, 1911 年，第 244 页)

Thorndike 称之为“效果定律”，因为它描述了强化事件对选择行为的趋势的影响。Thorndike 后来修改了定律，以更好地解释动物学习的后续数据（如奖励和惩罚的效果之间的差异），定律的各种形式在学习理论家中产生了相当大的争议（例如，参见 Gallistel, 2005; Herrnstein, 1970; Kimble, 1961, 1967; Mazur, 1994 年）。尽管如此，效果法则——以一种或另一种形式——被广泛认为是许多行为背后的基本原则（例如，Hilgard and Bower, 1975; Dennett, 1978; Campbell, 1960; Czikó, 1995）。它是 Clark Hull (1943, 1952) 有影响的学习理论和 B. F. Skinner (1938) 有影响的实验方法的基础。

在动物学习的上下文中，术语“强化”是在 Thorndike 表达效果定律之后就开始很好地使用，据我们所知，它最早出现在巴甫洛夫条件反射专著的 1927 年英译本中。巴甫洛夫将强化描述为一种行为模式的增强，这是由于动物接受了刺激——一种强化——与另一种刺激或反应有适当的时间关系。一些心理学家将强化的概念扩展到包括行为的削弱和增强，并将强化的概念扩展到可能包括省略或终止刺激。要被认为是强化，增强或削弱必须在强化退出后持续存在；仅仅吸引动物注意或激励其行为而不产生持久变化的刺激将不被认为是强化。

在计算机中实施试错学习的想法出现在关于人工智能可能性的最早想法中。在 1948 年的一份报告中，阿兰·图灵描述了一种“愉悦-痛苦系统”的设计，它遵循效果定律：

当到达动作未确定的配置时，将对丢失的数据进行随机选择，并在描述中临时应用适当的条目。当疼痛刺激发生时，所有暂定条目都会被取消，当快乐刺激发生时，它们都会被永久化。（图灵，1948 年）

许多巧妙的机电机器被建造出来验证了试错学习。最早的可能是 Thomas Ross (1933) 建造的一台机器，它能够通过一个简单的迷宫找到自己的路，并记住通过开关设置的路径。1951 年，W.Grey Walter 建造了能够进行简单学习的“机械乌龟”(Walter, 1950) 的一个版本。1952 年，克劳德·香农演示了一只名叫 Teseus 的迷宫老鼠通过试错找到迷宫中的

路，迷宫本身通过地板下的磁铁和继电器记住成功的方向（另见香农，1951）。J.A.Deutsch (1954) 基于他的行为理论 (Deutsch, 1953) 描述了一种迷宫求解机器，它与基于模型的强化学习的一些共同性质（第 8 章）。在他的博士学位论文中，Marvin Minsky (1954) 讨论了强化学习的计算模型，并描述了他构建的模拟机器，该机器由他称为 SNARC (Stochastic Neural-Analog Reinforcement Calculators, 随机神经模拟强化计算器) 的组件组成，旨在模拟类似于大脑中可修改的突触连接（第 15 章）。网站 [cyberneticzoo.com](http://cyberneticzoo.com) 包含了大量关于这些和许多其他机电学习机器的信息。

建造机电学习机器让位于对数字计算机编程来完成各种类型的学习，其中一些实现了试错学习。Farley 和 Clark (1954) 描述了一个通过试错学习的神经网络学习机的数字模拟。但是他们的兴趣很快就从试错学习转移到泛化和模式识别，也就是说，从强化学习转移到监督学习 (Clark 和 Farley, 1955)。这开始了一种关于这些学习类型之间关系的混乱模式。许多研究人员似乎认为他们在研究强化学习，但实际上是在研究监督学习。例如，人工神经网络的前驱，如 Rosenblatt (1962) 和 Widrow and Hoff (1960)，显然受到强化学习的激励——他们使用奖励和惩罚的语言——但是他们研究的系统是适用于模式识别和知觉学习的监督学习系统。即使在今天，一些研究者和教科书也淡化或模糊了这些学习类型之间的区别。例如，一些人工神经网络教科书使用“试错”一词来描述从训练样本中学习的网络。这是一种可以理解的混淆，因为这些网络使用错误信息来更新连接权重，但这忽略了试错学习的本质特征，即基于评估性反馈选择动作，而不依赖于正确动作的知识。

部分由于这些困惑，在 20 世纪 60 年代和 70 年代，对真正的试错学习的研究变得稀少，尽管也有明显的例外。在 20 世纪 60 年代，“强化”和“强化学习”这两个术语第一次被用在工程文献中来描述试错学习的工程应用（例如，Waltz 和 Fu, 1965; Mendel, 1966; Fu, 1970; Mendel 和 McClaren, 1970）。特别有影响的是 Minsky 的论文《迈向人工智能的步骤》(Minsky, 1961)，它讨论了几个与试错学习相关的问题，包括预测、期望，以及他所说的复杂强化学习系统的基本信用分配问题：如何在可能涉及到成功的许多决策中分配成功的信用？在某种意义上，我们在这本书中讨论的所有方法都是为了解决这个问题。Minsky 的论文今天很值得阅读。

在接下来的几个段落中，我们讨论其他一些例外情况和部分例外情况，这些例外情况相对忽视了 20 世纪 60 年代和 70 年代对真正的试错学习的计算和理论研究。

一个例外是新西兰研究人员 John Andreae 的工作，他开发了一种名为 STeLLA 的系统，该系统在与环境的互动中通过试错来学习。这个系统包括一个世界的内部模型，以及后来处理隐藏状态问题的“内部独白” (Andreae, 1963, 1969a, b)。Andreae 后来的工作 (1977) 更多地强调向老师学习，但仍然包括在试错中学习，产生新事件是该系统的目标之一。这项工作一个特点是“回漏过程”，在 Andreae (1998) 中进行了更详细的阐述，它实现了一种类似于我们描述的备份更新操作的信用分配机制。不幸的是，他的开创性研究并不广为人知，也没有对后续的强化学习研究产生重大影响。最近的总结可用 (Andreae, 2017a, b)。

更有影响力的是 Donald Michie 的工作。在 1961 年和 1963 年，他描述了一个简单的

试错学习系统，用于学习如何玩井字游戏（或 **naut** 和 **cross**），称为 **MENACE**（for **Matchbox Educable Naughts and Crosses Engine**）。它由每个可能的游戏位置的火柴盒组成，每个火柴盒包含许多彩色珠子，从那个位置的每个可能的移动都有不同的颜色。通过从对应于当前游戏位置的火柴盒中随机抽取一个珠子，可以确定 **MENACE** 的移动。当游戏结束时，珠子会被添加到游戏中使用的盒子中或从盒子中移除，以奖励或惩罚 **MENACE** 的决定。**Michie** 和 **Chambers** (1968) 描述了另一种井字游戏强化学习器，称为 **GLEE**（**Game Learning Expectimaxing Engine**，游戏学习预期引擎）和强化学习控制器，称为 **BOXES**。他们将 **BOXES** 应用于学习平衡铰接在可移动手推车上的杆子的任务，这是基于只有当杆子掉落或手推车到达轨道末端时才会出现的失败信号。这项任务是根据 **Widrow** 和 **Smith** (1964) 的早期工作改编的，他使用监督学习方法，并假设已经能够平衡杆子的老师进行了指导。**Michie** 和 **Chambers** 的杆子平衡版本是在知识不完全的情况下强化学习任务的最好的早期例子之一。它影响了后来的强化学习工作，从我们自己的一些研究开始（**Barto**, **Sutton** 和 **Anderson**, 1983; **Sutton**, 1984）。**Michie** 一直强调试错以及学习是人工智能必不可少的方面（**Michie**, 1974）。

**Widrow**, **Gupta** 和 **Maitra** (1973) 修改了 **Widrow** 和 **Hoff** (1960) 的最小均方 (**LMS**) 算法，以生成强化学习规则，该规则可以从成功和失败信号中学习，而不必从训练样本中学习。他们称这种学习形式为“选择性自举适应”，并将其描述为“与批评家一起学习”，而不是“与老师一起学习”。他们分析了这个规则，并展示了如何学习 21 点。这是 **Widrow** 对强化学习的一次孤立尝试，他对监督学习的贡献更具影响力。我们对“评论家”一词的使用源自 **Widrow**, **Gupta** 和 **Maitra** 的论文。**Buchanan**, **Mitchell**, **Smith** 和 **Johnson** (1978) 在机器学习的上下文中独立使用了“批评家”一词（另请参见 **Dietterich** 和 **Buchanan**, 1984 年），但是对于他们来说，批评家是一个专家系统，能够做的不仅仅是评估性能。

学习自动机的研究对导致现代强化学习研究的试错主线有更直接的影响。这些方法用于解决非关联的纯选择性学习问题，称为 **k** 臂赌博机，类似于老虎机或除了使用 **k** 杠杆外的“单臂赌博机”（参见第 2 章）。学习自动机是简单的，低内存的机器，可以提高在这些问题中获得奖励的概率。学习自动机起源于 20 世纪 60 年代俄国数学家和物理学家 **M.L.Tsetlin** 及其同事（于 1973 年在 **Tsetlin** 去世后出版），此后在工程领域得到了广泛的发展（参见 **Narendra** 和 **Thathachar**, 1974 年, 1989 年）。这些发展包括对随机学习自动机的研究，这是一种基于奖励信号更新动作概率的方法。尽管不是按照随机学习自动机的传统进行发展，但 **Harth** 和 **Tzanakou** (1974) 的 **Alopex** 算法（for **Algorithm of pattern extraction**，用于模式提取算法）是一种随机方法，用于检测动作和强化之间的相关性，这影响了我们的一些早期研究（**Barto**, **Sutton** 和 **Brouwer**, 1981）。随机学习自动机被心理学的早期工作所预示，首先是 **William Estes** (1950) 致力于学习的统计理论，后来又被其他人进一步发展（例如，**Bush** 和 **Mosteller**, 1955; **Sternberg**, 1963）。

在心理学领域发展起来的统计学习理论被经济学研究人员所采用，从而导致了该领域致力于强化学习的研究主线。这项工作始于 1973 年，当时将 **Bush** 和 **Mosteller** 的学习理论应用于一系列经典经济模型中（**Cross**, 1973 年）。这项研究的一个目标是研究比传统理想化的经济 **agent** 更像真实人的人工 **agent**（**Arthur**, 1991）。这种方法扩展到博弈



论背景下的强化学习研究。经济学中的强化学习在很大程度上独立于人工智能中的强化学习的早期工作，尽管博弈论仍然是这两个领域都感兴趣的话题（超出了本书的范围）。Camerer (2011) 讨论了经济学中的强化学习传统，而 Nowe, Vrancx 和 De Hauwere (2012) 从多 agent 扩展的角度对本书中介绍的方法进行了概述。博弈论中的强化与用于玩井字游戏、跳棋和其他休闲游戏的强化学习的主题大不相同。有关强化学习和游戏这方面的概述，请参见，例如，Szita (2012)。

John Holland (1975) 概述了基于选择原理的自适应系统的一般理论。他的早期工作主要涉及非关联形式的试错，例如进化方法和  $k$  臂赌博机。1976 年，更全面的是 1986 年，他引入了分类器系统，这是真正的强化学习系统，包括关联和值函数。Holland 分类器系统的一个关键组成部分是用于信用分配的“bucket-brigade”算法，该算法与我们的井字游戏示例中使用的时间差分算法紧密相关，并在第 6 章中进行了讨论。另一个关键组成部分是遗传算法，一种进化方法，其作用是发展有用的表示形式。分类器系统已被许多研究人员广泛发展，形成了强化学习研究的主要分支 (Urbanowicz 和 Moore, 2009 年进行了综述)，但是遗传算法，我们不认为它们本身就是强化学习系统，受到了更多关注，就像其他用于进化计算的方法一样（例如，Fogel, Owens 和 Walsh, 1966; Koza, 1992）。

在人工智能中，最有责任将试错的主线重新引向强化学习的人是 Harry Klopff (1972, 1975, 1982)。Klopff 认识到，随着学习研究者几乎完全专注于监督学习，适应性行为的基本方面正在消失。Klopff 认为，缺少的是行为的快乐方面，从环境中获得某种结果的动力，控制环境达到期望的目标而远离不期望的目标 (见第 15.9 节)。这是反复试验学习的基本思想。Klopff 的思想对作者们尤其有影响，因为我们对它们的评估 (Barto and Sutton, 1981a) 使我们对监督学习和强化学习之间的区别有所了解，并最终使我们着重于强化学习。我们和同事完成的许多早期工作都是为了表明强化学习和监督学习确实是不同的 (Barto, Sutton 和 Brouwer, 1981; Barto 和 Sutton, 1981b; Barto 和 Anandan, 1985)。其他研究表明强化学习如何解决人工神经网络学习中的重要问题，特别是它如何产生多层网络的学习算法 (Barto, Anderson 和 Sutton, 1982; Barto 和 Anderson, 1985; Barto, 1985, 1986; Barto, 1985, 1986)。Barto 和 Jordan, 1987 年；请参阅第 15.10 节)。

我们现在转到强化学习历史的第三条主线，那就是关于时间差分学习。时间差分学习方法的独特之处在于受同一量的时间连续估计值之间的差异所驱动，例如在井字游戏示例中获胜的概率。这条主线比其他两条主线更小、更不明显，但它在该领域发挥了特别重要的作用，部分原因是时间差分方法似乎是强化学习的新方法和独特的方法。

时间差分学习的起源部分源于动物学习心理学，尤其是在辅助补强的概念中。次要强化是一种与主要强化（例如食物或疼痛）配对的刺激物，因此具有类似的强化特性。Minsky (1954) 可能是第一个意识到这种心理学原理对人工学习系统很重要的人。Arthur Samuel (1959) 是第一个提出并实现一种学习方法的人，该方法包括时间差分思想，这是他著名的跳棋比赛程序的一部分 (第 16.2 节)。

Samuel 没有提及 Minsky 的工作，也没有提及与动物学习的可能联系。他的灵感显然来自克劳德·香农 (1950 年) 的建议，即可以对一台计算机进行编程，使其使用评估函数下棋，并且可以通过在线修改此函数来改善其游戏性。(香农的这些想法可能也影响



了贝尔曼，但我们对此没有任何证据)。Minsky (1961) 在他的“步骤”论文中广泛讨论了 Samuel 的工作，提出了与自然和人为的次要强化理论的联系。

正如我们已经讨论过的，在 Minsky 和 Samuel 工作之后的十年中，关于试错学习的计算工作很少，并且显然没有针对时间差分学习进行任何计算工作。1972 年，Klopf 将试错学习与时间差分学习的重要组成部分结合在一起。Klopf 感兴趣的是可以扩展到大型系统中学习的原理，因此对局部强化的概念很感兴趣，因此整个学习系统的各个子组件可以相互强化。他提出了“广义强化”的概念，其中每个组成部分（名义上是每个神经元）都以强化术语看待其所有输入：兴奋性输入为奖励，抑制性输入为惩罚。这与我们现在称为时间差分学习的想法不同，回想起来，它比 Samuel 的研究还远。另一方面，Klopf 将这一想法与试错学习联系起来，并将其与庞大的动物学习心理学经验数据库联系起来。

Sutton (1978a, b, c) 进一步发展了 Klopf 的思想，特别是与动物学习理论的联系，描述了由时间连续预测的变化驱动的学习规则。他和 Barto 完善了这些观点，并建立了基于时间差分学习的经典条件作用的心理学模型 (Sutton 和 Barto, 1981a; Barto 和 Sutton, 1982)。随后是基于时间差分学习的经典条件作用的其他几种有影响力的心理学模型 (例如 Klopf, 1988; Moore 等, 1986; Sutton and Barto, 1987, 1990)。此时发展的一些神经科学模型在时间差分学习方面得到了很好的解释 (Hawkins 和 Kandel, 1984; Byrne, Gingrich 和 Baxter, 1990; Gelperin, Hopfield 和 Tank, 1985; Tesauro, 1986; Friston 等, 1994 年)，尽管在大多数情况下没有历史联系。

我们关于时间差分学习的早期工作受到动物学习理论和 Klopf 工作的强烈影响。与 Minsky 的“步骤”论文和 Samuel 的跳棋玩家之间的关系直到后来才被认可。但是，到 1981 年，我们已经完全意识到上述所有先前的工作都是时间差分 and 试错主线的一部分。目前，我们开发了一种将时间差分学习与试错学习相结合的方法，称为 actor-critic 体系结构，并将该方法应用于 Michie 和 Chambers 的杠杆平衡问题 (Barto, Sutton 和 Anderson, 1983)。这种方法在 Sutton (1984) 博士学位论文中得到了广泛的研究，并扩展到在 Anderson (1986) 博士学位论文中使用反向传播神经网络。大约在这个时候，Holland (1986) 以他的 bucket-brigade 算法的形式，将时间差分思想明确地纳入了他的分类器系统。Sutton (1988) 采取了关键步骤，将时间差分学习与控制分离开来，将其作为通用的预测方法。该论文还介绍了 TD( $\lambda$ ) 算法并证明了其收敛性。

当我们在 1981 年完成关于 actor-critic 体系结构的工作时，我们发现了 Ian Witten (1977, 1976a) 的一篇论文，该论文似乎是时间差分学习规则的最早出版物。他提出了我们现在称为表格 TD(0) 的方法，该方法用作解决 MDP 的自适应控制器的一部分。这项工作于 1974 年首次提交给期刊出版，也出现在 Witten 的 1976 年博士学位论文中。Witten 的工作是 Andreae 早期使用 STeLLA 和其他试错学习系统进行实验的后继型工作。因此，Witten 于 1977 年发表的论文涵盖了强化学习研究的两个主要主线，即试错学习和最优控制，同时为时间差分学习做出了明显的早期贡献。

时间差分 and 最优控制主线在 1989 年与 Chris Watkins 的 Q-learning 发展完全结合在一起。这项工作扩展并整合了强化学习研究的所有三个主线的先前工作。Paul Werbos (1987) 通过争论试错学习和动态规划的收敛性从 1977 年开始为这种集成做出了贡献。

Watkins 工作之时，强化学习研究已经有了巨大的发展，主要是在人工智能的机器学习子领域，也包括人工神经网络和更广泛的人工智能。1992 年，Gerry Tesauro 的双陆棋游戏程序 TD-Gammon 取得了巨大的成功，给这个领域带来了更多的关注。

自本书第一版出版以来，神经科学的一个蓬勃发展的子领域开始关注强化学习算法与神经系统强化学习之间的关系。造成这种情况的最主要原因是时间差分算法的行为与大脑中产生多巴胺的神经元活动之间的不可思议的相似性，正如许多研究人员所指出的那样（Friston 等，1994；Barto，1995a；Houk，Adams，和 Barto，1995 年；Montague，Dayan 和 Sejnowski，1996 年；Schultz，Dayan 和 Montague，1997 年）。第 15 章介绍了强化学习的这一令人兴奋的方面。在强化学习的近代历史中，其他重要的贡献太多，在这个简短的叙述中无法提及；我们在它们出现的各个章节的末尾引用了其中的许多内容。

## 书目

有关强化学习的其他一般知识，请读者参考 Szepesvari (2010)，Bertsekas 和 Tsitsiklis (1996)，Kaelbling (1993a) 以及 Sugiyama，Hachiya 和 Morimura (2013) 的书。从控制或运筹学的角度来看的书籍包括 Si，Barto，Powell 和 Wunsch (2004)，Powell (2011)，Lewis 和 Liu (2012)，和 Bertsekas (2012)。Cao (2009) 的综述将强化学习置于随机动态系统的学习和优化的其他方法的上下文中。《机器学习》期刊上的三期特刊侧重于强化学习：Sutton (1992a)，Kaelbling (1996) 和 Singh (2002)。Barto (1995b) 提供了有用的综述；Kaelbling，Littman 和 Moore (1996)；以及 Keerthi 和 Ravindran (1997)。Weiring 和 van Otterlo (2012) 编辑的卷很好地概述了最近的发展。

**1.2** 本章中 Phil 早餐的例子是受 Agre (1988) 启发的。

**1.5** 井字游戏示例中使用的时间差分方法在第 6 章进行了阐述。

## 第一部分

# 表格解决方法

在本书的这一部分中，我们以最简单的形式描述了强化学习算法的几乎所有核心思想：其中状态和动作空间足够小，可以将近似值函数表示为数组或表格。在这种情况下，这些方法通常可以找到精确的解决方案，也就是说，它们通常可以精确地找到最优值函数和最优策略。这与本书下一部分中描述的近似方法形成对比，后者只能找到近似解，但反过来可以有效地应用于更大的问题。

本书这一部分的第一章描述了强化学习问题的一种特殊情况解决方法，在这种情况下，只有一种状态，称为赌博机问题。第二章描述了贯穿本书其余部分的一般问题表述——有限马尔可夫决策过程——以及它的主要思想，包括贝尔曼方程和值函数。

接下来的三章描述了用于解决有限马尔可夫决策问题的方法的三个基本类别：动态规划，蒙特卡洛方法和时间差分学习。每类方法都有其优点和缺点。动态规划方法在数学上已经很好地开发，但是需要完整且准确的环境模型。蒙特卡洛方法不需要模型，概念上也很简单，但并不适合逐步进行增量计算。最后，时间差分方法不需要模型，并且是完全增量的，但分析起来更复杂。这些方法在效率和收敛速度方面也有一些不同。

剩下的两章描述了如何将这三类方法结合起来，以获得每种方法的最佳特性。在其一章中，我们描述了如何通过多步自举方法将蒙特卡罗方法的优点与时间差分方法的优点结合起来。在本书这部分的最后一章中，我们将展示时间差分学习方法如何与模型学习和规划方法（如动态规划）相结合，以完整和统一地解决表格强化学习问题。

## 第二章 多臂赌博机

强化学习区别于其他类型的学习的最重要特征是，它使用的训练信息是评价所采取的动作，而不是通过给出正确的动作来进行指导。这就产生了主动探索和明确搜索良好行为的需要。纯粹的评价性反馈表明所采取的动作有多好，但并不能说明它是可能的最好或最差的动作。另一方面，纯粹的指导性反馈指出了要采取的正确动作，而与实际采取的动作无关。这种反馈是监督学习的基础，包括模式分类、人工神经网络和系统识别的大部分内容。以它们的纯粹形式，这两种反馈是截然不同的：评价性反馈完全取决于所采取的动作，而指导性反馈则与所采取的动作无关。

在本章中，我们将在一种简化的设置下研究强化学习的评价方面，这一方面不涉及学习在一种以上情况下的行动。这种非关联的设置是其中大多数涉及评价性反馈的先前工作都已完成的设置，它避免了全部强化学习问题的大部分复杂性。通过研究这种情况，我们可以最清楚地看到评价性反馈与指导性反馈有何不同，并且可以与之结合。

我们探讨的特定的非关联性、评价性反馈问题是  $k$  臂赌博机问题的一个简单版本。我们使用这个问题来介绍许多基本的学习方法，这些方法将在后面的章节中扩展以应用于全部的强化学习问题。在本章的最后，我们向全部的强化学习问题迈进一步，通过讨论当赌博机问题变得关联性时，即在不只一种情况下采取行动时会发生的情况。

### 2.1 $k$ 臂赌博机问题

考虑以下学习问题。您反复面临着  $k$  个不同选项或动作中的一个选择。在每一次选择之后，您会得到一个从平稳概率分布中选出的数值奖励，该奖励取决于您选择的动作。您的目标是在某个时间段内最大化期望的总奖励，例如，超过 1000 个动作选择或时间步。

这是  $k$  臂赌博机问题的原始形式，因此被类比为老虎机或“单臂赌博机”，除了它有  $k$  个杠杆而不是一个。每个动作的选择都像是玩老虎机的一个杠杆的游戏，而奖励就是击中大奖的回报。通过反复的动作选择，您可以将动作集中在最佳杠杆上，从而最大程度地赢取奖金。另一个类比是医生为一系列重病患者选择实验治疗。每个动作都是治疗的选择，每个奖励都是患者的幸存或健康。如今，有时会使用“赌博机问题”术语来概括上述问题，但在本书中，我们仅将其用于指代这种简单案例。

在我们的  $k$  臂赌博机问题中，只要选择了该动作， $k$  个动作中的每个动作都有期望或平均的奖励；让我们称之为动作的价值。我们将在时间步  $t$  上选择的动作表示为  $A_t$ ，并将相应的奖励表示为  $R_t$ 。假定选择了  $a$ ，则任意动作  $a$  的值  $q_*(a)$  是期望的奖励：

$$q_*(a) \doteq \mathbb{E}[R_t | A_t = a].$$

如果您知道每个动作的价值，那么解决  $k$  臂赌博机问题将是微不足道的：您将始终选择

价值最高的动作。我们假设您不确定地知道动作值，尽管您可能有估计。我们将时间步  $t$  处动作  $a$  的估计值表示为  $Q_t(a)$ 。我们希望  $Q_t(a)$  接近  $q_*(a)$ 。

如果您维护动作值的估计，则在任何时间步上至少都有一个动作的估计值最大。我们称这些为贪婪的动作。当您选择这些动作之一时，我们说您正在利用对这些动作的值的现有知识。相反，如果您选择其中一种非贪婪动作，则表示您正在探索，因为这使您可以改善对非贪婪动作价值的估计。利用是在一步之内最大化期望回报的正确做法，但从长远来看，探索可能会产生更大的总奖励。例如，假设贪婪动作的价值是确定的，而其他几项动作的评估结果几乎相同，但存在很大的不确定性。这种不确定性使得这些其他动作中的至少一个实际上可能比贪婪的动作要好，但是您不知道哪个。如果您还有很多时间步可以选择动作，那么探索非贪婪的动作并发现其中哪个比贪婪的动作更好可能会更好。在短期内，探索过程中的奖励较低，但从长期来看，奖励较高，因为发现更好的动作后，您可以多次利用它们。由于不可能通过任何一个单一的动作选择来进行探索和利用，因此人们经常提到探索与利用之间的“冲突”。

在任何特定情况下，是进行探索还是利用更好，都将以复杂的方式取决于估计的精确值，不确定性以及剩余时间步的数量。对于  $k$  臂赌博机和相关问题的特殊数学公式，有许多复杂的方法可以平衡探索和利用。但是，大多数这些方法都对平稳性和先验知识做出了强有力的假设，这些假设在应用以及我们将在后续章节中考虑的全面强化学习问题中被违反或无法验证。当这些方法的理论假设不适用时，对这些方法的最优性或有界损失的保证就不太满意。

在本书中，我们不用担心以复杂的方式平衡探索与利用；我们只担心如何平衡它们。在本章中，我们介绍了几种用于解决  $k$  臂赌博机问题的简单平衡方法，并表明它们比总是利用的方法好得多。平衡探索与利用的需要是强化学习中出现的一个独特挑战；我们的  $k$  臂赌博机问题的简单性使我们能够以特别清晰的形式展示这一点。

## 2.2 动作值方法

我们首先更仔细地研究用于估计动作值的方法，以及用于使用估计值来做出动作选择决策的方法，我们统称为动作值方法。回想一下，动作的真实值是选择该动作时的平均奖励。估计这一点的一种自然方法是对实际收到的奖励进行平均：

$$Q_t(a) \doteq \frac{t \text{ 之前采取 } a \text{ 时的奖励总和}}{t \text{ 之前采取 } a \text{ 的次数}} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{I}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{I}_{A_t=a}}, \quad (2.1)$$

其中  $\mathbb{I}_{\text{predicate}}$  表示随机变量，如果 *predicate* 为真则为 1，否则为 0。如果分母为零，那么我们将  $Q_t(a)$  定义为某个默认值，例如 0。随着分母趋于无穷大，根据大数定律， $Q_t(a)$  收敛至  $q_*(a)$ 。我们称其为估计动作价值的样本平均方法，因为每个估计都是相关奖励样本的平均值。当然，这只是估计动作价值的一种方法，不一定是最佳方法。尽管如此，现在让我们继续使用这种简单的估计方法，然后转向如何将估计值用于选择动作的问题。


最简单的动作选择规则是选择估计值最高的动作之一，也就是上一节定义的贪婪动作之一。如果有多个贪婪动作，则以某种任意方式（可能是随机的）在其中进行选择。我



们将此贪婪动作选择方法写为

$$A_t \doteq \arg \max_a Q_t(a), \quad (2.2)$$

其中  $\arg \max_a$  表示动作  $a$ ，该动作让其后的表达式被最大化（同样，平局任意）。贪婪的动作选择总是利用当前的知识来最大化立即的奖励。它不花时间对所有明显劣等的动作进行抽样，以查看它们是否会真的更好。一个简单的替代方法是在大多数时间贪婪地表现，但每隔一段时间，比如说小概率  $\varepsilon$ ，而是从所有具有相等概率的动作中随机选择，而与动作值估计无关。我们将这种接近贪婪的动作选择规则的方法称为  $\varepsilon$ -greedy 方法。这些方法的一个优点是，在极限中，随着步数的增加，每个动作将被采样无限次，从而确保所有  $Q_t(a)$  收敛到  $q_*(a)$ 。当然，这意味着选择最优动作的概率收敛到大于  $1-\varepsilon$ ，即接近确定性。然而，这些只是渐近保证，而对于这些方法的实际有效性却很少提到。

 **练习 2.1** 在  $\varepsilon$ -greedy 动作选择中，对于两个动作和  $\varepsilon = 0.5$  的情况，选择贪婪动作的概率是多少？ □

## 2.3 10 臂试验

为了粗略地评估贪婪和  $\varepsilon$ -greedy 动作值方法的相对有效性，我们在一组测试问题中对它们进行了数值比较。这是一组 2000 个随机生成的  $k$  臂赌博机问题， $k = 10$ 。对于每个赌博机问题，如图 2.1 所示，动作值  $q_*(a)$ ， $a = 1, \dots, 10$  根据均值为 0 和方差为 1 的正态（高斯）分布进行选择。然后，当学习方法应用于在时间步  $t$  选择动作  $A_t$  的问题时，从均值为  $q_*(A_t)$  和方差为 1 的正态分布中选择实际奖励  $R_t$ 。这些分布在图 2.1 中以灰色表示。我们称这套测试任务为 10 臂试验。对于任何一种学习方法，我们都可以测量它的性能和行为，因为它在应用于其中一个赌博机问题时，随着 1000 多个时间步的经验而有所改善。这就构成了一次运行。重复进行 2000 次独立运行，每次运行都有一个不同的赌博机问题，我们得到了学习算法的平均行为的度量。

图 2.2 在 10 臂试验上比较了贪婪方法和上述两种贪婪方法（ $\varepsilon = 0.01$  和  $\varepsilon = 0.1$ ）。所有这些方法均使用样本平均技术形成其动作值估算值。上方的图显示了预期奖励随着经验的增加。贪婪方法在开始时的改进速度比其他方法稍快，但随后在较低的水平上趋于平稳。它获得的每一步奖励只有大约 1，相比之下，在这个试验上，最好的可能是大约 1.55。从长远来看，贪婪方法的表现要差得多，因为它经常陷入执行次优动作的困境。下图显示了贪婪方法仅在大约三分之一的任务中找到了最优动作。在其他三分之二中，其最佳动作的初始样本令人失望，并且从未恢复。 $\varepsilon$ -greedy 方法最终表现更好，因为它们不断探索并提高了识别最优动作的机会。 $\varepsilon = 0.1$  方法进行了更多探索，通常会更早地找到最优动作，但它从未选择该动作的时间超过了 91%。 $\varepsilon = 0.01$  方法的改进速度较慢，但在图中所示的两个性能指标上最终都比  $\varepsilon = 0.1$  方法更好。还可以随着时间的推移而减小  $\varepsilon$ ，以尝试获得高值和低值中的最佳值。

$\varepsilon$ -greedy 方法胜于贪婪方法的优势取决于任务。例如，假设奖励方差更大，比如 10 而不是 1。在奖励更嘈杂的情况下，需要更多的探索才能找到最优动作， $\varepsilon$ -greedy 方法应

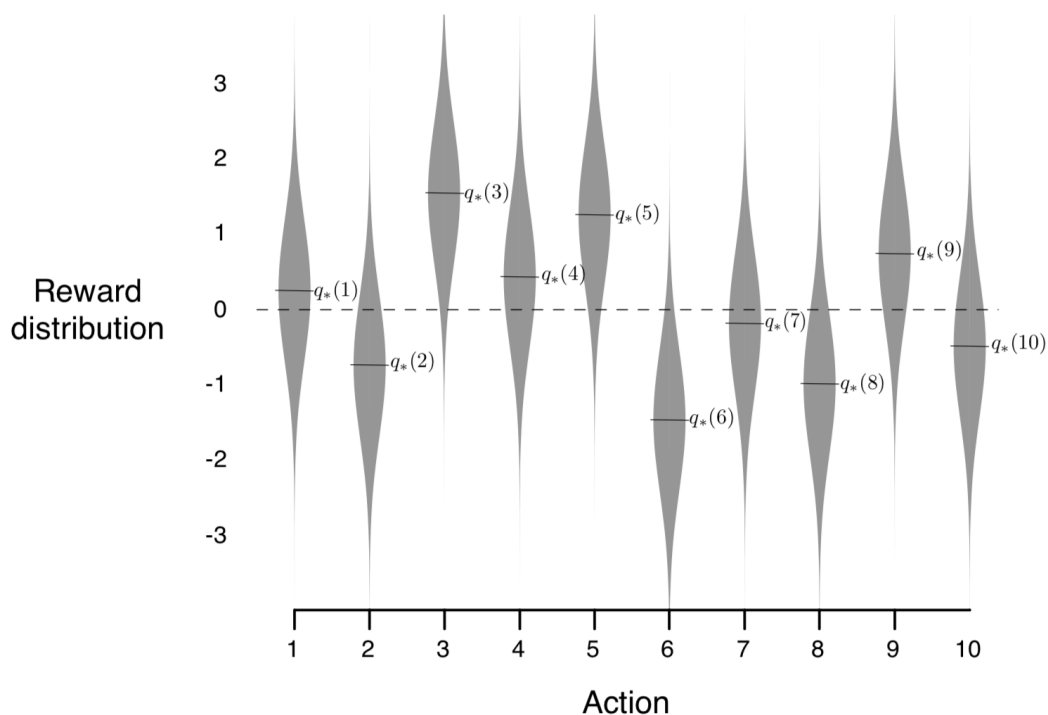


图 2.1: 10 臂试验中的赌博机问题示例。根据均值为零和单位方差的正态分布选择十个动作中每个动作的真实值  $q_*(a)$ , 然后根据这些灰色分布所表明的均值为  $q_*(a)$  和单位方差的正态分布选择实际奖励。

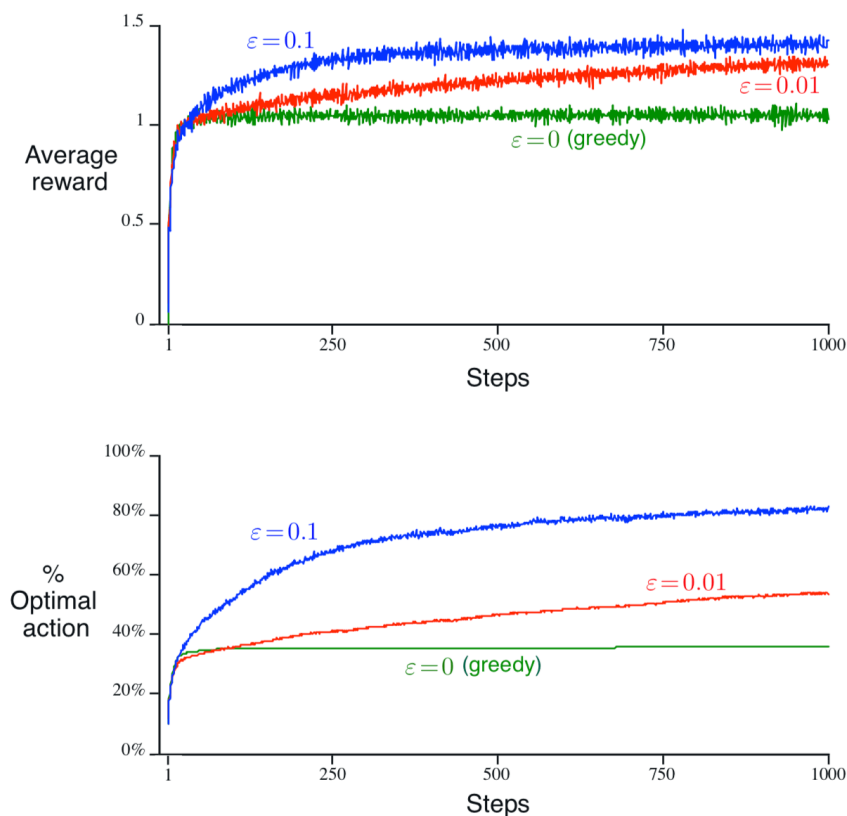




图 2.2:  $\epsilon$ -greedy 动作值方法在 10 臂试验上的平均性能。这些数据是具有不同的赌博机问题的 2000 多次运行的平均值。所有方法都使用样本平均值作为它们的动作值估计。

该比贪婪方法更好。另一方面，如果奖励方差为零，那么贪婪的方法在尝试一次之后就会知道每个动作的真实价值。在这种情况下，贪婪的方法实际上可能执行得最好，因为它很快就会找到最优动作，然后再也不用去探索。但是，即使在确定性的情况下，如果我们弱化一些其他假设，探索也有很大的优势。例如，假设赌博机任务是非平稳的，也就是说，动作的真实值随着时间的推移而变化。在这种情况下，甚至在确定性的情况下也需要探索，以确保非贪婪的动作之一不会变得比贪婪的动作更好。正如我们将在接下来的几章中看到的，非平稳性是强化学习中最常见的情况。即使潜在的任务是平稳的和确定的，学习者也面临着一组赌博机般的决策任务，随着学习的进行和智能体的决策策略的变化，每个任务都会随着时间的推移而变化。强化学习需要在探索和利用之间取得平衡。

 **练习 2.2: Bandit example** 考虑一个具有  $k = 4$  个动作的  $k$  臂赌博机问题，分别表示为 1、2、3 和 4。考虑对该问题应用赌博机算法，该算法使用  $\epsilon$ -greedy 动作选择，样本平均动作值估计和对于所有  $a$ ， $Q_1(a) = 0$ 。假设动作和奖励的初始序列为  $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$ 。在某些时间步上， $\epsilon$  情况可能已经发生，导致随机选择一个动作。这肯定发生在哪些时间步？在哪些时间步这可能已经发生？ □

 **练习 2.3** 在图 2.2 所示的比较中，就累积奖励和选择最佳动作的概率而言，哪种方法在长期内表现最好？它会好多少？量化地表达你的答案。 □

## 2.4 渐增实现

到目前为止，我们已经讨论过的动作值方法都将动作值估计为观察到的奖励的样本平均值。现在，我们转向如何以计算有效的方式来计算这些平均值的问题，尤其是在恒定内存和恒定的每时间步计算的情况下。

为了简化表示法，我们集中在单个动作上。现在让  $R_i$  表示第  $i$  次选择该动作后收到的奖励，让  $Q_n$  表示其被选择  $n - 1$  次后其动作值的估计，我们现在可以简单地写为

$$Q_n \doteq \frac{R_1 + R_2 + \dots + R_{n-1}}{n - 1}.$$

显而易见的实现是维护所有奖励的记录，然后在需要估计值的时候执行此计算。然而，如果这样做了，那么内存和计算需求将随着时间的推移而增长，正如我们所看到的那样。每一笔额外的奖励都需要额外的内存来存储它，并需要额外的计算来计算分子中的总和。

正如您可能怀疑的那样，这并不是真正必要的。利用处理每个新奖励所需的小而恒定的计算来设计更新平均值的递增公式很容易。给定  $Q_n$  和第  $n$  次奖励  $R_n$ ，所有  $n$  个奖励的新平均值可由下式计算

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^n R_i$$

$$\begin{aligned}
&= \frac{1}{n} \left( R_n + \sum_{i=1}^{n-1} R_i \right) \\
&= \frac{1}{n} \left( R_n + (n-1) \frac{1}{n-1} \sum_{i=1}^{n-1} R_i \right) \\
&= \frac{1}{n} (R_n + (n-1)Q_n) \\
&= \frac{1}{n} (R_n + nQ_n - Q_n) \\
&= Q_n + \frac{1}{n} [R_n - Q_n], \tag{2.3}
\end{aligned}$$

对于  $n = 1$ ，它仍然成立，对于任意  $Q_1$ ，获得  $Q_2 = R_1$ 。此实现只需要  $Q_n$  和  $n$  的内存，每个新的奖励只需要很小的计算 (2.3)。

此更新规则 (2.3) 的形式在本书中经常出现。一般形式是

$$NewEstimate \leftarrow OldEstimate + StepSize [Target - OldEstimate]. \tag{2.4}$$

表达式  $[Target - OldEstimate]$  是估计中的错误。它是通过向“**Target**”迈进一步来减少。假定 **target** 可能指示移动的理想方向，尽管它可能会有噪声。在上述情况下，例如，**target** 是第  $n$  个奖励。

请注意，增量方法 (2.3) 中使用的步长参数 ( $StepSize$ ) 随时间步而变化。在处理动作  $a$  的第  $n$  个奖励时，该方法使用步长参数  $\frac{1}{n}$ 。在本书中，我们用  $\alpha$  或更一般地用  $\alpha_t(a)$  表示步长参数。

下面的框中显示了使用递增计算的样本平均和  $\epsilon$ -greedy 动作选择的完整赌博机算法的伪代码。假定函数  $bandit(a)$  采取动作并返回相应的奖励。

#### 一个简单的赌博机算法

初始化,  $a$  从 1 到  $k$ :

$$Q(a) \leftarrow 0$$

$$N(a) \leftarrow 0$$

永久循环:

$$A \leftarrow \begin{cases} \arg \max_a Q(a) & \text{概率为 } 1 - \epsilon \\ \text{随机动作} & \text{概率为 } \epsilon \end{cases}$$

$$R \leftarrow bandit(A)$$

$$N(A) \leftarrow N(A) + 1$$

$$Q(A) \leftarrow Q(A) + \frac{1}{N(A)} [R - Q(A)]$$

## 2.5 非平稳问题

到目前为止讨论的平均方法适用于平稳的赌博机问题，即奖励概率不随时间变化的赌博机问题，如前所述，我们经常遇到实际上完全非平稳的强化学习问题。在这种情况下，将更多的权重放在近期的奖励上，而不是放在过去的长期奖励上，这是有意义的。最流行的方法之一是使用恒定步长参数。例如，用于更新  $n-1$  个过去奖励的平均  $Q_n$  的增量更新规则 (2.3) 被修改为

$$Q_{n+1} \doteq Q_n + \alpha [R_n - Q_n], \quad (2.5)$$

其中步长参数  $\alpha \in (0, 1]$  是常量。这导致  $Q_{n+1}$  是过去奖励和初始估计  $Q_1$  的加权平均值：

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha [R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha) Q_n \\ &= \alpha R_n + (1 - \alpha) [\alpha R_{n-1} + (1 - \alpha) Q_{n-1}] \\ &= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \cdots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \\ &= (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} R_i. \end{aligned} \quad (2.6)$$

我们称其为加权平均值，是因为权重之和为  $(1 - \alpha)^n + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} = 1$ ，您可以自己检查一下。请注意，给予奖励  $R_i$  的权重  $\alpha (1 - \alpha)^{n-i}$  取决于  $n - i$  之前观察到的奖励数量。 $1 - \alpha$  小于 1，因此给予  $R_i$  的权重随着介入奖励数量的增加而减少。实际上，权重根据  $1 - \alpha$  的指数呈指数衰减。（如果  $1 - \alpha = 0$ ，则由于权重为  $0^0 = 1$ ，所有的权重都会在最后一个奖励  $R_n$  上）。因此，有时将其称为指数近期加权平均。

有时可以很方便地逐步改变步长参数。令  $\alpha_n(a)$  表示步长参数，该参数用于处理在第  $n$  次选择动作  $a$  后收到的奖励。正如我们已经注意到的那样，选择  $\alpha_n(a) = \frac{1}{n}$  会导致样本平均方法，通过大数定律可以保证该方法收敛到真实的动作值。但是，当然不能保证对序列  $\{\alpha_n(a)\}$  的所有选择都收敛。随机逼近理论中的一个著名结果为我们提供了以概率 1 收敛的条件：

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \quad \text{和} \quad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty \quad (2.7)$$

第一个条件以确保步幅足够大，以最终克服任何初始条件或随机波动。第二个条件保证最终步长变得足够小以确保收敛。

注意，对于样本平均情况， $\alpha_n(a) = \frac{1}{n}$ ，这两个收敛条件都满足，但是对于恒定步长参数， $\alpha_n(a) = \alpha$ ，则情况不满足。在后一种情况下，不满足第二个条件，这表明估计值从未完全收敛，而是随着最近收到的奖励而不断变化。如上所述，这在非平稳环境中实际上是可取的，有效非平稳问题是强化学习中最常见的问题。此外，满足条件 (2.7) 的步长参数序列通常会收敛很慢，或者需要进行大量调整才能获得令人满意的收敛速度。尽管满足这些收敛条件的步长参数序列在理论工作中经常使用，但是在应用和实证研究中



却很少使用它们。

🔥 **练习 2.4** 如果步长参数  $\alpha_n$  不恒定，则估计值  $Q_n$  是先前接收的奖励的加权平均值，其权重与 (2.6) 给出的权重不同。就步长参数的序列而言，对于一般情况，类似于 (2.6)，每个先前奖励的权重是多少？ □

🔥 **练习 2.5** (编程) 设计并进行实验，以证明样本平均方法对于解决非平稳问题的困难。使用 10 臂试验的修改版本，其中起初所有  $q_*(a)$  均相等，然后进行独立的随机游走（比如在每一步对所有  $q_*(a)$  加上均值为零且标准差为 0.01 的正态分布增量）。绘制类似图 2.2 所示的图，为使用样本平均值进行增量计算的动作值方法，和另一使用恒定步长参数  $\alpha = 0.1$  的动作值方法去准备图。使用  $\varepsilon = 0.1$  和更长的运行时间，比如 10,000 步。 □

## 2.6 乐观初始值

到目前为止，我们讨论的所有方法都在某种程度上依赖于初始动作值估计  $Q_1(a)$ 。按统计学的说法，这些方法与它们最初的估计是有偏的。对于样本平均方法，只要至少选择了一次所有动作，偏差就会消失，但对于  $\alpha$  恒定的方法，偏差是永久性的，尽管如 (2.6) 所示，随着时间的推移，偏差会减少。在实践中，这种偏差通常不是问题，有时会非常有帮助。缺点是，初始估计实际上变成了必须由用户选择的一组参数，哪怕只是将它们全部设置为零。好处是，它们提供了一种简单的方法来提供一些关于期望奖励水平的先验知识。

初始动作值也可以用作鼓励探索的简单方式。假设我们没有像在 10 臂试验上那样将初始动作值设置为零，而是全部设置为 5。回想一下，这个问题中的  $q_*(a)$  是从均值为 0 和方差为 1 的正态分布中选择的。因此，初始估计值为 5 是非常乐观的。但这种乐观主义鼓励对动作值方法的探索。无论最初选择哪种动作，奖励都会低于初始估计；学习者会切换到其他动作，对所获得的奖励感到“失望”。结果是，在值估计收敛之前，所有动作都会尝试几次。即使总是选择贪婪的动作，系统也会进行大量的探索。

图 2.3 显示了对于所有  $a$  在使用  $Q_1(a) = +5$  的贪婪方法的 10 臂赌博机试验中的性能。为了进行比较，还显示了  $Q_1(a) = 0$  的  $\varepsilon$ -greedy 方法。起初，乐观方法的性能较差，因为它会进行更多的探索，但最终效果会更好，因为其探索会随着时间的推移而减少。我们称此种鼓励探索的技术为乐观的初始值。我们认为这是一个简单的技巧，可以有效解决平稳问题，但远不是鼓励探索的普遍有用方法。例如，它不适用于非平稳问题，因为它的探索动力本来就是暂时的。如果任务发生变化，从而产生了新的探索需求，则此方法无济于事。确实，任何以任何特殊方式关注初始条件的方法都不太可能解决一般的非平稳情况。时间的开始只发生一次，因此我们不应该过多地关注它。这种批评也适用于样本平均方法，该方法也将时间的开始视为特殊事件，用相同的权重对所有随后的奖励进行平均。但是，所有这些方法都非常简单，并且其中一种（或它们的一些简单组合）在实践中通常就足够了。在本书的其余部分中，我们将频繁使用这些简单的探索技术中的几种。

🔥 **练习 2.6 : Mysterious Spikes** 图 2.3 中显示的结果应该是相当可靠的，因为它们是多

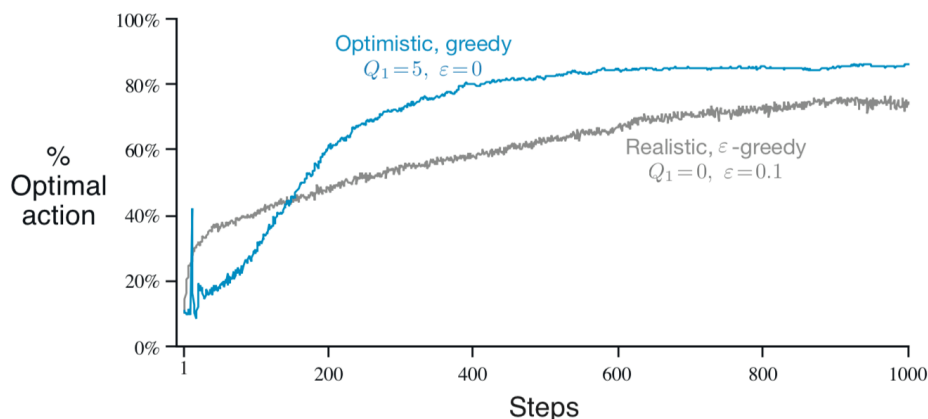



图 2.3: 乐观的初始动作值估计对 10 臂试验的影响。两种方法都使用恒定的步长参数  $\alpha = 0.1$ 。

个单独的、随机选择的 10 臂赌博机任务的平均值。那么，为什么乐观方法的曲线的早期部分会有振荡和尖峰呢？换句话说，是什么让这种方法在特定的早期步骤上表现得更好或更差呢？

 **练习 2.7: Unbiased Constant-Step-Size Trick** 在本章的大部分内容中，我们使用样本平均来估计动作值，因为样本平均不会产生恒定步长所产生的初始偏差（参见导致 (2.6) 的分析）。然而，样本平均并不是一个完全令人满意的解决方案，因为它们在非平稳问题上的表现可能很差。是否有可能避免固定步长的偏差，同时保持它们在非平稳问题上的优势？一种方法是使用步长为

$$\beta_n \doteq \alpha / \bar{o}_n, \quad (2.8)$$

来处理特定动作的第  $n$  次奖励，其中  $\alpha > 0$  是常规的恒定步长，而  $\bar{o}_n$  是从 0 开始的跟踪：

$$\bar{o}_n \doteq \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1}), \quad \text{对于 } n \geq 0, \quad \bar{o}_0 \doteq 0. \quad (2.9)$$

进行类似 (2.6) 中的分析，表明  $Q$  值是没有初始偏差的指数近期加权平均。

## 2.7 置信上限动作选择

探索是必要的，因为动作值估计的准确性总是存在不确定性。贪婪的动作是那些目前看起来最好的动作，但其他一些动作实际上可能更好。 $\varepsilon$ -greedy 的动作选择迫使人们尝试非贪婪的动作，但不加区别，不偏好那些近乎贪婪或特别不确定的动作。最好根据非贪婪动作的实际最优潜力进行选择，同时考虑到它们的估计与最大值有多接近，以及这些估计中的不确定性。这样做的一种有效方法是根据

$$A_t \doteq \arg \max_a \left[ Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right], \quad (2.10)$$

选择动作，其中  $\ln t$  表示  $t$  的自然对数（必须将  $e \approx 2.71828$  的数字提高到等于  $t$ ）， $N_t(a)$  表示在时间  $t$  之前选择动作  $a$  的次数（(2.1) 中的分母），而数字  $c > 0$  控制探索的程度。如果  $N_t(a) = 0$ ，那么  $a$  被认为是最大化动作。

此置信上限（upper confidence bound, UCB）动作选择的思路是平方根项是对  $a$  的值估计中的不确定性或方差的度量。因此，最大化的量是动作  $a$  可能真值的一种上界，并且  $c$  确定置信水平。每选择  $a$  一次，不确定度都可能减小： $N_t(a)$  递增，并且如分母所示，不确定度项减小。另一方面，每次选择除  $a$  以外的动作时， $t$  增加，但  $N_t(a)$  不增加；因为  $t$  出现在分子中，所以不确定性估计增加。使用自然对数意味着增加随着时间的推移变小，但是无界的；所有动作最终都将被选择，但是具有较低估计值的动作，或者已经被频繁选择的动作，将随着时间的推移以递减的频率被选择。

图2.4显示了在 10 臂试验上使用 UCB 的结果。UCB 通常表现良好，如此处所示，但是要比  $\epsilon$ -greedy 从赌博机扩展到本书其余部分中考虑的更一般的强化学习环境要困难得多。一个难题在于处理非平稳问题；需要比第2.5节中介绍的方法更复杂的方法。另一个困难是处理大型状态空间，尤其是在使用本书第二部分中开发的函数逼近时。在这些更高级的设置中，UCB 动作选择的想法通常不切实际。

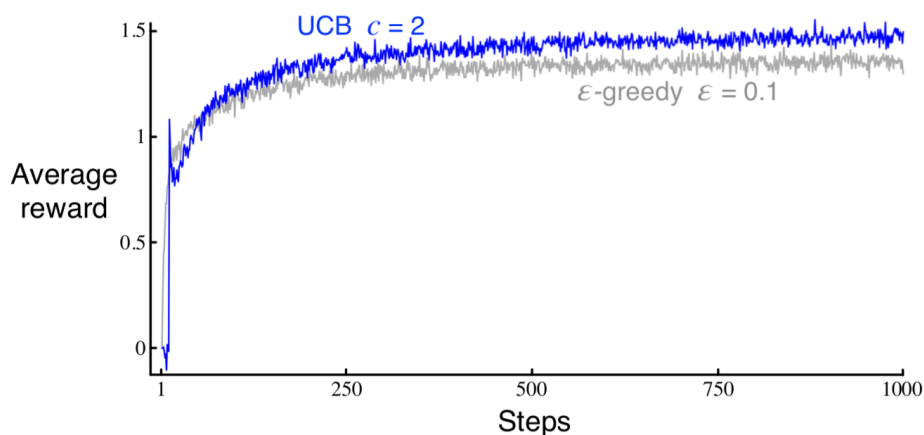



图 2.4: 10 臂试验上 UCB 动作选择的平均性能。如图所示，UCB 通常比  $\epsilon$ -greedy 动作选择表现得更好，除了在前  $k$  个步中，它在尚未尝试的动作中随机选择。

 **练习 2.8:** *UCB Spikes* 在图2.4中，UCB 算法在第 11 步显示了明显的峰值性能。这是为什么？请注意，为了让你的答案完全令人满意，必须解释为什么奖励在第 11 步增加，为什么在随后的步减少。提示：如果  $c = 1$ ，则尖峰不那么突出。 □


## 2.8 梯度赌博机算法

到目前为止，在本章中，我们已经考虑了估计动作值并使用这些估计来选择动作的方法。这通常是一种很好的方法，但不是唯一可能的方法。在本节中，我们考虑学习每个动作  $a$  的数值偏好（preference），我们将其命名为  $H_t(a)$ 。偏好越大，采取该动作的频率就越高，但偏好没有奖励方面的解释。只有一个动作相对于另一个动作的相对偏好是重要的；如果我们将 1000 加到所有动作偏好上，则不会影响动作概率，这些概率是根据

如下的 soft-max 分布（即 Gibbs 分布或 Boltzman 分布）确定的：

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^k e^{H_t(b)}} \doteq \pi_t(a), \quad (2.11)$$

这里我们还引入了一个有用的新符号， $\pi_t(a)$ ，用来表示在时间  $t$  采取动作  $a$  的概率。最初，所有动作偏好都是相同的（例如，对于所有  $a$ ， $H_1(a) = 0$ ），使得所有动作被选择的概率相等。

 **练习 2.9** 证明了在两种动作情况下，soft-max 分布与统计学和人工神经网络中常用的 logistic 函数或 sigmoid 函数的分布相同。□

基于随机梯度上升的思想，给出了 soft-max 动作偏好的自然学习算法。在每个步骤中，在选择动作  $A_t$  并接收奖励  $R_t$  之后，将通过以下方式更新动作偏好：

$$\begin{aligned} H_{t+1}(A_t) &\doteq H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)), & \text{and} \\ H_{t+1}(a) &\doteq H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a), & \text{for all } a \neq A_t, \end{aligned} \quad (2.12)$$

其中  $\alpha > 0$  是步长参数， $\bar{R}_t \in \mathbb{R}$  是到时间  $t$ （但不包括时间  $t$ ）为止的所有奖励的平均值，它可以按照第 2.4 节（如果问题是非平稳的，则是第 2.5 节）中描述的方式递增计算。 $\bar{R}_t$  用作比较奖励的基准线。如果奖励高于基线，那么在将来采取  $A_t$  的概率就会增加，如果奖励低于基线，那么这种可能性就会降低。未选择的动作以相反的方向移动。

图 2.5 显示了在 10 臂试验的变体上使用梯度赌博机算法的结果，在该变型中，根据均值为 4 而不是零的正态分布选择真实的期望奖励（并且与以前一样具有单位方差）。由于奖励基线项的存在，所有奖励的这种上移对梯度赌博机算法完全没有影响，它会立即适应新的水平。但是，如果省略基线（即，如果在 (2.12) 中将  $\bar{R}_t$  视为常量零），则性能将显著降低，如图所示。

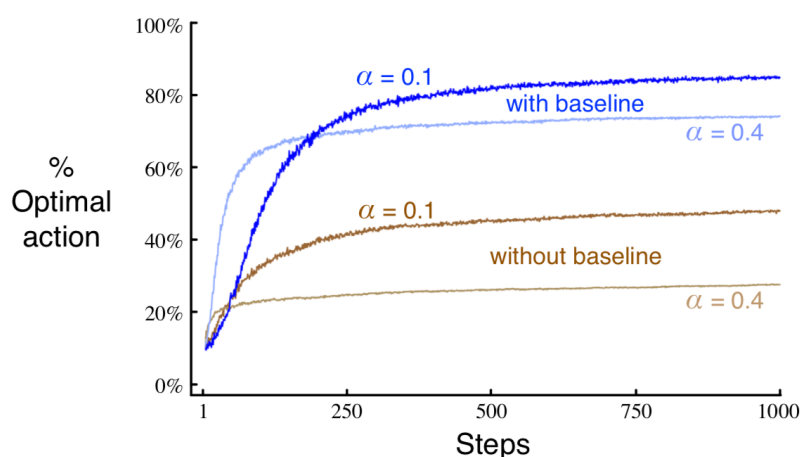


图 2.5: 在 10 臂试验上，当  $q_*(a)$  被选择为接近 4 而不是接近零时，具有和不具有奖励基线的梯度赌博机算法的平均性能。

### 随机梯度上升的赌博机梯度算法

通过将梯度赌博机算法理解为梯度上升的随机近似，可以更深入地了解梯度赌博机算法。在精确梯度上升中，每个动作偏好  $H_t(a)$  将与增量对性能的影响成正比地递增：

$$H_{t+1}(a) \doteq H_t(a) + \alpha \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}, \quad (2.13)$$

这里的性能衡量标准是期望奖励：

$$\mathbb{E}[R_t] = \sum_x \pi_t(x) q_*(x), \quad (2.14)$$

增量影响的度量是该性能度量相对于动作偏好的偏导数。当然，由于假设我们不知道  $q_*(x)$ ，因此在我们的情况下不可能完全实现梯度上升，但是实际上我们的算法 (2.12) 的更新在期望值上等于 (2.13)，使该算法成为随机梯度上升的一个实例。计算表明，此过程仅需进行初次演算，但需要执行几个步骤。首先，我们仔细研究一下确切的性能梯度：

$$\begin{aligned} \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \left[ \sum_x \pi_t(x) q_*(x) \right] \\ &= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\ &= \sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)}, \end{aligned}$$

其中  $B_t$ ，称为基线，可以是不依赖于  $x$  的任何标量。我们可以在这里包含一个基线，而不改变等式，因为在所有的动作中梯度总和为零， $\sum_x \frac{\partial \pi_t(x)}{\partial H_t(a)} = 0$ ；当  $H_t(a)$  改变时，一些动作的概率上升，一些下降，但是变化的总和一定为零，因为概率的总和总是 1。

接下来，我们将总和的每一项乘以  $\pi_t(x)/\pi_t(x)$ ：

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_x \pi_t(x) (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(x).$$

方程现在是期望的形式，将随机变量  $A_t$  的所有可能值  $x$  相加，然后乘以取这些值的概率。因此：

$$\begin{aligned} &= \mathbb{E} \left[ (q_*(A_t) - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \right] \\ &= \mathbb{E} \left[ (R_t - \bar{R}_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \right], \end{aligned}$$

这里我们选择基线  $B_t = \bar{R}_t$ ，并用  $R_t$  代替  $q_*(A_t)$ ，这是允许的，因为  $\mathbb{E}[R_t|A_t] =$



$q_*(A_t)$ 。不久，我们将建立  $\frac{\partial \pi_t(x)}{\partial H_t(a)} = \pi_t(x)(\mathbb{I}_{a=x} - \pi_t(a))$ ，如果  $a = x$ ，则将  $\mathbb{I}_{a=x}$  定义为 1，否则定义为 0。现在，我们有

$$\begin{aligned} &= \mathbb{E}[(R_t - \bar{R}_t)\pi_t(A_t)(\mathbb{I}_{a=A_t} - \pi_t(a))/\pi_t(A_t)] \\ &= \mathbb{E}[(R_t - \bar{R}_t)(\mathbb{I}_{a=A_t} - \pi_t(a))] \end{aligned}$$

回想一下，我们的计划一直是将性能梯度作为我们可以在每一步中采样的期望，就像我们刚才所做的那样，然后在每一步中与样例成比例地更新。用上述期望的样例代替 (2.13) 中的性能梯度：

$$H_{t+1}(a) = H_t(a) + \alpha(R_t - \bar{R}_t)(\mathbb{I}_{a=A_t} - \pi_t(a)), \quad \text{for all } a,$$

您可能会认为它等同于我们的原始算法 (2.12)。

因此，正如我们假设的那样，它仅仅是为了表明  $\frac{\partial \pi_t(x)}{\partial H_t(a)} = \pi_t(x)(\mathbb{I}_{a=x} - \pi_t(a))$ 。回忆一下导数的标准商规则：

$$\frac{\partial}{\partial x} \left[ \frac{f(x)}{g(x)} \right] = \frac{\frac{\partial f(x)}{\partial x} g(x) - f(x) \frac{\partial g(x)}{\partial x}}{g(x)^2}$$

使用这个，我们可以写出

$$\begin{aligned} \frac{\partial \pi_t(x)}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \pi_t(x) \\ &= \frac{\partial}{\partial H_t(a)} \left[ \frac{e^{H_t(x)}}{\sum_{y=1}^k e^{H_t(y)}} \right] \\ &= \frac{\frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_{y=1}^k e^{H_t(y)} - e^{H_t(x)} \frac{\partial \sum_{y=1}^k e^{H_t(y)}}{\partial H_t(a)}}{(\sum_{y=1}^k e^{H_t(y)})^2} \quad (\text{by the quotient rule}) \\ &= \frac{\mathbb{I}_{a=x} e^{H_t(x)} \sum_{y=1}^k e^{H_t(y)} - e^{H_t(x)} e^{H_t(a)}}{(\sum_{y=1}^k e^{H_t(y)})^2} \quad (\text{because } \frac{\partial e^x}{\partial x} = e^x) \\ &= \frac{\mathbb{I}_{a=x} e^{H_t(x)}}{\sum_{y=1}^k e^{H_t(y)}} - \frac{e^{H_t(x)} e^{H_t(a)}}{(\sum_{y=1}^k e^{H_t(y)})^2} \\ &= \mathbb{I}_{a=x} \pi_t(x) - \pi_t(x) \pi_t(a) \\ &= \pi_t(x)(\mathbb{I}_{a=x} - \pi_t(a)) \end{aligned}$$

我们刚刚证明了梯度赌博机算法的期望更新等于期望奖励的梯度，从而证明了该算法是随机梯度上升的一个实例。这保证了算法具有鲁棒的收敛性。


请注意，除了不依赖于所选动作之外，我们不需要奖励基线的任何属性。例如，我们可以将它设置为 0 或 1000，算法仍将是随机梯度上升的实例。基线的选择不会影响算法的期望更新，但会影响更新的方差，从而影响收敛速度 (如图 2.5 中所示)。选择它作为奖励的平均值可能不是最好的，但它很简单，在实践中效果很好。

## 2.9 关联搜索（上下文赌博机）

到目前为止，我们在本章中只考虑了非关联任务，即不需要将不同的动作与不同的情况相关联的任务。在这些任务中，学习者要么在任务平稳时试图找到单一的最佳动作，要么在任务非平稳时试图跟踪随着时间的推移而变化的最佳动作。然而，在一般的强化学习任务中，有不只一种情况，学习策略的目标是：从情况到在这些情况下最好的动作的映射。为了为整个问题做好准备，我们简要讨论将非关联任务扩展到关联环境的最简单方法。

举个例子，假设有几个不同的  $k$  臂赌博机任务，每一步你都会随机选择其中一个。因此，赌博机任务一步一步地随机变化。在您看来，这是一个单一的非平稳  $k$  臂赌博机任务，其真实动作值随步骤的不同而随机变化。您可以尝试使用本章中介绍的可以处理非平稳性的方法，但除非真正的动作值缓慢更改，否则这些方法不会很好地工作。但是，现在假设，当您选择了一个赌博机任务时，您会得到一些关于其身份（但不是其动作值）的独特线索。也许您面对的是一台实际的老虎机，它会随着动作值的改变而改变其显示颜色。现在，您可以学习一种策略，它将每个任务（由您看到的颜色表示）与面对该任务时要采取的最佳动作相关联-例如，如果是红色，则选择臂 1；如果是绿色，则选择臂 2。在没有任何区分赌博机任务的信息的情况下，使用正确的策略通常可以做得更好。

这是关联搜索任务的一个例子，之所以被称为关联搜索任务，是因为它既涉及到寻找最佳动作的试错学习，也涉及到将这些动作与它们最好的情况相关联。在文献中，关联搜索任务现在通常被称为上下文赌博机。关联搜索任务介于  $k$  臂赌博机问题和完全强化学习问题之间。它们类似于完全强化学习问题，因为它们涉及到学习策略，但又像我们版本的  $k$  臂赌博机问题，因为每个动作只影响立即奖励。如果动作被允许影响在下一情况以及奖励，那么我们就有了完全的强化学习问题。我们将在下一章提出这个问题，并在本书的其余部分考虑其影响。

 **练习 2.10** 假设你面对的是一个 2 臂赌博机任务，其真实动作值随时间步而随机变化。具体地说，假设对于任何时间步，动作 1 和 2 的真值分别为 0.1 和 0.2，概率为 0.5（情况 A），以及 0.9 和 0.8，概率为 0.5（情况 B）。如果你在任何一步都不能说出你面对的是哪一种情况，你能达到的最好的成功期望是什么，你应该如何行动来实现它？现在假设在每一步中都被告知您面对的是情况 A 还是情况 B（尽管您仍然不知道真实的动作值）。这是一个关联搜索任务。在这个任务中，你能达到的最好的成功期望是什么？你应该如何行动才能达到这个目标？ □

## 2.10 总结

在本章中，我们介绍了几种平衡探索和利用的简单方法。 $\epsilon$ -greedy 方法是一小部分时间随机选择动作，而 UCB 方法是确定性地选择，但通过在每个步骤巧妙地偏爱到目前为止收到的样本较少的动作来实现探索。梯度赌博机算法不是估计动作值，而是动作偏好，并且使用 *softmax* 分布以分级的、概率的方式偏爱更喜欢的动作。乐观地初始化估

计的简单计策甚至会导致贪婪的方法进行显著的探索。

自然会问这些方法中哪种方法最好。尽管这通常是一个很难回答的问题，但我们可以在本章中使用 10 臂试验运行它们，并比较它们的性能。麻烦的是它们都有一个参数。为了获得有意义的比较，我们必须将其性能视为其参数的函数。到目前为止，我们的图形显示了每种算法和参数设置随时间的学习过程，以生成该算法和参数设置的学习曲线。如果我们为所有算法和所有参数设置绘制学习曲线，则该图将过于复杂且拥挤而无法进行清晰的比较。相反，我们根据 1000 步的平均值总结出完整的学习曲线。该值与学习曲线下的面积成比例。图 2.6 显示了针对本章中的各种赌博机算法的度量，每种度量都是其自身参数的函数，在 x 轴上以单个刻度显示。这种图称为参数研究。请注意，参数值相差 2 倍，并以对数刻度显示。还要注意每种算法性能的倒 U 形特征；所有算法在其参数的中间值上表现最好，既不能过大也不能过小。在评估一种方法时，我们不仅应关注其最佳参数设置的性能，而且还应关注其对参数值的敏感程度。所有这些算法都是相当不敏感的，它们在大约相差一个数量级的参数值范围内表现良好。总体而言，在这个问题上，UCB 似乎表现最好。

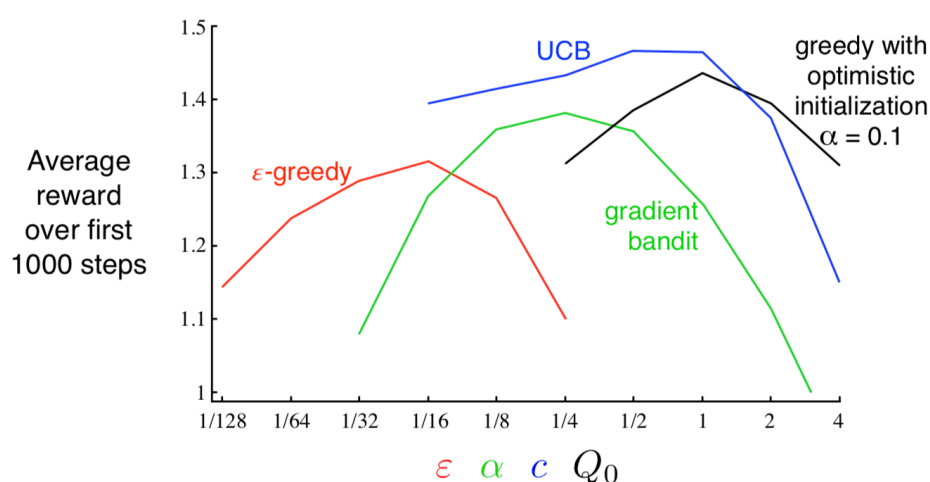


图 2.6: 本章介绍的各种赌博机算法的参数研究。每一个点都是在特定的参数设置下，用特定的算法通过 1000 步获得的平均奖励。

尽管它们很简单，但我们认为本章中介绍的方法可以被认为是最先进技术。有更复杂的方法，但是它们的复杂性和假设使其对于我们真正关注的完全强化学习问题不切实际。从第 5 章开始，我们将介绍解决完全强化学习问题的学习方法，这些方法部分使用了本章中探讨的简单方法。

虽然本章探讨的简单方法可能是我们目前所能做的最好的，但它们远不是平衡探索和利用问题的完全令人满意的解决方案。

一种经过充分研究的平衡  $k$  臂赌博机问题中的探索与利用的方法是计算一种特殊的作用值，称为 Gittins 指数。在某些重要的特殊情况下，尽管确实需要对可能出现的问题的先验分布的完整知识，但我们通常认为这是不可用的，但这种计算是易于处理的，并直接导致最优解。另外，这种方法的理论和计算可扩展性似乎都不能推广到我们在本书其余部分中考虑的完全强化学习问题。

Gittins 指数方法是贝叶斯方法的一个实例，贝叶斯方法假设动作值的初始分布是已知的，然后在每个步骤之后精确地更新分布（假设真实的动作值是固定的）。通常，更新计算可能非常复杂，但是对于某些特殊分布（称为共轭先验），它们很容易。一种可能性是，然后根据每个步骤的动作是最佳动作的后验概率来选择它们。这种方法有时称为后验采样或汤普森采样，其执行效果通常与本章介绍的最佳无分布方法相似。

在贝叶斯环境中，甚至可以考虑计算探索与利用之间的最佳平衡。可以为任何可能的动作计算每个可能的立即奖励的概率以及动作值的后验分布。这种不断变化的分布成为问题的信息状态。给定 1000 个步骤的时间范围，可以考虑所有 1000 个步骤的所有可能动作，所有可能的结果奖励，所有可能的下一个动作，所有下一个奖励，等等。有了这些假设，就可以确定每个可能事件链的奖励和概率，并且只需要选择最佳事件即可。但是，树的可能性展得非常迅速。即使只有两个动作和两个奖励，树也会有 22000 片叶子。精确地执行这种庞大的计算通常是不可行的，但是也许可以进行科学地近似。这种方法将有效地将赌博机问题变成完全强化学习问题的一个实例。最后，我们也许可以使用近似的强化学习方法（如本书第二部分中介绍的方法）来寻求最佳解决方案。但这是属于研究的范畴，不在本入门书的讨论范围之内。

 **练习 2.11**（编程）对于练习 2.5 中概述的非平稳情况，制作类似于图 2.6 的图。包括  $\alpha = 0.1$  的步长不变的  $\epsilon$ -greedy 算法。使用 200,000 步的运行，并作为每个算法和参数设置的性能度量，使用最近 100,000 步的平均奖励。□

## 2.11 书目与历史评论

**2.1** 在统计，工程和心理学领域都研究过赌博机问题。从统计学上讲，赌博机问题属于“实验的顺序设计”，由 Thompson（1933，1934）和 Robbins（1952）提出，并由 Bellman（1956）研究。Berry 和 Fristedt（1985）从统计学的角度提供了对赌博机问题的广泛论述。Narendra and Thathachar（1989）从工程学的角度来看待赌博机问题，对关注它们的各种理论传统进行了很好的讨论。在心理学中，赌博机问题在统计学习理论中发挥了作用（例如，Bush 和 Mosteller，1955 年；Estes，1950 年）。

启发式搜索文献中经常使用贪婪一词（例如 Pearl，1984 年）。在控制工程中，探索与利用之间的冲突称为识别（或估计）与控制之间的冲突（例如 Witten，1976b）。Feldbaum（1965）将其称为双重控制问题，指的是在试图控制不确定性系统时需要同时解决识别和控制两个问题。在讨论遗传算法的各个方面时，Holland（1975）强调了这种冲突的重要性，称其为开发需求与新信息需求之间的冲突。

**2.2** Thathachar 和 Sastry（1985）首次提出了解决我们的  $k$  臂赌博机问题的动作值方法。在自动机学习文献中，这些通常称为估计算法。术语动作值归因于 Watkins（1989）。最早使用  $\epsilon$ -greedy 方法的人可能也是 Watkins（1989，187 页），但是这个想法是如此简单，以至于可能会更早地使用。

**2.4-2.5** 该材料属于随机迭代算法的一般标题，Bertsekas 和 Tsitsiklis（1996）对此进行了很好的介绍。

**2.6** Sutton (1996) 在强化学习中使用了乐观初始化。

**2.7** Lai 和 Robbins (1985), Kaelbling (1993b) 和 Agrawal (1995) 进行了使用置信度上界估计来选择动作的早期工作。我们在本文中介绍的 UCB 算法在文献中称为 UCB1, 最早由 Auer, Cesa-Bianchi 和 Fischer (2002) 开发。

**2.8** 梯度赌博机算法是 Williams (1992) 引入的基于梯度的强化学习算法的特例, 后来又发展为我们在本书稍后将讨论的 *actor-critic* 和策略梯度算法。Balaraman Ravindran (个人交流) 对我们的发展产生了影响。那里以及 Greensmith, Bartlett, Baxter (2002, 2004) 和 Dick (2015) 提供了有关基线选择的进一步讨论。此类算法的早期系统研究是由 Sutton (1984) 完成的。

动作选择规则 (2.11) 的术语“soft-max”归因于 Bridle (1990)。该规则似乎是 Luce (1959) 首次提出的。

**2.9** Barto, Sutton 和 Brouwer (1981) 引入了术语“关联搜索”和相应的问题。关联强化学习这个词也已经用于关联搜索 (Barto 和 Anandan, 1985), 但是我们更喜欢保留这个术语作为完全强化学习问题的同义词 (如 Sutton, 1984)。(而且, 正如我们指出的那样, 现代文献也使用“上下文赌博机”一词来解决这个问题。) 我们注意到, Thorndike 的“效果定律”(在第 1 章中引用) 通过提及情境 (状态) 之间的关联链接的形成来描述关联搜索。) 和动作。根据操作性或工具性调节的术语学 (例如, Skinner, 1938), 判别性刺激是发信号通知存在特定的强化意外事件的刺激。用我们的术语来说, 不同的鉴别刺激对应于不同的状态。

**2.10** Bellman (1956) 首次展示了如何使用动态规划来在问题的贝叶斯公式内计算探索与利用之间的最佳平衡。Gittins 指数方法归功于 Gittins 和 Jones (1974)。Duff (1995) 展示了如何通过强化学习来学习针对赌博机问题的 Gittins 指数。Kumar (1985) 的调查很好地讨论了贝叶斯方法和非贝叶斯方法解决这些问题。信息状态一词来自有关部分可观察的 MDP 的文献。参见, 例如 Lovejoy (1991)。

其他理论研究则集中在探索的效率上, 通常表示为一种算法能够以多快的速度处理最佳决策策略。形式化探索效率的一种方式, 是通过适应性学习来监督学习算法的样本复杂性, 这是算法在学习目标函数时需要达到所需准确度的训练示例数量。强化学习算法的探索样本复杂度的定义是时间步数, 在该时间步中算法没有选择接近最优的动作 (Kakade, 2003 年)。Li (2012) 在对强化学习中探索科学的理论方法的调查中讨论了这种方法和和其他几种方法。Russo, VanRoy, Kazerouni, Osband 和 Wen (2018) 提供了汤普森采样的现代方法。



## 第三章 有限马尔可夫决策过程



## 特别声明

---

在过去的 2019 年，Elegant $\text{\LaTeX}$  系列模板均逐步上线 GitHub、CTAN、Overleaf 以及 Gitee 上。截止到 2019 年底，ElegantNote、ElegantBook、ElegantPaper 三个模板在 GitHub 上的收藏数达到了 194、333 和 220，从 2019 年 5 月开启捐赠之后收到了用户 33 笔合计超过 1500 元的捐赠，用户群人数也超过了 400 人。这些数字的背后，反映出 Elegant $\text{\LaTeX}$  越来越受用户的喜爱，在此非常感谢大家。

但是，我想声明的是：

由于某些原因，Elegant $\text{\LaTeX}$  项目 不再接受 任何非我本人预约的提交。

我是一个理想主义者，关于这个模板，我有自己的想法。我所关心的是，我周围的人能方便使用  $\text{\LaTeX}$  以及此模板，我自己会为自己的东西感到开心。如果维护模板让我不开心，那我就不会再维护了。诚然这个模板并不是完美的，但是相比 2.x 好很多了，这些改进离不开大家的反馈、China $\text{\TeX}$  和逐鹿人的鼓励以及支援人员的帮助！

如果你无法认同我的想法，建议直接删除本模板。

Ethan Deng

February 10, 2020

## 第四章 Elegant $\text{\LaTeX}$ 系列模板介绍

Elegant $\text{\LaTeX}$  项目组致力于打造一系列美观、优雅、简便的模板方便用户使用。目前由 **ElegantNote**, **ElegantBook**, **ElegantPaper** 组成, 分别用于排版笔记, 书籍和工作论文。强烈推荐使用最新正式版本! 本文将介绍本模板的一些设置内容以及基本使用方法。如果您有其他问题, 建议或者意见, 欢迎在 **GitHub** 上给我们提交 **issues** 或者邮件联系我们。

我们的联系方式如下, 建议加入用户 QQ 群提问, 这样能更快获得准确的反馈, 加群时请备注  $\text{\LaTeX}$  或者 Elegant $\text{\LaTeX}$  相关内容。

- 官网: <https://elegantlatex.org/>
- GitHub 网址: <https://github.com/ElegantLaTeX/>
- CTAN 地址: <https://ctan.org/pkg/elegantbook>
- 文档 Wiki: <https://github.com/ElegantLaTeX/ElegantBook/wiki>
- 下载地址: 正式发行版, 最新版
- 微博: Elegant $\text{\LaTeX}$
- 微信公众号: Elegant $\text{\LaTeX}$
- 用户 QQ 群: 692108391
- 邮件: [elegantlatex2e@gmail.com](mailto:elegantlatex2e@gmail.com)

### 4.1 ElegantBook 更新说明

此次更新主要有

1. 增加数学字体选项 **math**, 可选项为 **newtx** 和 **cm**。  
**重要提示:** 原先通过 **newtxmath** 宏包设置的数学字体改为  $\text{\LaTeX}$  默认数学字体, 如果保持原来的字体, 需要显式声明数学字体 (**math=newtx**);
2. 新增中文字体选项 **chinesefont**, 可选项为 **ctexfont**、**founder** 和 **nofont**。
3. 将封面作者信息设置为可选, 并且增加自定义信息命令 **\bioinfo**;
4. 在说明文档中增加版本历史, 新增 **\datechange** 命令和 **change** 环境;
5. 增加汉化章节选项 **scheme**, 可选项为汉化 **chinese**;
6. 由于 **\lvert** 问题已经修复, 重新调整 **ctex** 宏包和 **amsmath** 宏包位置。
7. 修改页眉设置, 去除了 **\lastpage** 以避免 page anchor 问题, 加入 **\frontmatter**。
8. 修改参考文献选项 **cite**, 可选项为数字 **number**、作者-年份 **authoryear** 以及上标 **super**。
9. 新增参考文献样式选项 **bibstyle**, 并将英文模式下参考文献样式 **apalike** 设置为默认值, 中文仍然使用 **gbt7714** 宏包设置。



**笔记** 如果你之前使用了本模板, 在使用新版本时, 需要删除文档中的 **\hypersetup{pageanchor=true}**, 并且在 **\maketitle** 和 **\tableofcontents** 之间添加 **\frontmatter**。

2.x 版本的用户请仔细查看[跨版本转换](#)。

## 4.2 模板安装与更新

你可以通过免安装的方式使用本模板，包括在线使用和本地（文件夹内）使用两种方式，也可以通过  $\text{\TeX}$  发行版安装使用。

### 4.2.1 在线使用模板

我们把三套模板全部上传到 [Overleaf](#) 上了，网络便利的用户可以直接通过 [Overleaf](#) 在线使用我们的模板。使用 [Overleaf](#) 的好处是无需安装  $\text{\TeX}$  Live 2019，可以随时随地访问自己的文件。查找模板，请在 [Overleaf](#) 模板库里面搜索 `elegantlatex` 即可，你也可以直接访问[搜索结果](#)。选择适当的模板之后，将其 `Open as Template`，即可把模板存到自己账户下，然后可以自由编辑以及和别人一起协作。更多关于 [Overleaf](#) 的介绍和使用，请参考 [Overleaf](#) 的[官方文档](#)。

**注** [Overleaf](#) 上，中文需要使用  $\text{\XeLaTeX}$  进行编译，英文建议使用  $\text{\pdfLaTeX}$  编译。

### 4.2.2 本地免安装使用

免安装使用方法如下，从 [GitHub](#) 或者 [CTAN](#) 下载最新（正式）版文件，严格意义上只需要类文件 `elegantbook.cls`。然后将模板文件放在你的工作目录下即可使用。这样使用的好处是，无需安装，简便；缺点是，当模板更新之后，你需要手动替换 `cls` 文件。

### 4.2.3 发行版安装使用

如果你是  $\text{\TeX}$  Live 2019 用户，我们推荐你直接进行安装和更新。你可以通过  $\text{\TeX}$  Live 2019 自带的 `tlshel1`<sup>1</sup> 进行安装。安装非常简单，步骤如下，搜索并打开 `tlshel1`，然后通过 `File -> Load Default Repository` 加载远程仓库，如果你不想使用默认的仓库，你可以通过 `Options` 下的菜单设置远程仓库。设置好仓库之后，等待仓库加载完毕，你可以在下面的搜索栏搜索 `elegantbook`，然后选择进行安装与更新。

### 4.2.4 更新问题

如果使用 `tlshel1` 无法更新模板，请使用命令行全部更新全部宏包或者使用免安装的方法使用本模板。

通过命令行（管理员权限）输入下面的命令对 `tlmgr` 自身和全部宏包进行更新。

```
tlmgr update --self
tlmgr update --all
```

更多的内容请参考 [How do I update my  \$\text{\TeX}\$  distribution?](#)

---

<sup>1</sup>也叫  $\text{\TeX}$  Live Manager

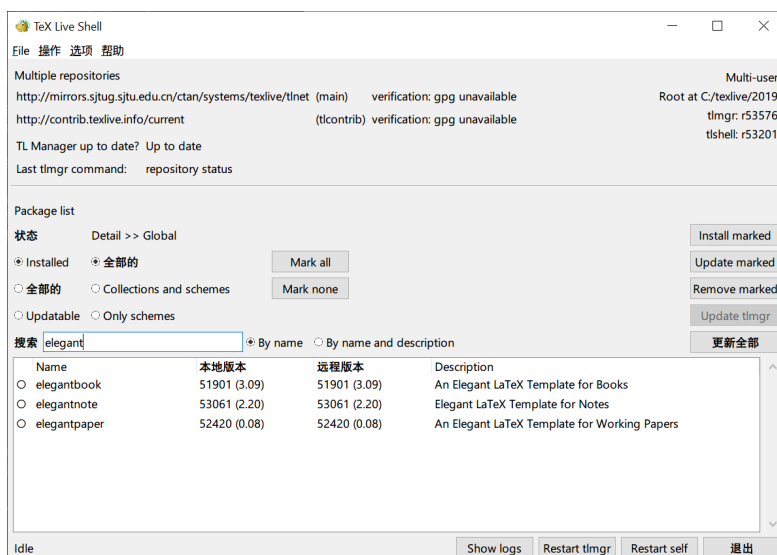


图 4.1: 使用 TeX Live Shell 安装 ElegantBook 模板

### 4.2.5 其他发行版本

如果你是 TeX Live 2018 的用户，由于 2018 难以更新到 2019，建议卸载 2018 重装 2019。如果嫌麻烦，你可以手动安装模板，将 `elegantbook.cls` 复制到你的 TeX Live 目录下，默认安装目录为 `C:\texlive\2019\texmf-dist\tex\latex\elegantbook`，然后通过命令行（管理员权限），运行 `texhash` 即可。

由于宏包版本问题，本模板不支持 CT<sub>EX</sub> 套装。更多关于 TeX Live 2019 的安装使用以及 CT<sub>EX</sub> 与 TeX Live 的兼容、系统路径问题，请参考官方文档以及嘯行的一份简短的 [安装 L<sup>A</sup>T<sub>E</sub>X 的介绍](#)。

## 4.3 用户作品计划

ElegantL<sup>A</sup>T<sub>E</sub>X 系列模板从创立至今已经有 9 年了，我们的模板也受到了很多用户的喜爱，在此，为了促进模板用户之间的交流，了解用户需求，完善本模板，我们将建立一个区域专门展示用户的文档，包括但不限于 GitHub 和官网等。如果你愿意将自己的作品展示出来，请邮件或者其他方式联系我们。如果自己代码已经传到 GitHub 或者 Gitee 等网站，可以提供对应网址。[用户文档中心](#)目前有下面一些作品：

1. 唐绍东：微积分笔记
2. 曲豆豆：超甜微积分习题集
3. 王世强：《化工数值计算与 MATLAB》复习指南
4. 李晨迪：Fluid Mechanics Notes
5. 肖明顺：地球物理勘查 常用规范汇编
6. 白衣卿相：期末复习笔记-拓扑学摘要



## 4.4 关于提交

出于某些因素的考虑，Elegant $\text{\LaTeX}$  项目自 2019 年 5 月 20 日开始，**不再接受任何非作者预约性质的提交** (pull request)! 如果你想改进模板，你可以给我们提交 issues，或者可以在遵循协议 (LPPL-1.3c) 的情况下，克隆到自己仓库下进行修改。

## 4.5 协作人员招募

招募 Elegant $\text{\LaTeX}$  的协作人员 (志愿者)，没有报酬。工作内容：翻译 Elegant $\text{\LaTeX}$  系列模板文档，维护模板的维基，如果有公众号文稿写作经历的话，也可以帮忙写微信稿。本公告长期有效。目前 Elegant $\text{\LaTeX}$  共有 4 名协作人员，在此感谢他们无私的奉献!

- 官方文档翻译: [YPY](#);
- GitHub 维基维护: [Ingo Zinningo](#)、[追寻原风景](#);
- QQ 群管理、FAQ 整理: [Sikouhju](#).

另外，也感谢 [Xiangdong Zeng](#)、[逐鹿人](#) 等人帮忙群管理。

## 4.6 致谢

2019 年 5 月 20 日，ElegantBook 模板在 GitHub 上的收藏数 (star) 达到了 100<sup>2</sup>。在此特别感谢 China $\text{\TeX}$  以及  $\text{\LaTeX}$  工作室对于本系列模板的大力宣传与推广。 $\text{\LaTeX}$  工作室网站上有很多精彩的帖子和精致的模板，欢迎大家去挖掘里面的宝藏，这也是国内最全面的  $\text{\LaTeX}$  相关的网站。

如果你喜欢我们的模板，你可以在 GitHub 上收藏我们的模板。

## 4.7 捐赠

如果您非常喜爱我们的模板或者我，你还可以选择捐赠<sup>3</sup>以表达您对我们模板和我的支持。本模板自 3.08 版本发布了捐赠信息之后，收到了超过 1500 元的捐赠 (四舍五入就是一个亿)，非常感谢!



微信



支付宝

---

<sup>2</sup>截止 3.10 版本正式发布，star 数为 374。

<sup>3</sup>最好在捐赠时备注信息。

赞赏费用的使用解释权归 **Elegant $\text{\LaTeX}$**  所有，并且不接受监督，请自愿理性打赏。  
10 元以上的赞赏，我们将列入捐赠榜，谢谢各位金主！

表 4.1: Elegant $\text{\LaTeX}$  系列模板捐赠榜

捐赠者	金额	时间	渠道	捐赠者	金额	时间	渠道
Lerh	10 RMB	2019/05/15	微信	越过地平线	10 RMB	2019/05/15	微信
银桑	20 RMB	2019/05/27	微信	* 空	10 RMB	2019/05/30	微信
latexstudio.net	666 RMB	2019/06/05	支付宝	A*n	40 RMB	2019/06/15	微信
* 夏	22 RMB	2019/06/15	微信	* 倩	21 RMB	2019/06/15	微信
Cassis	11 RMB	2019/06/30	微信	* 君	10 RMB	2019/07/23	微信
P*u	50 RMB	2019/07/30	微信	* 萌	19 RMB	2019/08/28	微信
曲豆豆	10 RMB	2019/08/28	微信	李博	100 RMB	2019/10/06	微信
Njustsl	10 RMB	2019/10/11	微信	刘志阔	99.99 RMB	2019/10/15	支付宝
* 韬	16 RMB	2019/10/17	微信	赤霓	12 RMB	2019/10/17	支付宝
追寻原风景	10 RMB	2019/10/28	微信	郭德良	88 RMB	2019/11/03	微信
自强不息	20 RMB	2019/11/04	支付宝	读书之虫	20 RMB	2019/11/18	微信
* 等	10 RMB	2019/11/18	微信	* 哲	20 RMB	2019/11/18	微信
佚名	10 RMB	2019/11/24	微信	Jiye Qian	66 RMB	2019/12/04	微信
* 阳	20 RMB	2019/12/05	微信	Catcher	11 RMB	2019/12/08	支付宝
希尔波特门徒	10 RMB	2019/12/09	支付宝	* 伟	10 RMB	2019/12/09	微信
Simon	20 RMB	2019/12/11	支付宝	流殇、浅忆	66.60 RMB	2019/12/18	支付宝
羽	10 RMB	2019/12/20	支付宝	* 琛	15 RMB	2019/12/20	微信
随风	20 RMB	2019/12/27	支付宝	Ws	23.30 RMB	2019/12/28	微信
初八	100 RMB	2020/01/02	支付宝	p*e	20 RMB	2020/01/03	微信
Shunmx	100 RMB	2020/01/03	微信	hj	10 RMB	2020/01/03	微信
F*5	10 RMB	2020/01/03	微信	S*m	20.20 RMB	2020/01/03	微信
二代青雉	13 RMB	2020/01/14	支付宝	*?	66 RMB	2020/01/15	微信
Mr. Xiong	20 RMB	2020/01/17	微信	* 博	15 RMB	2020/01/18	微信
* 者	10 RMB	2020/02/02	微信	Jackie	88.80 RMB	2020/02/09	微信

另外，为了表示感谢，我们制作了捐赠纪念证，欢迎大家来信告知邮箱以及姓名（艺名），我们将通过邮件发送电子版纪念证。



## 第五章 ElegantBook 设置说明

本模板基于基础的 `book` 文类，所以 `book` 的选项对于本模板也是有效的（纸张无效，因为模板有设备选项）。默认编码为 UTF-8，推荐使用  $\text{T}_{\text{E}}\text{X}$  Live 编译。本文编写环境为 Win10 (64bit) +  $\text{T}_{\text{E}}\text{X}$  Live 2019，支持  $\text{pdfL}_{\text{A}}\text{T}_{\text{E}}\text{X}$  以及  $\text{X}_{\text{L}}\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$  编译。中文请尽量使用  $\text{X}_{\text{L}}\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$  编译。

### 5.1 语言模式

本模板内含两套语言环境，改变语言环境会改变图表标题的引导词（图，表），文章结构词（比如目录，参考文献等），以及定理环境中的引导词（比如定理，引理等）。不同语言模式的启用如下：

```
\documentclass[cn]{elegantbook}
\documentclass[lang=cn]{elegantbook}
```

**注** 只有中文环境（`lang=cn`）才可以输入中文。另外如果抄录环境（`lstlisting`）中有中文字符，请务必使用  $\text{X}_{\text{L}}\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$  编译。

### 5.2 设备选项

最早我们在 `ElegantNote` 模板中加入了设备选项（`device`），后来，我们觉得这个设备选项的设置可以应用到 `ElegantBook` 中<sup>1</sup>，而且 `Book` 一般内容比较多，如果在 `iPad` 上看无需切边，放大，那用户的阅读体验将会得到巨大提升。你可以使用下面的选项将版面设置为 `iPad` 设备模式<sup>2</sup>

```
\documentclass[pad]{elegantbook} %or
\documentclass[device=pad]{elegantbook}
```

### 5.3 颜色主题

本模板内置 5 组颜色主题，分别为 `green`<sup>3</sup>、`cyan`、`blue`（默认）、`gray`、`black`。另外还有一个自定义的选项 `nocolor`。调用颜色主题 `green` 的方法为





















```
\documentclass[green]{elegantbook} %or
\documentclass[color=green]{elegantbook}
```

<sup>1</sup>不过因为 `ElegantBook` 模板封面图片的存在，在修改页面设计时，需要对图片进行裁剪。

<sup>2</sup>默认为 `normal` 模式，也即 A4 纸张大小。

<sup>3</sup>为原先默认主题。

表 5.1: ElegantBook 模板中的颜色主题

	green	cyan	blue	gray	black	主要使用的环境
structure						chapter section subsection
main						definition exercise problem
second						theorem lemma corollary
third						proposition

如果需要自定义颜色的话请选择 `nocolor` 选项或者使用 `color=none`，然后在导言区定义 `structurecolor`、`main`、`second`、`third` 颜色，具体方法如下：

```
\definecolor{structurecolor}{RGB}{0,0,0}
\definecolor{main}{RGB}{70,70,70}
\definecolor{second}{RGB}{115,45,2}
\definecolor{third}{RGB}{0,80,80}
```

5.4 封面

5.4.1 封面个性化

从 3.10 版本开始，封面更加弹性化，用户可以自行选择输出的内容，包括 `\title` 在内的所有封面元素都可为空。目前封面的元素有

- 标题： `\title`
- 副标题： `\subtitle`
- 作者： `\author`
- 机构： `\institute`
- 日期： `\date`
- 版本： `\version`
- 箴言： `\extrainfo`
- 封面图： `\cover`
- 徽标： `\logo`

另外，额外增加一个 `\bioinfo` 命令，有两个选项，分别是信息标题以及信息内容。比如需要显示 `User Name: 111520`，则可以使用

```
\bioinfo{User Name}{111520}
```



### 5.4.2 封面图

本模板使用的封面图片来源于 [pixabay.com](https://pixabay.com)<sup>4</sup>, 图片完全免费, 可用于任何场景。封面图片的尺寸为  $1280 \times 1024$ , 更换图片的时候请严格按照封面图片尺寸进行裁剪。推荐一个免费的在线图片裁剪网站 [fotor.com](https://fotor.com)。用户 QQ 群内有一些合适尺寸的封面, 欢迎取用。

### 5.4.3 徽标

本文用到的 Logo 比例为 1:1, 也即正方形图片, 在更换图片的时候请选择合适的图片进行替换。

### 5.4.4 自定义封面

另外, 如果需要使用自定义的封面, 比如 Adobe illustrator 或者其他软件制作的 A4 PDF 文档, 请把 `\maketitle` 注释掉, 然后借助 `pdfpages` 宏包将自制封面插入即可。如果使用 `titlepage` 环境, 也是类似。如果需要 2.x 版本的封面, 请参考 `etitlepage`。

## 5.5 章标标题

本模板内置 2 套章标题显示风格, 包含 `hang` (默认) 与 `display` 两种风格, 区别在于章标题单行显示 (`hang`) 与双行显示 (`display`), 本说明使用了 `hang`。调用方式为

```
\documentclass[hang]{elegantbook} %or
\documentclass[titlestyle=hang]{elegantbook}
```

在章标题内, 章节编号默认是以数字显示, 也即第 1 章, 第 2 章等等, 如果想要把数字改为中文, 可以使用

```
\documentclass[chinese]{elegantbook} %or
\documentclass[scheme=chinese]{elegantbook}
```

## 5.6 数学环境简介

在我们这个模板中, 我们定义了两种不同的定理模式 `mode`, 包括简单模式 (`simple`) 和炫彩模式 (`fancy`), 默认为 `fancy` 模式, 不同模式的选择为

```
\documentclass[simple]{elegantbook} %or
\documentclass[mode=simple]{elegantbook}
```

本模板定义了四大类环境

- 定理类环境, 包含标题和内容两部分, 全部定理类环境的编号均以章节编号。根据格式的不同分为 3 种

<sup>4</sup>感谢 ChinaTeX 提供免费图源网站, 另外还推荐 [pexels.com](https://pexels.com)。



- **definition** 环境，颜色为 **main**；
- **theorem**、**lemma**、**corollary** 环境，颜色为 **second**；
- **proposition** 环境，颜色为 **third**。
- 示例类环境，有 **example**、**problem**、**exercise** 环境（对应于例、例题、练习），自动编号，编号以章节为单位，其中 **exercise** 有提示符。
- 提示类环境，有 **note** 环境，特点是：无编号，有引导符。
- 结论类环境，有 **conclusion**、**assumption**、**property**、**remark**、**solution** 环境<sup>5</sup>，三者均以粗体的引导词为开头，和普通段落格式一致。

### 5.6.1 定理类环境的使用

由于本模板使用了 **tcolorbox** 宏包来定制定理类环境，所以和普通的定理环境的使用有些许区别，定理的使用方法如下：

```
\begin{theorem}{theorem name}{label}
The content of theorem.
\end{theorem}
```

第一个必选项 **theorem name** 是定理的名字，第二个必选项 **label** 是交叉引用时所用到的标签，交叉引用的方法为 `\ref{thm:label}`。请注意，交叉引用时必须加上前缀 **thm:**。

其他相同用法的定理类环境有：

表 5.2: 定理类环境

环境名	标签名	前缀	交叉引用
definition	label	def	<code>\ref{def:label}</code>
theorem	label	thm	<code>\ref{thm:label}</code>
lemma	label	lem	<code>\ref{lem:label}</code>
corollary	label	cor	<code>\ref{cor:label}</code>
proposition	label	pro	<code>\ref{pro:label}</code>

### 5.6.2 其他环境的使用

其他三种环境没有选项，可以直接使用，比如 **example** 环境的使用方法与效果：

```
\begin{example}
This is the content of example environment.
\end{example}
```

这几个都是同一类环境，区别在于

- 示例环境（**example**）、练习（**exercise**）与例题（**problem**）章节自动编号；
- 注意（**note**），练习（**exercise**）环境有提醒引导符；

<sup>5</sup>本模板还添加了一个 **result** 选项，用于隐藏 **solution** 和 **proof** 环境，默认为显示（**result=answer**），隐藏使用 **result=noanswer**。

- 结论 (conclusion) 等环境都是普通段落环境，引导词加粗。

## 5.7 装饰物

本模板为章节后和页面下方的装饰物 (base) 添加了隐藏选项，有 `show` (默认) 和 `hide` 两个选项。

```
\documentclass[hide]{elegantbook} %or
\documentclass[base=hide]{elegantbook}
```

## 5.8 列表环境

本模板借助于 `tikz` 定制了 `itemize` 和 `enumerate` 环境，其中 `itemize` 环境修改了 3 层嵌套，而 `enumerate` 环境修改了 4 层嵌套 (仅改变颜色)。示例如下

- |                           |                             |
|---------------------------|-----------------------------|
| • first item of nesti;    | 1. first item of nesti;     |
| • second item of nesti;   | 2. second item of nesti;    |
| • first item of nestii;   | (a). first item of nestii;  |
| • second item of nestii;  | (b). second item of nestii; |
| • first item of nestiii;  | I. first item of nestiii;   |
| • second item of nestiii. | II. second item of nestiii. |

## 5.9 参考文献

此模板使用了 `BibTeX` 来生成参考文献，在中文示例中，使用了 `gbt7714` 宏包。参考文献示例：[1-3] 使用了中国一个大型的 P2P 平台 (人人贷) 的数据来检验男性投资者和女性投资者在投资表现上是否有显著差异。

你可以在谷歌学术，Mendeley，Endnote 中获得文献条目 (bib item)，然后把它们添加到 `reference.bib` 中。在文中引用的时候，引用它们的键值 (bib key) 即可。注意需要在编译的过程中添加 `BibTeX` 编译。如果你想添加未引用的文献，可以使用

```
\nocite{EINAV2010,Havrylchuk2018} %or include some bibitems
\nocite{*} %include all the bibitems
```

本模板还添加了 `cite=numbers`、`cite=super` 和 `cite=authoryear` 三个参考文献选项，用于设置参考文献格式的设置，默认为 `numbers`。据我们所知，理工科类一般使用数字形式 `numbers` 或者上标形式 `super`，而文科类使用作者-年份 `authoryear` 比较多，所以我们将 `numbers` 作为默认格式。如果需要改为 `cite=super` 或者 `authoryear`，可以使用

```
\documentclass[cite=super]{elegantbook} % set super style ref style
\documentclass[super]{elegantbook}
\documentclass[cite=authoryear]{elegantbook} % set author year ref style
```

```
\documentclass[authoryear]{elegantbook}
```

为了方便文献样式修改，模板引入了 `bibstyle` 选项，默认为 `apalike`，更多的选择可以参考 [BibTeX Bibliography Styles](#)。用法为

```
\documentclass[bibstyle=apalike]{elegantbook}
```

## 5.10 添加序章

如果你想在第一章前面添序章，不改变原本章节序号，可以在第一章内容前面使用

```
\chapter*{Introduction}
\markboth{Introduction}{Introduction}
The content of introduction.
```

## 5.11 目录选项与深度

本模板添加了一个目录选项 `toc`，可以设置目录为单栏 (`onecol`) 和双栏 (`twocol`) 显示，比如双栏显示可以使用

```
\documentclass[twocol]{elegantbook}
\documentclass[toc=twocol]{elegantbook}
```

默认本模板目录深度为 1，你可以在导言区使用

```
\setcounter{tocdepth}{2}
```

将其修改为 2 级目录（章与节）显示。

## 5.12 章节摘要

模板新增了一个章节摘要环境 (`introduction`)，使用示例

```
\begin{introduction}
  \item Definition of Theorem
  \item Ask for help
  \item Optimization Problem
  \item Property of Cauchy Series
  \item Angle of Corner
\end{introduction}
```

效果如下：



## 内容提要

- ☐ Definition of Theorem
- ☐ Ask for help
- ☐ Optimization Problem
- ☐ Property of Cauchy Series
- ☐ Angle of Corner

环境的标题文字可以通过这个环境的可选参数进行修改，修改方法为：

```
\begin{introduction}[Brief Introduction]
...
\end{introduction}
```

## 5.13 章后习题

前面我们介绍了例题和练习两个环境，这里我们再加一个，章后习题（`problemset`）环境，用于在每一章结尾，显示本章的练习。使用方法如下

```
\begin{problemset}
  \item exercise 1
  \item exercise 2
  \item exercise 3
\end{problemset}
```

效果如下：

### 第五章习题

1. exercise 1
2. exercise 2
3. exercise 3
4. 测试数学公式

$$a^2 + b^2 = c_{2i}(1, 2)[1, 23] \quad (5.1)$$

**注** 如果你想把 `problemset` 环境的标题改为其他文字，你可以类似于 `introduction` 环境修改 `problemset` 的可选参数。另外，目前这个环境会自动出现在目录中，但是不会出现在页眉页脚信息中（待解决）。

**解** 如果你想把 `problemset` 环境的标题改为其他文字，你可以类似于 `introduction` 环境修改 `problemset` 的可选参数。另外，目前这个环境会自动出现在目录中，但是不会出现在页眉页脚信息中（待解决）。

## 5.14 旁注

在 3.08 版本中，我们引入了旁注设置选项 `marginpar=margintrue` 以及测试命令 `\elegantpar`，但是由此带来一堆问题。我们决定在 3.09 版本中将其删除，并且，在旁注

命令得到大幅度优化之前, 不会将此命令再次引入书籍模板中。对此造成各位用户的不方便, 非常抱歉! 不过我们保留了 `marginpar` 这个选项, 你可以使用 `marginpar=margintrue` 获得保留右侧旁注的版面设计。然后使用系统自带的 `\marginpar` 或者 `marginnote` 宏包的 `\marginnote` 命令。

**注** 在使用旁注的时候, 需要注意的是, 文本和公式可以直接在旁注中使用。

```
% text
\marginpar{margin paragraph text}

% equation
\marginpar{
  \begin{equation}
    a^2 + b^2 = c^2
  \end{equation}
}
```

但是浮动体 (表格、图片) 需要注意, 不能用浮动体环境, 需要使用直接插图命令或者表格命令环境。然后使用 `\captionof` 为其设置标题。为了得到居中的图表, 可以使用 `\centerline` 命令或者 `center` 环境。更多详情请参考: [Caption of Figure in Marginpar](#)。

```
% graph with centerline command
\marginpar{
  \centerline{
    \includegraphics[width=0.2\textwidth]{logo.png}
  }
  \captionof{figure}{your figure caption}
}

% graph with center environment
\marginpar{
  \begin{center}
    \includegraphics[width=0.2\textwidth]{logo.png}
    \captionof{figure}{your figure caption}
  \end{center}
}
```



## 第六章 字体选项

字体选项独立成章的原因是，我们希望本模板的用户关心模板使用的字体，知晓自己使用的字体以及遇到字体相关的问题能更加便捷地找到答案。

**重要提示：**从 3.10 版本更新之后，沿用至今的 `newtx` 系列字体被重新更改为 `cm` 字体。并且新增中文字体（`chinesefont`）选项。

### 6.1 数学字体选项

本模板定义了一个数学字体选项（`math`），可选项有三个：

1. `math=cm`（默认），使用 `LATEX` 默认数学字体（推荐，无需声明）；
2. `math=newtx`，使用 `newtxmath` 设置数学字体（潜在问题比较多）。
3. `math=mtpro2`，使用 `mtpro2` 宏包设置数学字体，要求用户已经成功安装此宏包。

### 6.2 使用 `newtx` 系列字体

如果需要使用原先版本的 `newtx` 系列字体，可以通过显示声明数学字体：

```
\documentclass[math=newtx]{elegantbook}
```

#### 6.2.1 连字符

如果使用 `newtx` 系列字体宏包，需要注意下连字符的问题。

$$\int_{R^q} f(x,y)dy.off \quad (6.1)$$

的代码为

```
\begin{equation}
\int_{R^q} f(x,y) dy.\emph{of \kern0pt f}
\end{equation}
```

#### 6.2.2 宏包冲突

另外在 3.08 版本中，有用户反馈模板在和 `yhmath` 以及 `esvect` 等宏包搭配使用的时候会出现报错：

```
LaTeX Error:
Too many symbol fonts declared.
```

原因是使用 `newtxmath` 宏包时，重新定义了数学字体用于大型操作符，达到了最多 16 个数学字体的上限，在调用其他宏包的时候，无法新增数学字体。为了减少调用非常用宏包，在此给出如何调用 `yhm` 以及 `esvect` 宏包的方法。

请在 `elegantbook.cls` 内搜索 `yhm` 或者 `esvect`，将你所需要的宏包加载语句取消注释即可。

```
%%% use yhm pkg, uncomment following code
% \let\oldwidering\widering
% \let\widering\undefined
% \RequirePackage{yhm}
% \let\widering\oldwidering

%%% use esvect pkg, uncomment following code
% \RequirePackage{esvect}
```

## 6.3 中文字体选项

模板从 3.10 版本提供中文字体选项 `chinesefont`，可选项有

1. `ctexfont`：默认选项，使用 `ctex` 宏包根据系统自行选择字体，可能存在字体缺失的问题，更多内容参考 `ctex` 宏包[官方文档](#)<sup>1</sup>。
2. `founder`：方正字体选项，调用 `ctex` 宏包并且使用 `fontset=none` 选项，然后设置字体为方正四款免费字体，方正字体下载注意事项见后文。
3. `nofont`：调用 `ctex` 宏包并且使用 `fontset=none` 选项，不设定中文字体，用户可以自行设置中文字体，具体见后文。

**注** 使用 `founder` 选项或者 `nofont` 时，必须使用  $\text{\LaTeX}$  进行编译。

### 6.3.1 方正字体选项

由于使用 `ctex` 宏包默认调用系统已有的字体，部分系统字体缺失严重，因此，用户希望能够使用其它字体，我们推荐使用方正字体。方正的方正书宋、方正黑体、方正楷体、方正仿宋四款字体均可免费试用，且可用于商业用途。

用户可以自行从[方正字体官网](#)下载此四款字体，在下载的时候请务必注意选择 GBK 字符集

也可以使用  $\text{\LaTeX}$  工作室提供的方正字体，提取码为：njy9 进行安装。安装时，Win 10 用户请右键选择为全部用户安装，否则会找不到字体。

### 6.3.2 其他中文字体

如果你想完全自定义字体<sup>2</sup>，你可以选择 `chinesefont=nofont`，然后在导言区设置

<sup>1</sup>可以使用命令提示符，输入 `texdoc ctex` 调出本地 `ctex` 宏包文档

<sup>2</sup>这里仍然以方正字体为例。

全部字体订单

待付款

已完成

字体/订单号

搜索

如果订单中包含方正黑体、方正书宋、方正仿宋、方正楷体这四款字体，针对“商业发布”使用方式免费，其它字体仅用于“个人非商业”使用

字体名称	编码	单价	实付价	交易状态	操作
订单号：C20200204164821OW1F			2020-02-04 16:48:21		
方正仿宋_GBK <div>免费商用</div>	简繁扩展(GBK)	¥ 0.00	免费	已完成	<div>下载字体</div>
方正黑体_GBK <div>免费商用</div>	简繁扩展(GBK)	¥ 0.00			
方正书宋_GBK <div>免费商用</div>	简繁扩展(GBK)	¥ 0.00			
方正楷体_GBK <div>免费商用</div>	简繁扩展(GBK)	¥ 0.00			

```

\setCJKmainfont [BoldFont={FZHei-B01},ItalicFont={FZKai-Z03}]{FZShuSong-Z01}
\setCJKsansfont [BoldFont={FZHei-B01},ItalicFont={FZHei-B01}]{FZHei-B01}
\setCJKmonofont [BoldFont={FZHei-B01},ItalicFont={FZHei-B01}]{FZFangSong-Z02}
\setCJKfamilyfont{zh song}{FZShuSong-Z01}
\setCJKfamilyfont{zh hei}{FZHei-B01}
\setCJKfamilyfont{zh kai}{FZKai-Z03}
\setCJKfamilyfont{zh fs}{FZFangSong-Z02}
\newcommand*{\songti}{\CJKfamily{zh song}}
\newcommand*{\heiti}{\CJKfamily{zh hei}}
\newcommand*{\kaishu}{\CJKfamily{zh kai}}
\newcommand*{\fangsong}{\CJKfamily{zh fs}}

```

## 第七章 ElegantBook 写作示例

### 内容提要

- 积分定义 7.1
- Fubini 定理 7.1
- 最优性原理 7.1
- 柯西列性质 7.1.1
- 韦达定理

### 7.1 Lebesgue 积分

在前面各章做了必要的准备后，本章开始介绍新的积分。在 Lebesgue 测度理论的基础上建立了 Lebesgue 积分，其被积函数和积分域更一般，可以对有界函数和无界函数统一处理。正是由于 Lebesgue 积分的这些特点，使得 Lebesgue 积分比 Riemann 积分具有在更一般条件下的极限定理和累次积分交换积分顺序的定理，这使得 Lebesgue 积分不仅在理论上更完善，而且在计算上更灵活有效。

Lebesgue 积分有几种不同的定义方式。我们将采用逐步定义非负简单函数，非负可测函数和一般可测函数积分的方式。

由于现代数学的许多分支如概率论、泛函分析、调和分析等常常用到一般空间上的测度与积分理论，在本章最后一节将介绍一般的测度空间上的积分。

#### 7.1.1 积分的定义

我们将通过三个步骤定义可测函数的积分。首先定义非负简单函数的积分。以下设  $E$  是  $\mathcal{R}^n$  中的可测集。

##### 定义 7.1. 可积性

设  $f(x) = \sum_{i=1}^k a_i \chi_{A_i}(x)$  是  $E$  上的非负简单函数，其中  $\{A_1, A_2, \dots, A_k\}$  是  $E$  上的一个可测分割， $a_1, a_2, \dots, a_k$  是非负实数。定义  $f$  在  $E$  上的积分为  $\int_a^b f(x)$

$$\int_E f dx = \sum_{i=1}^k a_i m(A_i) \pi \alpha \beta \sigma \gamma \nu \xi \epsilon \epsilon. \int_a^b \oint_a^b \prod_{i=1}^n \quad (7.1)$$

一般情况下  $0 \leq \int_E f dx \leq \infty$ 。若  $\int_E f dx < \infty$ ，则称  $f$  在  $E$  上可积。




一个自然的问题是，Lebesgue 积分与我们所熟悉的 Riemann 积分有什么联系和区别？在 4.4 在我们将详细讨论 Riemann 积分与 Lebesgue 积分的关系。这里只看一个简单的例子。设  $D(x)$  是区间  $[0, 1]$  上的 Dirichlet 函数。即  $D(x) = \chi_{Q_0}(x)$ ，其中  $Q_0$  表示  $[0, 1]$  中

的有理数的全体。根据非负简单函数积分的定义,  $D(x)$  在  $[0, 1]$  上的 Lebesgue 积分为

$$\int_0^1 D(x)dx = \int_0^1 \chi_{Q_0}(x)dx = m(Q_0) = 0 \quad (7.2)$$

即  $D(x)$  在  $[0, 1]$  上是 Lebesgue 可积的并且积分值为零。但  $D(x)$  在  $[0, 1]$  上不是 Riemann 可积的。

有界变差函数是与单调函数有密切联系的一类函数。有界变差函数可以表示为两个单调递增函数之差。与单调函数一样, 有界变差函数几乎处处可导。与单调函数不同, 有界变差函数类对线性运算是封闭的, 它们构成一线空间。练习题 7.1 是一个性质的证明。

 **练习 7.1** 设  $f \notin L(\mathcal{R}^1)$ ,  $g$  是  $\mathcal{R}^1$  上的有界可测函数。证明函数

$$I(t) = \int_{\mathcal{R}^1} f(x+t)g(x)dx \quad t \in \mathcal{R}^1 \quad (7.3)$$

是  $\mathcal{R}^1$  上的连续函数。

**解** 即  $D(x)$  在  $[0, 1]$  上是 Lebesgue 可积的并且积分值为零。但  $D(x)$  在  $[0, 1]$  上不是 Riemann 可积的。

**证明** 即  $D(x)$  在  $[0, 1]$  上是 Lebesgue 可积的并且积分值为零。但  $D(x)$  在  $[0, 1]$  上不是 Riemann 可积的。

#### 定理 7.1. Fubini 定理

(1) 若  $f(x, y)$  是  $\mathcal{R}^p \times \mathcal{R}^q$  上的非负可测函数, 则对几乎处处的  $x \in \mathcal{R}^p$ ,  $f(x, y)$  作为  $y$  的函数是  $\mathcal{R}^q$  上的非负可测函数,  $g(x) = \int_{\mathcal{R}^q} f(x, y)dy$  是  $\mathcal{R}^p$  上的非负可测函数。并且

$$\int_{\mathcal{R}^p \times \mathcal{R}^q} f(x, y)dxdy = \int_{\mathcal{R}^p} \left( \int_{\mathcal{R}^q} f(x, y)dy \right) dx. \quad (7.4)$$

(2) 若  $f(x, y)$  是  $\mathcal{R}^p \times \mathcal{R}^q$  上的可积函数, 则对几乎处处的  $x \in \mathcal{R}^p$ ,  $f(x, y)$  作为  $y$  的函数是  $\mathcal{R}^q$  上的可积函数, 并且  $g(x) = \int_{\mathcal{R}^q} f(x, y)dy$  是  $\mathcal{R}^p$  上的可积函数。而且 7.4 成立。



**笔记** 在本模板中, 引理 (lemma), 推论 (corollary) 的样式和定理 7.1 的样式一致, 包括颜色, 仅仅只有计数器的设置不一样。

我们说一个实变或者复变量的实值或者复值函数是在区间上平方可积的, 如果其绝对值的平方在该区间上的积分是有限的。所有在勒贝格积分意义下平方可积的可测函数构成一个希尔伯特空间, 也就是所谓的  $L^2$  空间, 几乎处处相等的函数归为同一等价类。形式上,  $L^2$  是平方可积函数的空间和几乎处处为 0 的函数空间的商空间。

#### 命题 7.1. 最优性原理

如果  $u^*$  在  $[s, T]$  上为最优解, 则  $u^*$  在  $[s, T]$  任意子区间都是最优解, 假设区间为  $[t_0, t_1]$  的最优解为  $u^*$ , 则  $u(t_0) = u^*(t_0)$ , 即初始条件必须还是在  $u^*$  上。

我们知道最小二乘法可以用来处理一组数据, 可以从一组测定的数据中寻求变量之间的依赖关系, 这种函数关系称为经验公式。本课题将介绍最小二乘法的精确定义及如何寻求点与点之间近似成线性关系时的经验公式。假定实验测得变量之间的  $n$  个数据, 则在平面上, 可以得到  $n$  个点, 这种图形称为“散点图”, 从图中可以粗略看出这些点大致



散落在某直线近旁, 我们认为其近似为一线性函数, 下面介绍求解步骤。

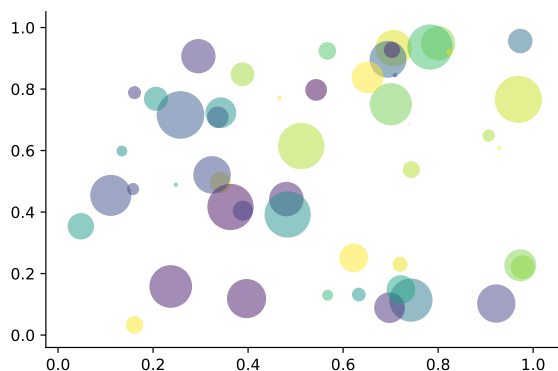


图 7.1: 散点图示例  $\hat{y} = a + bx$

以最简单的一元线性模型来解释最小二乘法。什么是一元线性模型呢? 监督学习中, 如果预测的变量是离散的, 我们称其为分类 (如决策树, 支持向量机等), 如果预测的变量是连续的, 我们称其为回归。回归分析中, 如果只包括一个自变量和一个因变量, 且二者的关系可用一条直线近似表示, 这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量, 且因变量和自变量之间是线性关系, 则称为多元线性回归分析。对于二维空间线性是一条直线; 对于三维空间线性是一个平面, 对于多维空间线性是一个超平面。

**性质** 柯西列的性质

1.  $\{x_k\}$  是柯西列, 则其子列  $\{x_k^i\}$  也是柯西列。
2.  $x_k \in \mathcal{R}^n$ ,  $\rho(x, y)$  是欧几里得空间, 则柯西列收敛,  $(\mathcal{R}^n, \rho)$  空间是完备的。

**结论** 回归分析 (regression analysis) 是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。运用十分广泛, 回归分析按照涉及的变量的多少, 分为一元回归和多元回归分析; 按照因变量的多少, 可分为简单回归分析和多重回归分析; 按照自变量和因变量之间的关系类型, 可分为线性回归分析和非线性回归分析。

## 第七章习题

1. 设  $A$  为数域  $K$  上的  $n$  级矩阵。证明: 如果  $K^n$  中任意非零列向量都是  $A$  的特征向量, 则  $A$  一定是数量矩阵。
2. 证明: 不为零矩阵的幂零矩阵不能对角化。
3. 设  $A = (a_{ij})$  是数域  $K$  上的一个  $n$  级上三角矩阵, 证明: 如果  $a_{11} = a_{22} = \cdots = a_{nn}$ , 并且至少有一个  $a_{kl} \neq 0 (k < l)$ , 则  $A$  一定不能对角化。

## 第八章 常见问题集

---

我们根据用户社区反馈整理了下面一些常见的问题，用户在遇到问题时，应当首先查阅本手册和本部分的常见的问题。

1. 有没有办法章节用“第一章，第一节，(一)”这种？

见前文介绍，可以使用 `scheme=chinese` 设置。

2. 3.07 版本的 `cls` 的 `natbib` 加了 `numbers` 编译完了没变化，群主设置了不可更改了？  
之前在 3.07 版本中在引入 `gbt7714` 宏包时，加入了 `authoryear` 选项，这个使得 `natbib` 设置了 `numbers` 也无法生效。3.08 和 3.09 版本中，模板新增加了 `numbers`、`super` 和 `authoryear` 文献选项，你可以参考前文设置说明。

3. 大佬，我想把正文字体改为亮色，背景色改为黑灰色。

页面颜色可以使用 `\pagecolor` 命令设置，文本命令可以参考[这里](#)进行设置。

4. Package `ctex` Error: CTeX fontset ‘Mac’ is unavailable.

在 Mac 系统下，中文编译请使用  $\text{XeLaTeX}$ 。

5. ! LaTeX Error: Unknown option ‘`scheme=plain`’ for package ‘`ctex`’.

你用的  $\text{CTeX}$  套装吧？这个里面的 `ctex` 宏包已经是已经是 10 年前的了，与本模板使用的 `ctex` 宏集有很大区别。不建议  $\text{CTeX}$  套装了，请卸载并安装  $\text{TeX Live 2019}$ 。

6. 我该使用什么版本？

请务必使用**最新正式发行版**，发行版间不定期可能会有更新（修复 bug 或者改进之类），如果你在使用过程中没有遇到问题，不需要每次更新**最新版**，但是在发行版更新之后，请尽可能使用最新版（发行版）！最新发行版可以在 [GitHub](#) 或者  $\text{TeX Live 2019}$  内获取。

7. 我该使用什么编辑器？

你可以使用  $\text{TeX Live 2019}$  自带的编辑器  $\text{TeXworks}$  或者使用  $\text{TeXstudio}$ ， $\text{TeXworks}$  的自动补全，你可以参考我们的总结 [TeXworks 自动补全](#)。推荐使用  $\text{TeX Live 2019}$  +  $\text{TeXstudio}$ 。我自己用 VS Code 和 Sublime Text，相关的配置说明，请参考 [L<sup>A</sup>T<sub>E</sub>X 编译环境配置：Visual Studio Code 配置简介](#) 和 [Sublime Text 搭建 L<sup>A</sup>T<sub>E</sub>X 编写环境](#)。

8. 您好，我们想用您的 ElegantBook 模板写一本书。关于机器学习的教材，希望获得您的授权，谢谢您的宝贵时间。

模板的使用修改都是自由的，你们声明模板来源以及模板地址（[GitHub](#) 地址）即可，其他未尽事宜按照开源协议 LPPL-1.3c。做好之后，如果方便的话，可以给我们一个链接，我把你们的教材放在  $\text{ElegantL<sup>A</sup>T<sub>E</sub>X}$  用户作品集里。

9. 请问交叉引用是什么？

本群和本模板适合有一定  $\text{\LaTeX}$  基础的用户使用，新手请先学习  $\text{\LaTeX}$  的基础，理解各种概念，否则你将寸步难行。

10. 定义等环境中无法使用加粗命令么？

是这样的，默认中文并没加粗命令，如果你想在定义等环境中使用加粗命令，请使用 `\heiti` 等字体命令，而不要使用 `\textbf`。或者，你可以将 `\textbf` 重新定义为 `\heiti`。英文模式不存在这个问题。

11. 代码高亮环境能用其他语言吗？

可以的，ElegantBook 模板用的是 `listings` 宏包，你可以在环境 (`\lstlisting`) 之后加上语言（比如 Python 使用 `language=Python` 选项），全局语言修改请使用 `lsset` 命令，更多信息请参考宏包文档。

12. 群主，什么时候出 Beamer 的模板（主题），ElegantSlide 或者 ElegantBeamer？

由于 Beamer 中有一个很优秀的主题 `Metropolis`。在找到非常好的创意之前不会发布正式的 Beamer 主题，如果你非常希望得到 Elegant $\text{\LaTeX}$  “官方”的主题，请在用户 QQ 群内下载测试主题 `PreElegantSlide`。正式版制作计划在今年或者明年。



## 第九章 版本更新历史

---

根据用户的反馈，我们不断修正和完善模板。截止到此次更新，ElegantBook 模板在 GitHub 上有将近 100 次提交，正式发行版本 (release) 有 16 次。

---

### 2020/02/10 更新：版本 3.10 正式发布

- ① 增加数学字体选项 `math`，可选项为 `newtx` 和 `cm`。  
重要提示：原先通过 `newtxmath` 宏包设置的数学字体改为 L<sup>A</sup>T<sub>E</sub>X 默认数学字体，如果保持原来的字体，需要显式声明数学字体 (`math=newtx`)；
- ② 新增中文字体选项 `chinesefont`，可选项为 `ctexfont`、`founder` 和 `nofont`。
- ③ 将封面作者信息设置为可选，并且增加自定义信息命令 `\bioinfo`；
- ④ 在说明文档中增加版本历史，新增 `\datechange` 命令和 `change` 环境；
- ⑤ 增加汉化章节选项 `scheme`，可选项为汉化 `chinese`；
- ⑥ 由于 `\lvert` 问题已经修复，重新调整 `ctex` 宏包和 `amsmath` 宏包位置。
- ⑦ 修改页眉设置，去除了 `\lastpage` 以避免 page anchor 问题，加入 `\frontmatter`。
- ⑧ 修改参考文献选项 `cite`，可选项为数字 `number`、作者-年份 `authoryear` 以及上标 `super`。
- ⑨ 新增参考文献样式选项 `bibstyle`，并将英文模式下参考文献样式 `apalike` 设置为默认值，中文仍然使用 `gbt7714` 宏包设置。

---

### 2019/08/18 更新：版本 3.09 正式发布

- ① 由于 `\elegantpar` 存在问题，暂时性删除 `\elegantpar` 命令，并提示用户改用 `\marginnote` 和 `\marginpar` 旁注/边注命令。
- ② 积分操作符统一更改为 `esint` 宏包设置；
- ③ 新增目录选项 `toc`，可选项为单栏 `onecol` 和双栏 `twocol`；
- ④ 手动增加参考文献选项 `cite`，可选项为上标形式 `super`；
- ⑤ 修正章节习题 (`problemset`) 环境。

## 参考文献

---

- [1] QUADRINI V. Financial Frictions in Macroeconomic Fluctuations[J]. FRB Richmond Economic Quarterly, 2011, 97(3):209-254.
- [2] CARLSTROM C T, FUERST T S. Agency Costs, Net Worth, and Business Fluctuations: A Computable General Equilibrium Analysis[J]. The American Economic Review, 1997:893-910.
- [3] LI Q, CHEN L, ZENG Y. The Mechanism and Effectiveness of Credit Scoring of P2P Lending Platform: Evidence from Renrendai.com[J]. China Finance Review International, 2018, 8(3):256-274.
- [4] 方军雄. 所有制、制度环境与信贷资金配置[J]. 经济研究, 2007(12):82-92.
- [5] 刘凤良, 章潇萌, 于泽. 高投资、结构失衡与价格指数二元分化[J]. 金融研究, 2017(02):54-69.
- [6] 吕捷, 王高望. CPI 与 PPI “背离”的结构性解释[J]. 经济研究, 2015, 50(04):136-149.



## 附录 A 基本数学工具

---

本附录包括了计量经济学中用到的一些基本数学，我们扼要论述了求和算子的各种性质，研究了线性和某些非线性方程的性质，并复习了比例和百分数。我们还介绍了一些在应用计量经济学中常见的特殊函数，包括二次函数和自然对数，前 4 节只要求基本的代数技巧，第 5 节则对微分学进行了简要回顾；虽然要理解本书的大部分内容，微积分并非必需，但在一些章末附录和第 3 篇某些高深专题中，我们还是用到了微积分。

### A.1 求和算子与描述统计量

**求和算子**是用以表达多个数求和运算的一个缩略符号，它在统计学和计量经济学分析中扮演着重要作用。如果  $\{x_i : i = 1, 2, \dots, n\}$  表示  $n$  个数的一個序列，那么我们就把这  $n$  个数的和写为：

$$\sum_{i=1}^n x_i \equiv x_1 + x_2 + \cdots + x_n \quad (\text{A.1})$$