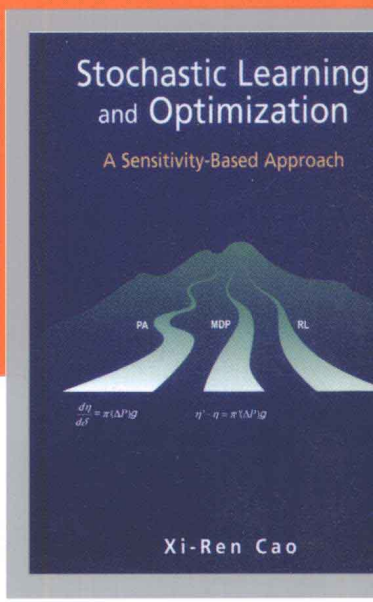
信息技术和电气工程学科国际知名教材中译本系列

Stochastic Learning and Optimization
A Sensitivity-Based Approach

随机学习与优化
——基于灵敏度的方法



曹希仁 著
Xi-Ren Cao
陈 曦 译

清华大学出版社
Springer

信息技术和电气工程学科国际知名教材中译本系列

Stochastic Learning and Optimization
——A Sensitivity-Based Approach

随机学习与优化
——基于灵敏度的方法

曹希仁 著
Xi-Ren Cao

陈 曦 译

清华大学出版社
北京

Springer

北京市版权局著作权合同登记号　图字：01-2009-2126

图书在版编目(CIP)数据

随机学习与优化：基于灵敏度的方法（第2版）/（美）曹希仁著；陈曦译. —北京：清华大学出版社，2011.2
书名原文：Stochastic Learning and Optimization: A Sensitivity—Based Approach
（信息技术和电气工程学科国际知名教材中译本系列）
ISBN 978-7-302-24292-5

Ⅰ. ①随… Ⅱ. ①曹… ②陈… Ⅲ. ①自动控制理论—教材 Ⅳ. ①TP13

中国版本图书馆CIP数据核字（2010）第252112号

责任编辑：王一玲　刘佩伟
责任校对：焦丽丽
责任印制：何　芊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦A座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010 62770175　　**邮　购**：010-62786544
投稿与读者服务：010 62795954，jsjjc@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：29　**字　数**：691千字
版　　次：2011年2月第1版　　**印　　次**：2011年2月第1次印刷
印　　数：1～3000
定　　价：49.00元

产品编号：027866 01

《信息技术和电气工程学科国际知名教材中译本系列》

出 版 说 明

三年多以前，2000 年 10 月，为了系统地参考和借鉴国外知名相关大学教材，推进我国大学的课程改革和我国大学教学的国际化进程，清华大学出版社策划、出版了《国际知名大学原版教材——信息技术学科与电气工程学科系列》，至今已经出版了 30 多种，深受高等院校信息技术与电气工程及相关学科师生和其他科技人员的欢迎和好评，在学术界和教育界产生了积极的影响。现在这个系列中的大部分教材都已经重印，并曾获得《2001 年引进版优秀畅销丛书奖》。在此期间，我们曾收到来自各地高校师生的很多反映，期望我们选择这个系列中的一些较为基础性和较为前沿性的教材译成中译本出版，以为更广大的院校师生和科技人员所选用。正是基于这种背景和考虑，清华大学出版社决定进一步推出《信息技术和电气工程学科国际知名教材中译本系列》。

这套国际知名教材中译本系列所选书目的范围，限于信息技术和电气工程学科所属各专业的技术基础课和主要专业课。教材原版本除了选自《国际知名大学原版教材——信息技术学科与电气工程学科系列》外，还将精选其他具有较大影响的国外知名的相关领域教材或教学参考书。教材内容适于作为我国普通高等院校相应课程的教材或主要教学参考书。

本国际知名教材中译本系列按分期分批的方式组织出版。为了便于使用这套国际知名教材中译本教材系列的相关师生和科技人员从学科和教学的角度对其在体系和内容上的特点和特色有所了解，在每种中译本教材中都附有我们约请的相关领域资深教授撰写的推荐说明，其中的一些直接取自于《国际知名大学原版教材——信息技术学科与电气工程学科系列》中的影印版序。

本国际知名教材中译本系列的读者对象为信息技术和电气工程学科所属各专业的本科生或研究生，同时兼顾其他工程学科专业的本科生或研究生。既可采用作为相应课程的教材，也可作为相应课程的教学参考书。此外，本国际知名教材中译本系列也可提供作为工作于各个技术领域的工程师和技术人员的自学读物。

感谢使用本国际知名教材中译本系列的广大师生和科技人员的支持。期望广大读者提出意见和建议。

郑大钟 教授
清华大学信息科学技术学院

序

From a humble and difficult beginning, the subject of Perturbation Analysis has grown into a important sub-discipline of the operations research and simulation literature. Here is a book by the person who was co-inventor and the leading expert of the topic. Not only has he written with the latest insight on the topic but he also integrated it with two other broad subjects - Markov Decision Process and Reinforcement Learning. This unified viewpoint and treatment make accessible a large body of system engineering knowledge to senior and beginning graduate students.

The translator has taught a course using the material of this book together with the author for the past several years at Tsinghua University and thus is uniquely qualified for the task.

Yu-Chi Ho

– Yu-Chi Ho Gordon McKay Professor of Systems Engineering, T. Jefferson Coolidge Chair of Applied Mathematics, Harvard University Member, National Academy of Engineering, USA

摄动分析经历了初出茅庐的艰辛，如今已发展成运筹学和仿真领域的一个重要分支. 本书的作者是这个学科的共同创立人和领军专家. 在本书中作者不仅展现了对该课题的最新领悟，并且将其与两个更广泛的学科——马尔可夫决策过程和强化学习，结合在了一起. 对于高年级和新入学的研究生来说，这个统一的视角和处理方式使得大部分的系统工程的知识变得明白易懂.

本书的译者在过去的几年曾与作者一起在清华大学开设课程讲授书中的内容，因而特别适合这本书的翻译工作.

何毓琦

作者中译本序

科学研究的精髓在于洞察复杂事物的本质并将其以简洁的方式系统地、逻辑地表述出来. 本人研究随机系统的优化理论与方法三十余年, 发现了一个简单而通用的原理, 即这个领域的不同学科都建基于一个基本概念 (性能势) 和两个基本公式 (性能微分与差分). 这个简单的原理使许多重要的结果的推导变得简易且直观, 并指引了新的研究方向. 本书在此基础上写成. 阅读本书仅需基本的概率论及矩阵运算知识, 不超过大学本科工科专业的数学教材. 书中一些例子则涉及较深一些的排队论的知识.

本书的英文版出版后在学术界获得一些好评, 今摘录一些于书尾供读者参考.

欣闻拙作将出中文版. 希望能对学习控制和优化领域的青年学者、研究生、大学生和工程师们有所裨益. 十分感谢陈曦教授为翻译此书花费的巨大精力, 感谢清华大学、香港科技大学历届学生在编译过程中的诸多帮助, 也感谢清华大学出版社及Springer出版社对出版中文版的大力支持.

曹希仁

2010年7月

上海交通大学

前言

现代工程系统包括通信（互联网和无线）、制造、机器人和物流等很多领域，无论在系统的设计还是运行阶段，性能优化都非常重要．然而，大多数工程系统过于复杂以致于无法对它们建模，或者不能轻易辨识其系统的参数．因此，我们不得不使用学习的技巧．

学习和优化简述

随机系统的学习和优化是一个多学科的领域，它已引起包括控制系统、运筹学和计算机科学等许多学科的研究人员的广泛注意．离散事件动态系统（discrete event dynamic systems）中的摄动分析（perturbation analysis），运筹学中的马尔可夫决策过程（markov decision processes），计算机科学中的强化学习（reinforcement learning），控制系统中的神经元动态规划（neuro-dynamic programming），辨识和自适应控制（identification and adaptive control）等等，都有一个共同的目标，就是做出“最好的决策”来优化系统性能．

不同的领域对目标相同的问题会采取不同的视角和不同的描述．本书从灵敏度的角度，为摄动分析、马尔科夫决策过程、强化学习、辨识和自适应控制等提供了一个统一的框架，并引入了新的方法，进而在这个基于灵敏度的框架下，提出了新的研究课题和方向．

粗略地讲，强化学习通过观测和分析系统现在的行为，学会如何在不知道甚至不需要估计系统结构和参数的情况下，作出决策以改进系统的性能；摄动分析通过观测和分析系统的行为来估计系统的性能关于参数的导数，而其优化则是通过性能导数的估计和其他优化技术（如随机逼近）的组合来改进系统的性能．马尔可夫决策过程提供了马尔可夫模型系统的性能优化的理论基础[21,216]．在自适应控制中，系统行为由微分或差分方程来描述；当系统参数未知时，通过观测到的数据来辨识．自适应控制与辨识相结合能获得与学习和优化相同的结果．

这些研究领域的目标相同，即利用观测或分析系统行为而“学习”到的信息，找出优化系统性能的策略．已知系统的状态或历史，策略决定了作用于系统的行动，而行动控制了系统的演化．在一些情况下，策略依赖于连续参数，而策略空间是连续的；在另外一些情况下，策略空间是离散的并且策略的数目巨大．

基于灵敏度的观点

近年来的研究表明，我们可以从基于策略空间的性能灵敏度这个统一的观点来解释学习和优化的不同学科[56]. 学习和优化的基本元素是两种类型的性能灵敏度公式，一类是在策略空间中的任一策略处的性能导数，另一类是策略空间中的两个策略之间的性能差分. 通过这两类灵敏度公式，可以用简单和直观的方法推导或解释不同领域里已有的结果以及它们之间的关系，可以引入新的方法，还可以用相同的方法来处理平均，折扣以及其他性能准则.

这个统一的框架是建立在一些简单而基本的事实之上的. 一般来讲，当系统的结构信息未知时，通过观测和分析系统在一个策略下的行为，当然不能了解该系统在其他策略下的性能；我们每次也只能比较两个策略的性能. 问题是，在这些基本的限制之下，如何能够用尽量少的系统结构信息和尽量少的计算量达到性能优化的目标?

沿着这个方向，我们发现可以做下面两件事. 首先，如果策略空间是连续的，*利用系统结构的一些知识（如，排队或马尔可夫）和摄动分析原理，可以通过观测分析一个策略下的系统行为来估计该策略在策略空间中任何方向上的性能导数*[70,62,69]. 这样就得到性能导数公式. 更进一步，通过分析当前策略的样本路径可以获得计算导数需要的所有的量. 性能导数公式也成为了摄动分析以及最近在强化学习研究领域中提出的“策略梯度”方法的基础.

第二，如果策略空间是离散的，性能差分公式便成了优化的基础. 差分公式比较两个策略下的系统性能. 但是，与导数公式不同的是，差分公式涉及到两个策略的量，利用差分公式也不可能仅通过观测或分析一个策略下的系统行为，就知道另一个策略的性能，或者同一个系统在这两个策略下的性能的差. 幸运的是，*在某些结构条件下，通过性能差分公式的特殊的因子分解形式，由观测分析一个策略下的系统行为所学习得到的信息，总能找到另一个策略，如果这样的策略存在，在此策略下，系统性能会更好*. 这样一来就产生了策略迭代：从一个策略中学习并找到另一个较好的策略，再从这个较好的策略出发，找出一个更好的策略，如此迭代下去. 因此，性能差分公式便成为性能优化策略迭代这类方法的基础. 我们将会说明，用此原则也可以获得在辨识和自适应控制中所得到的结果，该原则为此领域提供了基于学习的视角.

在这两种灵敏度公式中基本的量是性能势，它具有明确的物理意义：它度量了一个状态对系统性能的“潜在”贡献. 两个状态的性能势的差，度量了从一个状态变到另一个状态对系统性能的影响. 这种从一个状态到另一个状态的变化，在摄动分析中被称为摄动（或简单被称为样本路径上的“跳变”）. 在强化学习中，已经有许多有效的算法（如TD（λ）和Q-学习）来估计性能势和它的变形Q-因子，以及它们在最优策略下的值.

由性能势的物理解释可以得到摄动分析的基本原则：系统结构或参数的任何变化所产生的影响，可以分解成许多状态之间的跳变（或摄动）的影响之和. 依据这个原理，可以把性能势当成一个组成单元，对不同的问题直观地构建出新的灵敏度公式，这些问题可能不能用现有文献中的标准方法来描述[59]. 因此，灵敏度公式是学习和优化的基础，灵敏度的构造方法开辟了一个新的研究方向：基于这些新的灵敏度公式，可以开发出新的学习和优化方法，同时对系统的特性也能善加利用.

基于事件的优化就是利用灵敏度的构造开发出的一个新的优化方法. 该方法利用了由事件所描述的系统特性. 策略依赖于事件而非状态. 一个事件被定义为一组状态转化的集

合，因此，当前发生的事件包含下一个状态，也即属于未来的一些信息．在信息技术的许多现代工程系统中，在采取行动前可以获得这样的信息，但标准马尔可夫模型并不能描述这一特性．因此，在某些情况下，基于事件的策略可能比基于状态的策略表现得更出色．另一方面，事件的数量通常与系统的大小成正比，比状态的数量少很多，后者随系统的大小呈指数增长．所以，在一些条件下，这种方法为克服或减轻计算上的困难，即维数灾，提供了可能．另外，通过定义不同的事件来描述不同问题的特性，诸如部分可观测的马尔可夫决策过程，状态和时间集结，分层控制（混合系统），选项（options），以及奇异摄动等已有的方法，都可以看成是基于事件的优化的特例．

本书的独到之处

与学习和优化领域的其他著作相比，本书在以下方面具有独到之处．

1. 本书利用策略空间上灵敏度观点的统一框架，覆盖了学习和优化的各门学科，它们包括：摄动分析、马尔可夫决策过程、强化学习，以及辨识和自适应控制等，其中的许多结果可以简单地用两类基本的灵敏度公式来解释．
2. 本书强调物理解释而非数学推导．通过直观的物理解释，我们意在用性能势作为组成单元构造出新的灵敏度的公式．物理上的直观认识可以提供给我们完善已有方法的新思路．
3. 运用统一的框架和构造方法，我们引入了最近发展出来的基于事件的优化方法．该方法开辟了一个新的研究方向，通过利用系统的特性，可以克服或减轻维数灾．
4. 我们将基于性能差分的方法应用于所有的马尔可夫决策过程，它们包括：遍历系统和多链系统、平均性能准则和折扣性能准则，以及偏差优化和n阶偏差优化，等等．我们还将证明由n阶偏差优化策略将最终得出Blackwell优化策略．这种方法用统一的方式对马尔可夫决策过程中的这些问题，提供了一个简单、明了和全面的表述，在现有的论著中这是独一无二的．

本书的内容

第1章是引言，包含学习和优化中的各门学科的综述，以及基于事件的方法的讨论．该章相当于本书的路线图．本书其余内容由三部分组成．第一部分从第2章到第7章，描述了如何从策略空间的灵敏度观点推导出摄动分析，马尔可夫决策过程，强化学习，以及辨识和自适应控制等的主要概念和结果．第二部分包括第8章和第9章，展示了利用灵敏度观点在基于事件的学习和优化领域的最新研究成果．第三部分包含的三个附录为学习本书提供了必需的数学基础．

第一部分从第2章开始．以性能势或实现因子作为组成单元，推导出马尔可夫系统和排队系统的性能导数公式．摄动分析中基于样本路径的灵敏度的观点是本书的统一方法的核心．在第3章，讨论了性能势，开发出基于样本路径估计性能势，性能导数以及利用性能势进行优化的算法．在第4章，说明了如何可以轻易地从性能差分公式中导出单链和多链的马尔可夫决策过程中的策略迭代．这种方法同样适用于平均和折扣准则，以及偏差优化等．还定义和解决了n阶偏差优化问题．在第5章，利用由样本路径估计得到的性能势，开发出在线策

略迭代算法. 第6章给出了强化学习的基本结果，它本质上是随机优化和基于样本路径的性能势以及它们的变形Q-因子的估计的组合. 第7章说明在线策略迭代方法可以应用于包括线性系统和一些非线性系统的辨识和自适应控制问题.

第二部分中，第8章给出了基于事件的优化方法. 该方法为利用特殊的系统结构解决维数灾的难题提供了一个可能的途径. 在某些情况下，基于事件的策略可能比基于状态的策略有更好的性能. 第9章说明了在一般的问题中，如何用性能势作为组成单元构建灵敏度的公式.

如何使用本书

本书采用一个统一的方式，为有兴趣了解摄动分析、马尔可夫决策过程、强化学习、辨识和自适应控制、随机逼近等不同学科的学习和优化理论的相关方法以及它们之间的相互关系的研究生和工程师们提供了入门的材料，书中的新观点有助于读者发现新的研究课题. 因此，本书对希望在这些领域找到研究动机和促进学科之间的合作的研究人员非常有用. 另外，特别是信息技术领域的工程师们可能会发现本书所介绍的观点和方法在实际应用中对他们也很有帮助.

标有“*”的章节是补充阅读材料，初次阅读者可以略去. 每章包含的大量习题可以帮助学生加深理解. 其中一些习题，总结了过去的一些研究课题，或许有些难度. 习题的答案可以通过与我联系或在我的主页http://cfins.au.tsinghua.edu.cn/personalhg/caoxiren/ 中获得.

本书的早期版本曾作为香港科技大学和北京的清华大学研究生课程的教材使用. 一学期为十四周（每周三小时）的课时，建议课程安排如下.

章	节	小时	周次
A~C	A.1 ~ C.2	3	1
1	1.1 ~ 1.4	2	2/3
2	2.1	4	4/3
	2.2	1	1/3
	2.4	3	1
3	3.1 ~ 3.3	3	1
	复习	1	1/3
4	4.1	3	1
	4.2	3	1
5	5.1 ~ 5.2	3	1
6	6.1 ~ 6.4	4	4/3
7	7.1 ~ 7.3	2	2/3
8	8.1 ~ 8.5	5	5/3
9	9.1 ~ 9.2	1	1/3
	复习	1	1/3
	考试	3	1
		总计42	14

以下是对课程中每章内容的建议和评论：

0. 附录中的内容包含学习本课程的预备知识. 这三个附录中的详细内容用三小时来复习是不够的. 简单复习时，可以着重于概率论和马尔可夫链，它们和本书内容紧密相关. 附录B主要和第4章有关，附录C主要和2.4节有关. 有些结果也可以在讲授主要内容时再复习.
1. 对有控制背景的学生，1.1节中关于策略的部分可以讲得快一点. 1.2节到1.3节是为了让学生对不同学科有一个总的认识，应该在学完第一部分后再复习，以便加深理解.
2. 第2章的主要部分是2.1和2.4两节.
3. 3.2节在文献中比较新.
4. 4.1和4.2两节涵盖了方法论中的主要思想. 如果时间允许，可以讲一讲4.3节中的主要结果而忽略证明.
5. 5.2.3节中的证明很有趣，但多少有点技巧，且需要仔细思考.
6. 在第6章中，根据随机逼近原理，我们强调迭代算法背后的直觉，而不是要证明这些算法. 性能导数的估计算法是近年来新的研究课题.
7. 在第7章，比较容易让学生相信控制系统可以建模成一个马尔可夫决策过程. 马尔可夫决策过程从离散状态空间扩展到连续状态空间没有概念上的困难. 作为例子，可以只讲线性二次问题.
8. 8.1节给出基于事件的优化方法的一个概述. 如果希望避免学习繁冗的数学公式，可以通过其中的两个例子来清晰理解该方法.
9. 9.2节提供了构造性能差分公式的基本想法. 其他节阐明了该方法的灵活性，可供补充阅读.

九周（每周三小时）的课程建议安排如下：

章	节	小时	周次
A~C	A.1 ~ C.2	1.5	1/2
1	1.1 ~ 1.4	1.5	1/2
2	2.1	4	4/3
3	3.1 ~ 3.3	2	2/3
4	4.1	3	1
	4.2.1	2	2/3
5	5.1 ~ 5.2	2	2/3
6	6.1 ~ 6.4	3	1
7	7.1 ~ 7.3	2	2/3
8	8.1 ~ 8.5	3	1
9	9.1 ~ 9.2	1	1/3
	考试	2	2/3
		总计27	9

另有建议如下:

1. 我们没有时间讲授2.4节中排队系统的摄动分析，3.3节中对排队系统的基于摄动分析的优化，等等.
2. 在第4章中，可以只是简要地介绍n阶偏差的概念和n阶偏差优化的问题.
3. 基于事件的优化可以通过例子来介绍.

致谢

本书的大部分内容都是基于我本人对学习和优化的研究. 自20世纪80年代在哈佛大学从事摄动分析开始，我在这个领域的研究已有二三十年. 在这期间，有很多人通过各种方式为我的研究提供了帮助.

真诚感谢何毓琦教授对我不断的支持和鼓励. 他的洞察力和灵感对我的研究产生了巨大的影响. 我要感谢以下人士，他们在不同的阶段曾与我在与本书相关的课题上有过合作或深入的讨论：K. J. Åström, T. Başar, A. G. Barto, D. P. Bertsekas, R. W. Brockett, C. G. Cassandras, 陈翰馥, A. Ephremides, 方海涛, E. A. Feinberg, M. C. Fu, P. Glasserman, 龚维博, 郭先平, B. Heidergott, P. V. Kokotovic, F. L. Lewis, L. Ljung, 马大骏, S. I. Marcus, S. P. Meyn, G. Ch. Pflug, 丘立, 任志远, 司捷, R. Suri, J. N. Tsitsiklis, B. Van Roy, P. Varaiya, A. F. Veinott, Y. Wardi, 温日华和张俊玉. 还要感谢仔细阅读本书早期的草稿，并对文字表述提出有益建议，指出排版错误的人士：曹芳，陈翰馥，陈同文，郭先平，李泉林，李衍杰，史定华，夏俐，徐琰恺和张俊玉. 另外，特别感谢徐琰恺和张俊玉，他们承担了为本书（英文版）编制插图的Latex文件的繁冗工作. 也感谢V. Unkefer女士为本书大部分内容所做的技术编辑工作，同时感谢J. Q. Shen为本书绘制封面的初稿. 当然，所有的错误责任归我. 同时，我也对哈佛大学、美国数字设备公司和香港科技大学多年来为我提供财政上的支持以及优良的研究环境表示诚挚的感谢.

最后，真诚感谢我的夫人王正敏多年来无论在任何情况下对我的不断的支持和理解.

香港
2007年4月

曹希仁
香港科技大学
eecao@ust.hk

符号与缩写

符号

a	事件
α	行动
$\mathcal{A}$	行动空间
$\mathcal{A}(i)$	在状态i处可用的行动集
A_l	在时刻l时所采取的行动
$\boldsymbol{A}_l=(A_0,\cdots,A_l)$	到时刻l的行动历史
β	折扣因子
B	无穷小生成元
$c(\boldsymbol{n},i)$	在排队系统状态为$\boldsymbol{n}$时服务台i的单位摄动实现概率
$c^{(f)}(\boldsymbol{n},i)$	对于性能函数f, 在排队系统状态为$\boldsymbol{n}$时, 服务台i的单位摄动实现因子
d, h	策略
$\mathcal{D}$	策略空间
$\mathcal{D}_0$	收益最优的策略空间
$\mathcal{D}_n$, $n=1,2,\cdots$	n阶偏差最优的策略空间
δ, θ	参数
$e=(1,\cdots,1)^T$	所有分量均为1的列向量
T	D^T: 矩阵D的转置
$e_{\cdot i}$	除第i个分量为1外, 其余元素均为0的列向量
e_S	维数为S的单位向量
$e_c(k_2)$	可控事件, $k_2=1,\cdots,k_c$
$e_o(k_1)$	可观事件, $k_1=1,\cdots,k_o$
$e_t(k_3)$	自然转移事件, $k_3=1,\cdots,k_t$

$\mathcal{E}$	单一事件(状态转移)集
E	期望
E_l	在时刻l时发生的事件
$\boldsymbol{E}_l=(E_0,\cdots,E_l)$	到时刻l的事件历史
η	性能指标(长期平均报酬)
η_β	折扣报酬
η^*	最优收益
f	性能(报酬)函数
f^d	在策略d下的性能(报酬)函数
F_L	在离散时间情况下, 等于$\sum_{l=0}^{L-1}f(X_l)$;
	在连续时间情况下, 等于$\int_0^{T_L}f(X_t)dt$
g	性能势函数, 向量; 偏差
g_β	折扣性能势
g^d	在策略d下的性能势函数
$g_n,\ n=0,1,\cdots$	n阶性能势, n阶偏差
$g_n^*,\ n=0,1,\cdots$	最优n阶偏差
$g_n^d,\ n=0,1,\cdots$	在策略d下的n阶性能势, n阶偏差
$\Gamma=[\gamma(i,j)]$	摄动实现矩阵
$\boldsymbol{H}_l$	到时刻l的历史信息, $\boldsymbol{H}_l=\{\boldsymbol{Y}_l,\boldsymbol{A}_{l-1}\}$
κ	步长
$L(i\|j)$	马尔可夫链从初态j到状态i的首达时间
L_{ij}^*	始自不同初态i和j的两条马尔可夫链的汇合点
n_i	在服务台i的顾客数
$\boldsymbol{n}=(n_1,\cdots,n_M)$	排队网络的状态
$\boldsymbol{n}_{-i,+j}$	状态$\boldsymbol{n}$的邻近状态
$P=[p(j\|i)]$	转移概率矩阵
P^d	在策略d下的转移概率矩阵
$\Delta P=(\Delta P)^{d,h}$	等于P^h-P^d: 从P^d到P^h的方向
P_δ	等于$P+\delta\Delta P$或$P^d+\delta(\Delta P)^{d,h}$
P^*	矩阵P^n的Cesaro极限
$\mathcal{P}$	概率测度

π	稳态概率, 行向量
π^d	在策略d下的稳态概率
$Q=[q_{i,j}]$	排队网络中的路由概率矩阵
$Q(i,\alpha)$	状态i和行动α的Q因子
$\Re$	实数空间$(-\infty,\infty)$
ρ	服务强度, 在$M/M/1$中$\rho=\frac{\lambda}{\mu}$
$\rho(R)$	矩阵R的谱半径
$\bar{s}_i$	服务台i的平均服务时间
S	状态空间的状态数
$\mathcal{S}$	状态空间
T_l	连续时间马尔可夫过程的第l个转移时刻
$T(i\|j)$	马尔可夫过程从初态j到状态i的首达时间
T_{ij}^*	始自不同初态i和j的两个马尔可夫过程的汇合点
X_l	马尔可夫链$\boldsymbol{X}$在时刻l的状态
X_t	马尔可夫过程$\boldsymbol{X}$在时刻t的状态
$\boldsymbol{X}$	马尔可夫链或马尔可夫过程, 一条样本路径
$\boldsymbol{X}_l$	到时刻l的状态历史
$\boldsymbol{X}_\delta=\boldsymbol{X}_\delta^{d,h}$	转移概率矩阵为$P^d+\delta(\Delta P)^{d,h}$的马尔可夫链的样本路径
Y_l	在时刻l的观测
$\boldsymbol{Y}_l$	到时刻l的观测历史
$\times$	笛卡儿积
$\otimes$	张量积
$\langle i,j\rangle$	从状态i到状态j的转移
$\geqslant(\leqslant)$	两个向量满足$u\geqslant(\leqslant)\ v$意味着对于所有的i, $u(i)\geqslant(\leqslant)\ v(i)$
$\succeq(\preceq)$	两个向量满足$u\succeq(\preceq)\ v$意味着$u\geqslant(\leqslant)\ v$且$u\neq v$
$\#$	$B^{\#}$: 矩阵B的群逆

缩写

CR	共同实现(Common realization)
DEDS	离散事件动态系统(Discrete event dynamic system)
EAMC	等价的集结马尔可夫链 (Equivalent aggregated Markov chain)
FCFS	先到先服务(First come first serve)
GM	梯度方法(Gradient method)
GPI	通用策略迭代(Generalized policy iteration)
I&AC	辨识与自适应控制(Identification and adaptive control)
i.i.d.	独立同分布(Identically and independently distributed)
JLQ	跳变线性二次问题(Jump linear quadratic)
LCFS	后到先服务(Last come first serve)
LDQ	线性折扣二次问题(Linear discounted quadratic)
LQ	线性二次问题(Linear quadratic)
LQG	线性二次高斯问题(Linear quadratic Gaussian)
LR	似然比(Likelihood ratio)
MDPs	马尔可夫决策过程(Markov decision processes)
NDP	神经元动态规划(Neuro-dynamic programming)
PA	摄动分析(Perturbation analysis)
PASTA	Poisson arrivals see time average
PDF	性能差分公式(Performance difference formula)
PE	泊松方程(Poisson equation)
PG	策略梯度(Policy gradient)
PI	策略迭代(Policy iteration)
POMDPs	部分可观马尔可夫决策过程 (Partially observable Markov decision processes)
PRF	摄动实现因子(Perturbation realization factor)

PS	处理器共享(Processor sharing)
Q-L	Q学习(Q-learning)
RL	强化学习(Reinforcement learning)
RM	Robins-Monro算法(Robins-Monro algorithm)
SA	随机逼近(Stochastic approximation)
SARSA	State-action-reward-state-action
SD	标准差(Standard deviation)
SFM	随机流模型(Stochastic fluid model)
SMP	半马尔可夫过程(Semi-Markov process)
TD	瞬时差分(Temporal difference)
WD	弱导数(Weak derivative)
w.p.1	以概率1 (With probability 1)