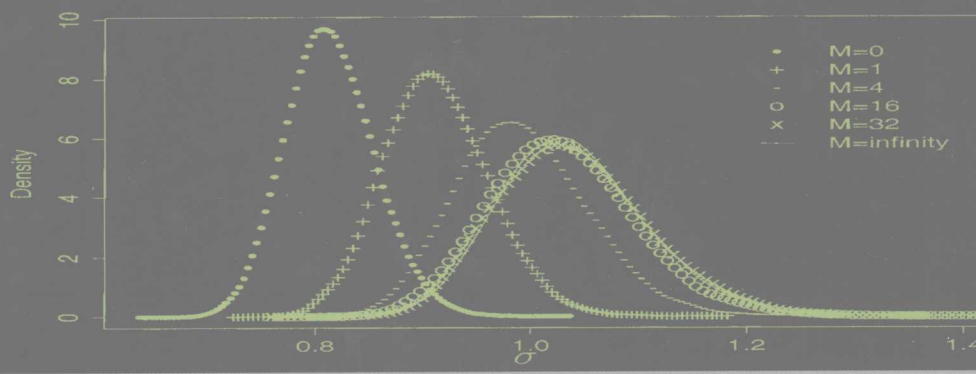
当代科学前沿论丛

NEW FRONTIERS OF SCIENCES

科学计算中的蒙特卡罗策略

Monne Carlo Strategies in Scientific Computin



刘 军 著　　唐年胜 周 勇 徐 亮 译

高等教育出版社 HIGHER EDUCATION PRESS

当代科学前沿论丛

科学计算中的蒙特卡罗策略

刘 军 著　唐年胜 周 勇 徐 亮 译

高等教育出版社

图字：01-2008-0393 号

内容简介

本书系统全面地介绍了蒙特卡罗方法的基本原理、序贯蒙特卡罗理论、行为中的序贯蒙特卡罗方法、Metropolis 算法及其推广、Gibbs 抽样、一般条件抽样等。此外，本书还详细阐述了这些理论和方法在物理学、生物学和化学等领域的具体应用，并且还辅以大量的模拟研究结果及其相关问题，便于教师组织教学和学生进行学习。

本书可作为统计学、生物遗传学、物理学、化学、教育心理学、社会科学和计算机科学等专业研究生的教学参考书，也可供相关专业的研究生、教师、统计工作者以及从事或者对蒙特卡罗算法研究感兴趣的科研人员参考。

图书在版编目（CIP）数据

科学计算中的蒙特卡罗策略 / 刘军著；唐年胜，周勇，徐亮译. —北京：高等教育出版社，2009.4
（当代科学前沿论丛）
书名原文：Monte Carlo Strategies in Scientific Computing
ISBN 978-7-04-025837-0

Ⅰ. 科… Ⅱ. ①刘…②唐…③周…④徐… Ⅲ. 蒙特卡罗法-教材 Ⅳ. O242.2

中国版本图书馆 CIP 数据核字（2009）第 013617 号

策划编辑	王丽萍	**责任编辑**	李华英	**封面设计**	于 涛	**责任绘图**	尹 莉
版式设计	马敬茹	**责任校对**	王 雨	**责任印制**	宋克学		

出版发行	高等教育出版社	**购书热线**	010 - 58581118
社　　址	北京市西城区德外大街 4 号	**免费咨询**	800 - 810 - 0598
邮政编码	100120	**网　　址**	http://www.hep.edu.cn
总　　机	010 - 58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经　　销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印　　刷	北京新华印刷厂	**畅想教育**	http://www.widedu.com
开　　本	787 × 1092　1/16	**版　　次**	2009 年 4 月第 1 版
印　　张	17.5	**印　　次**	2009 年 4 月第 1 次印刷
字　　数	420 000	**定　　价**	36.00 元

物料号　25837 - 00

前 言

蒙特卡罗计算的基本思想最初源于蒲丰在 1777 年提出的著名“蒲丰投针问题”的一项早期实验(Dörrie 1965). 在这个著名的实验中, 实验者向平行线网格间距为 D 的平面上投一长度为 l 的针 ($D > l$). 在理想条件下, 很容易计算出针与任意一条平行线相交的几率为 $2l/\pi D$. 因而, 如果假设 p_N 为 N 次投针实验中针与平行线“相交”的比率, 则

$$\hat{\pi} = \frac{2l}{p_N D}$$

可作为 π 的一个估计, 并且当 N 趋于无穷时, $\hat{\pi}$ 收敛到 π. 确实还真有一些研究者用此方法来估算 π 的值. 借助模拟随机过程来估计某一有兴趣的量的思想现已成为科学计算的重要组成部分.

蒙特卡罗方法在现实科学问题中的系统应用始于电子计算的早期时代 (1945—1955), 并伴随着世界上第一台可编程的“超大”计算机——MANIAC (数学分析机, 数值积分器和计算机)——于第二次世界大战期间在洛斯阿拉莫斯 (Los Alamos) 的发展而不断发展. 为了更好地使用这些具有快速计算能力的机器, 科学家们 (Stanislaw Ulam, John von Neumann, Nicholas Metropolis, Enrico Fermi 等) 提出了一种基于统计抽样技术的方法, 用以解决原子弹设计中有关易裂变物质的随机中子扩散的数值计算问题和估计 Schrödinger 方程中的特征根问题. 这一方法的基本思想首先由 Ulam 提出, 然后在他与 von Neumann 驾车从洛斯阿拉莫斯到拉米 (Lamy) 的途中, 经两人仔细考虑后得以正式提出. 据说, 是 Nick Metropolis 将此方法冠名为“蒙特卡罗”的, 该名称为推广使用这一方法起到了十分重要的作用.

早在 20 世纪 50 年代, 统计物理学家们 (N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller 和 E. Teller) 就为简单流体的模拟引入了基于马尔可夫链的动态蒙特卡罗方法. 这一方法随后被推广覆盖到越来越复杂的物理系统中, 包括自旋玻璃 (spin glass) 模型、谐波型晶体和多聚体模型等. 在 20 世纪 80 年代, 统计学家与计算机科学家发展了用以解决诸如组合优化、非参数统计推断 (如: 刀切法和自助法)、带有缺失观测值的似然计算、统计遗传分析、贝叶斯建模与计算等问题的基于蒙特卡罗的方法. 在 20 世纪 90 年代, 蒙特卡罗方法在计算生物学中开始发挥重要作用, 而且它被用来解决序列基序识别和复杂的谱系 (pedigree) 分析问题. 现在, 蒙

特卡罗方法的应用领域包括生物学 (Leach 1996, Karplus 和 Petsko 1990, Lawrence, Altschul, Boguski, Liu, Neuwald 和 Wootton 1993)、化学 (Alder 和 Wainwright 1959)、计算机科学 (Kirkpatrick, Gelatt 和 Vecchi 1983)、经济学与金融学 (Gouriérourx 和 Monfort 1977)、工程学 (Geman和Geman 1984)、材料科学 (Frenkel 和 Smit 1996)、物理学 (Metropolis, Rosenbluth, Rosenbluth, Teller 和 Teller 1953, Goodman 和 Sokal 1989, Marinari 和 Parisi 1992)、统计学 (Efron 1979, Gelfand 和 Smith 1990, Rubin 1987, Tanner 和 Wong 1987) 以及其他许多学科. 在所有的蒙特卡罗方法中, 马尔可夫链蒙特卡罗理论 (MCMC) 为处理复杂的随机系统提供了巨大的机会, 同时也是大分子学和其他物理系统研究中的中流砥柱. 最近, 由于 MCMC 理论和技术能使统计学家考虑更复杂、更现实的统计模型, 所以它引起了统计学家们的广泛关注.

许多不同科学领域的研究者由于受蒙特卡罗方法的高度灵活性和超强功效性的吸引, 都为它的发展做出了相应的贡献. 然而, 要了解任何一个领域的问题都需要大量丰富的特定专业领域的知识, 这就大大地限制了不同领域中研究者的相互交流. 近年来, 大量的研究工作致力于重新发现在其他领域中已经发展出的各项技术. 因此, 迫切需要发展一个相对通用的框架, 在此框架下, 每个领域的科学家,如: 理论化学家、统计物理学家、结构生物学家、统计学家、经济计量学家和计算机科学家, 既能相互比较各自的蒙特卡罗技术, 又可以相互学习. 许多把蒙特卡罗模拟和有关全局优化技术 (如: 模拟退火) 作为其研究工作中必不可少的工具的科学家和工程师们也需跟上蒙特卡罗方法最新的发展步伐, 同时也要了解各种蒙特卡罗方法的性质和联系. 本书的主要目的就是为读者提供一个有关蒙特卡罗方法的自成体系的、统一的和最新的处理模式.

本书主要面向三类读者: 一是专门从事蒙特卡罗算法研究的科研人员; 二是对应用先进的蒙特卡罗技术感兴趣的科学家; 三是想学习蒙特卡罗计算的统计学、计算生物学和计算机科学专业的研究生. 要了解本书所述的方法至少必须具备下列知识: 一个学期的概率理论课程 (Pitman 1993) 和一个学期的理论统计课程 (Rice 1994). 这两门课程的掌握都只需达到大学水平即可. 当然, 如果读者具有诸如人工智能、计算生物学、计算机视觉、工程学或者涉及繁重计算的贝叶斯统计等某一特定科学领域的一些背景知识, 那就更为理想了. 本书特别适合作为大学高年级或研究生学习有关蒙特卡罗方法课程的教材. 该书重点阐述了蒙特卡罗方法与科学和统计研究的关系.

在此, 特别感谢我的导师益友王永雄教授, 他为本书提出了许多宝贵的意见. 感谢他对蒙特卡罗方法和科学问题的极大热情以及一直以来对我的鼓励! 感谢 Persi Diaconis 教会了我包括马尔可夫链理论、群论和非参数贝叶斯方法等许多知识! 感谢 Susan Holmes 和 Persi 对马尔可夫链蒙特卡罗和其他相关问题的有启发性的谈话! 感谢 Donald B. Rubin 关于缺失数据的系统和贝叶斯思想的深入了解! 感谢 Jonathan Goodman 关于多次蒙特卡罗的有益的评论! 感谢 UCLA 的吴英年和朱颂纯关于晶格模拟的材料和条件抽样的思想! 感谢梁发明为本书提供了许多例子和图表! 感谢陈明辉和David van Dyk有建议性的评论. 感谢曾经在斯坦福大学和哈佛大学统计系学习过的几位研究生: 陈玉国, 陈玲语, Chiara Sabatti, Tanya Logvinenko, 秦朝辉和张俊尼, 他们对本书的出版做出了不同程度的贡献. 感谢 Helen Tombropoulos 女士为本书和

我早期出版的一些论文的编辑提供了帮助. 最后, 非常感谢妻子魏的爱以及她多年来对我的研究活动给予了一如既往的支持! 本书部分章节是作者在斯坦福大学统计系执教时所写. 本书的出版也得到了美国国家自然科学基金委员会 (DMS-9803649 和 DMS-0094613) 的部分资助.

刘军于美国麻州剑桥市

2001 年 3 月

译者的话

2007 年 8 月在云南大学召开国际统计前沿会议——高维数据分析期间, 中科院数学与系统科学研究院副院长陈敏研究员告诉我们, 他们希望在国内出版刘军教授在 Springer-Verlag 出版社出版的专著《科学计算中的蒙特卡罗策略》的翻译版, 并希望我们能承担此次翻译工作. 由于我们多年来一直都在研读此书并且基于此书的理论和方法成功地解决了一些复杂模型的参数估计和模型选择等问题, 同时我们也对刘军教授的学识和为人甚为钦佩, 因此, 我们当即就接受了这项翻译工作并希望能早日将此书的翻译版推荐给国内读者.

蒙特卡罗统计模拟方法 (简称为 MCMC 方法) 是 20 世纪 40 年代中期由于科学技术的发展和电子计算机的发明而提出的一种以概率统计理论为指导的一类非常重要的数值计算方法, 也是一种用于解决数值问题的基于计算机的统计抽样方法.目前, MCMC 方法已广泛用于诸如生物信息学、生物遗传学、物理学、计算机科学、材料科学、工程技术学、教育心理学、社会学、金融学和经济学等领域. 很多科学家、工程师和应用工作者都把蒙特卡罗统计模拟方法作为他们工作中最基本的最优化方法. 尽管如此, 但据我们所知目前国内仅有零星地介绍 MCMC 方法的论文和一些教材, 却没有一本全面且系统地介绍 MCMC 方法的中文教程. 而美国哈佛大学统计系和生物统计系终身教授、世界生物统计和生物信息学领域的著名专家、COPSS "总统奖" 得主刘军教授的专著《科学计算中的蒙特卡罗策略》不仅详细全面地介绍了 MCMC 方法的基本原理、序贯蒙特卡罗理论、行为中的序贯蒙特卡罗方法、Metropolis 算法及其推广、Gibbs 抽样、一般条件抽样、杂交蒙特卡罗方法等, 还重点分析了这些理论和方法在物理、生物和化学等领域的具体应用, 同时辅以大量的模拟研究结果和习题以帮助读者学以致用. 该专著内容丰富、思路清晰、层次分明、深入浅出、实用易读. 自 2001 年在 Springer-Verlag 出版社出版以来得到了国内外从事 MCMC 方法的理论研究者和实际应用者的高度评价. 许多专家和学者都将该专著推荐给自己的学生, 作为他们学习 MCMC 方法的入门教材. 因此,这是一本介绍 MCMC 方法的难得的好教材. 我们希望本书中文版的出版能弥补国内没有系统介绍 MCMC 方法的书的不足,也希望该书的出版能引起国内对 MCMC 方法感兴趣的理论研究者和实际应用者对此方法研究的热潮.

本书的翻译工作由唐年胜负责组织协调,是集体协作的结晶.其中周勇翻译了第 1~4 章, 唐年胜翻译了前言、第 5~8 章、第 13 章和附录、索引, 徐亮翻译了第 9~12 章. 唐年胜负责全书

的统稿工作和初核工作, 最后由刘军教授对全书进行校对. 在本书的翻译过程中我们得到了许多人的热情帮助. 在此, 我们要特别感谢原著者刘军教授、云南大学前任校长王学仁教授和中科院陈敏研究员自始至终对本书翻译工作的关心和支持! 香港浸会大学数学系邓文礼副教授也对本译著的初稿提出了许多中肯的意见, 在此我们一并向他们表示衷心的谢意. 我们还要感谢高等教育出版社的编辑们对本译著的出版所付出的辛勤工作,特别要感谢王丽萍同志对本译著的审定与出版给予的大力支持和帮助! 本书的翻译也得到了上海财经大学统计系硕士研究生袁媛、时秩、赵徽、朱亦兰和云南大学统计系硕士研究生徐登可、王倩、赵远英、丁宁、吴雪佳、杨洋的大力帮助, 他们分别帮我们进行了录入和部分章节的初译, 在此我们对他们表示衷心的谢意. 另外, 还要感谢参与校对工作的刘军教授的博士生范晓丹 (现在香港中文大学统计系任助理教授).

由于译者水平有限, 本译著中肯定存在不少的缺点和不妥之处, 敬请同行专家、学者和广大读者给予批评指正.

云南大学　唐年胜
中国科学院　周　勇
东南大学　徐　亮
2008 年 10 月

目 录

第 1 章

引言与实例

1.1 对蒙特卡罗技术的需求

许多科学问题的实质就是求积分

$$I = \int_D g(\mathbf{x}) d\mathbf{x},$$

其中 D 通常是一个高维空间中的区域, $g(\mathbf{x})$ 为有兴趣的目标函数. 如果我们能借助计算机从区域 D 中均匀地抽得独立同分布 (i.i.d.) 的随机样本 $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}$, 则可由下式得到 I 的近似值

$$\hat{I}_m = \frac{1}{m}\left\{g(\mathbf{x}^{(1)}) + \cdots + g(\mathbf{x}^{(m)})\right\}.$$

根据*大数定律*, 具有相同期望和有限方差的独立随机变量的平均值收敛于其共同的均值 (见附录), 即

$$\lim_{m\to\infty} \hat{I}_m = I, \text{ 以概率1收敛.}$$

中心极限定理 (CLT) 可用来评价其收敛速度:

$$\sqrt{m}(\hat{I}_m - I) \to N(0, \sigma^2), \text{ 依分布收敛,}$$

其中 $\sigma^2 = \mathrm{var}\{g(\mathbf{x})\}$. 因此, 蒙特卡罗近似的 "误差项" 是 $O(m^{-1/2})$, 它与 $\mathbf{x}$ 的维数无关. 这一基本特征奠定了蒙特卡罗方法在科学和统计领域中的潜在作用.

在最简单的情况下, 如当 $D = [0, 1]$ 和 $I = \int_0^1 g(x)dx$ 时, 我们可用下式近似 I:

$$\tilde{I}_m = \frac{1}{m}\{g(b_1) + \cdots + g(b_m)\},$$

其中 $b_j = j/m$, 这一方法通常被称为*黎曼逼近*. 如果 g 为某一适当光滑函数, 则黎曼逼近的误差率为 $O(m^{-1})$, 此即表明黎曼近似比蒙特卡罗方法近似要好一些. 还有很多较复杂的方法, 如: Simpson 法则和 Newton-Cotes 法则, 由这些方法可得到更好的数值近似 (Thisted 1988). 然

而, 这些确定性方法的一个致命的缺点是当 D 的维数增加时由这些方法很难得到好的数值近似结果. 例如, 在一个 $D=[0,1]^{10}$ 的 10 维空间中, 我们需要计算 $O(m^{10})$ 个网格点才能获得 I 的黎曼近似的 $O(m^{-1})$ 的精度. 相比之下, 从区域 D 中均匀抽取样本 $x^{(1)},\cdots,x^{(m)}$ 的蒙特卡罗方法至少在理论上可达到 $O(m^{-1/2})$ 的误差率, 该误差率与 D 的维数无关.

尽管蒙特卡罗积分的误差率在高维问题中维持不变, 但存在两个实质性的困难: (a)当 D 是高维空间中的较大区域时, 用来度量函数 g 在区域 D 中均匀程度的方差 σ^2 的值将特别大; (b)在高维空间的任意区域 D 中, 我们或许不能产生均匀随机样本. 为了克服这些困难, *重点抽样*的思想常被用来从非均匀分布—其概率密度的重心在状态空间 D 的“重要”部分—$\pi(\mathbf{x})$ 产生随机样本 $\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(m)}$. 于是, 可用下式来估计积分 I:

$$\hat{\hat{I}}=\frac{1}{m}\sum_{j=1}^{m}\frac{g(\mathbf{x}^{(j)})}{\pi(\mathbf{x}^{(j)})},$$

其中 $\hat{\hat{I}}$ 的方差为 $\sigma_\pi^2=\mathrm{var}_\pi\{g(\mathbf{x})/\pi(\mathbf{x})\}$. 如果 g 非负且 I 有限, 则可选 $\pi(\mathbf{x})\propto g(\mathbf{x})$, 基于此选择可得 I 的精确估计. 然而, 在蒙特卡罗方法的应用中上述情形几乎是不可能发生的. 更现实的想法是, 希望能找到一个好的“备选” π, 该备选分布有较大的概率在高 g 值区域. 此时, 从 π 中抽取随机样本是一个非常富有挑战性的问题.

在生物信息学、计算化学、物理学、结构生态学和统计学等领域的其他问题研究中常常需要从非均匀分布 π 中抽样. 在这些问题的研究中, 复杂系统的概率分布 $\pi(\mathbf{x})$ 来自于物理学和统计推断中的基本定律, 其中 $\mathbf{x}$ 常被称为*系统状态* (configuration). 例如, 在大分子研究中, $\mathbf{x}$ 或许表示分子中所有原子以三维坐标系形式表示的分子*结构*. 取目标概率分布为波尔兹曼分布 $\pi(\mathbf{x})=Z(T)e^{-h(\mathbf{x})/kT}$, 其中 k 是波尔兹曼常数, T 是系统温度, $h(\mathbf{x})$ 是能量函数, 而 $Z(T)$ 表示难以计算的*配分函数* (partition function). 科学家们的兴趣在于系统的某种“平均特征”, 一些平均特征在数学上可表示为 $E_\pi[g(\mathbf{x})]$, 其中 g 为某个适合的函数. 在贝叶斯统计推断中, $\mathbf{x}$ 常常表示缺失数据和参数值的联合状态, $\pi(\mathbf{x})$ 通常表示这些变量的后验分布. 为了对有兴趣的参数做合理的统计推断和对未来观测做有效预测, 我们不得不将多余参数和缺失数据从积分中积掉. 这些工作可看成是计算状态空间中的函数的期望.

有时, 最优化问题可阐述为蒙特卡罗抽样问题. 假设我们想找复杂状态空间中的目标函数 $h(\mathbf{x})$ 的最小值. 这个问题等价于找另一个函数 $q_T(\mathbf{x})=e^{-h(\mathbf{x})/T}$ (当 $T>0$ 时)的最大值. 此时, 当对所有 $T>0$ 函数 $q_T(\mathbf{x})$ 均可积时(这一情况在实际问题中是常见的), 我们可构造一个概率分布族:

$$\pi_T(\mathbf{x})\propto e^{-h(\mathbf{x})/T},\quad T>0.$$

当 T 充分小时, 从 $\pi_T(\mathbf{x})$ 中抽得的随机样本最有可能落在 $h(\mathbf{x})$ 的全局最小值点的附近. 这一思想是著名的模拟退火算法 (Kirkpatrick 等人1983) 的基础, 也是设计更有效的蒙特卡罗算法的*回火* (tempering) 技术的关键(参见第 10 章).

1.2 全书的范围及概要

蒙特卡罗方法的基本步骤是从概率分布函数 $\pi(\mathbf{x})$ 中产生随机 (伪随机) 样本, 其中概率分布函数的形式除一个归一化常数 (normalizing constant) 外全已知. 变量 $\mathbf{x}$ 通常在 $\mathbb{R}^k$ 中取值, 但偶尔也会在诸如置换或变换群 (Diaconis 1988, Liu 和 Wu 1999) 等其他空间中取值. 在大多