

哈密顿蒙特卡洛采样算法在贝叶斯统计教学中的探索

赵 谦

西安交通大学数学与统计学院, 陕西 西安

收稿日期: 2025年7月30日; 录用日期: 2025年8月29日; 发布日期: 2025年9月10日

摘 要

贝叶斯统计是统计学的重要专业课程, 而后验计算是本课程的重要组成部分。现有教材主要聚焦传统马尔可夫链蒙特卡洛(Markov Chain Monte Carlo, MCMC)采样算法, 缺少对于一些新近发展起来的、计算效果更好的MCMC采样算法的讨论。为此, 本文探讨了将哈密顿蒙特卡洛(Hamiltonian Monte Carlo, HMC)采样算法引入课程体系, 以期提升教学内容的前沿性与实用性, 并锻炼学生的编程实践能力。从教学实践看, 这一尝试取得了良好的教学效果, 有效契合了人工智能背景下统计学人才的培养需求。

关键词

贝叶斯统计, 后验计算, 马尔可夫链蒙特卡洛, 哈密顿动力系统

Exploring Hamiltonian Monte Carlo Sampling in Bayesian Statistics Teaching

Qian Zhao

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an Shaanxi

Received: Jul. 30th, 2025; accepted: Aug. 29th, 2025; published: Sep. 10th, 2025

Abstract

Bayesian Statistics is a crucial specialized course in statistics, and posterior computation constitutes a significant component of this curriculum. Existing textbooks primarily focus on traditional Markov Chain Monte Carlo (MCMC) sampling algorithms, lacking discussion of some recently developed, more computationally effective MCMC sampling algorithms. To address this, this paper explores incorporating the Hamiltonian Monte Carlo (HMC) sampling algorithm into the course structure. This aims to enhance the course content's relevance to current advancements and practical

文章引用: 赵谦. 哈密顿蒙特卡洛采样算法在贝叶斯统计教学中的探索[J]. 教育进展, 2025, 15(9): 600-611.

DOI: 10.12677/ae.2025.1591713

utility, while also strengthening students' practical programming skills. Teaching practice demonstrates that this initiative has achieved positive educational outcomes, effectively aligning with the training objectives for statistics professionals in the era of artificial intelligence.

Keywords

Bayesian Statistics, Posterior Computation, Markov Chain Monte Carlo, Hamiltonian Dynamics

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

贝叶斯统计是统计学专业的一门专业课,作为非经典统计学课程,为基于频率思想的经典统计课程提供了重要补充。贝叶斯统计起源于英国学者托马斯·贝叶斯(Thomas Bayes, 约 1701~1761)的一篇在他死后才发表的一篇论文《论有关基于问题的求解》(“An Essay Towards Solving a Problem in the Doctrine of Chances”)。在这篇论文中,他提出了著名的贝叶斯公式并建立贝叶斯统计思想的雏形。此后,虽然有拉普拉斯(P. Laplace)等数学家在其基础上做出了一些有意义的工作,但贝叶斯统计方法并未被普遍接受。20 世纪中叶后,随着方法和理论上的不断完善,以及在实际问题中的成功应用,贝叶斯学派逐渐发展为一个有重要影响的统计学派。

应用贝叶斯统计方法的关键在于后验分布的计算,而计算后验分布在实际问题中通常是困难问题。现有的本科贝叶斯统计教材中通常也会花费一定的篇幅专门介绍计算后验分布的方法,主要讨论基于蒙特卡洛马尔科夫链(Markov Chain Monte Carlo, MCMC)采样的方法。然而,现有教材针对 MCMC 采样算法的讨论主要针对较为经典的 Metropolis-Hastings 算法,而缺少对于一些新近发展起来的、计算效果更好的 MCMC 采样算法的介绍。对此,笔者在教学实践中尝试引入哈密顿蒙特卡洛(Hamiltonian Monte Carlo, HMC)采样方法[1] [2],并布置相应的编程作业,力求扩展学生在贝叶斯统计方面的相关知识,提升实践技能。此外,由于 HMC 相关算法在机器学习中有广泛应用,该部分教学内容的引入也呼应了人工智能背景下统计学专业的培养需求。

从更宏观的角度出发,贝叶斯后验计算属于统计计算的范畴。笔者注意到近年来,关于统计计算已经有了一些教学研究[3]-[7],但对于贝叶斯后验计算的教学尚无专门讨论。因此,本文的探索在某种意义上也填补了这方面的空白。

2. 贝叶斯后验计算

在贝叶斯统计中,总体分布中的参数(或参数向量)被看作随机变量,一切的统计推断与统计决策均基于对参数的后验分布进行,因此有效计算后验分布是应用贝叶斯统计方法的先决条件。

假设总体 X 的概率密度(方便起见,以下讨论均假定总体和参数为连续型随机变量,离散情形可相应推广)为 $f(x|\theta)$, 其中 θ 为总体分布的参数,其先验概率密度为 $\pi(\theta)$, $\mathbf{x}=(x_1, x_2, \dots, x_n)$ 是来自总体分布的样本。根据贝叶斯公式, θ 的后验概率密度 $\pi(\theta|\mathbf{x})$ 可通过下式进行计算:

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int_{\theta \in \Theta} f(\mathbf{x}|\theta)\pi(\theta)d\theta}, \quad (1)$$

其中, $f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)$ 为似然函数。上式中, 当先验分布与总体分布确定后, 分子部分容易计算。但分母部分由于涉及积分, 通常难以计算, 特别是在参数为高维向量的情况下。

为此, 在实际应用中, 通常会采用某种方式来“逼近”真正的后验分布。最为常用的有两类方法, 即近似推断与采样。近似推断的核心思想是首先选取特定的分布族来作为后验分布的近似, 然后在某种准则下来优化分布族中的参数, 例如常用的变分推断的优化准则为近似后验与真实后验之间的 KL 距离。该类方法的优势是计算效率较高, 但由于需要取定近似后验的形式, 通常不能保证计算结果收敛到真实后验分布。采样方法的核心思想是不直接计算后验分布, 而试图获得(近似)服从真实后验分布的样本, 以样本的经验分布刻画真实后验分布。该类方法通常有较好的理论性质, 例如其中最具代表性的 MCMC 采样算法, 可在一定条件下保证所构造的马尔可夫链的平稳分布就是所求的后验分布, 但其缺点是计算代价较大。

现有贝叶斯统计教材(如文献[8] [9])在讨论后验分布计算时, 大多主要以 Metropolis-Hastings 系列算法为代表介绍 MCMC 采样算法, 例如 Gibbs 采样等。但正如前面所讨论的, 这些方法虽然理论性质较好, 但计算代价较大。此外, 从实践上看, MH 系列算法在一些情况下的收敛性质并不理想。因此, 有必要在教学中引入新近发展起来的, 计算效果更好的采样算法。

3. 哈密顿蒙特卡洛采样算法与教学设计

3.1. 哈密顿蒙特卡洛采样算法简介

哈密顿蒙特卡洛(Hamiltonian Monte Carlo, HMC)采样算法基于附录 A.1 节中介绍的哈密顿动力系统(Hamiltonian Dynamics 或 Hamiltonian System)构造。其基本思想是将待采样的参数向量看作哈密顿系统中的位置向量, 再引入动量向量作为辅助, 然后通过“运行”动力系统(通常通过附录 A.3 节介绍的离散化的 Leapfrog 迭代实现)产生新的采样候选点, 最后计算接收概率决定是否将采样候选点作为新的采样点。具体算法如表 1 所示。

Table 1. (Discrete) Hamiltonian Monte Carlo sampling

表 1. (离散化)哈密顿蒙特卡洛采样

算法 1: 离散化哈密顿蒙特卡洛(HMC)采样

输入: 样本 $\mathbf{x}$, $U(\theta) = -\ln f(\mathbf{x}|\theta)\pi(\theta)$, $K(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T\mathbf{M}^{-1}\mathbf{z}$, Leapfrog 步长 ε 、步数 L

输出: 以 $\pi(\theta|\mathbf{x})$ 为平稳分布的马氏链采样 $\{\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(t)}, \dots\}$

- 1: 产生初始 $\theta^{(0)}$, $t \leftarrow 0$
- 2: 重复下面过程, 直至马氏链接近平稳状态
- 3: 从 $\mathcal{N}(0, \mathbf{M})$ 中产生 $\mathbf{z}^{(t)}$
- 4: 由 $(\theta^{(t)}, \mathbf{z}^{(t)})$ 以 ε 为步长执行 L 步 Leapfrog 迭代于 $H(\theta, \mathbf{z}) = U(\theta) + K(\mathbf{z})$ 至 $(\theta', \mathbf{z}')$
- 5: 以 $\alpha(\theta^{(t)}, \mathbf{z}^{(t)}, \theta', \mathbf{z}') = \min\left\{1, \exp\left(-H(\theta', \mathbf{z}') + H(\theta^{(t)}, \mathbf{z}^{(t)})\right)\right\}$ 的概率接受 $(\theta', \mathbf{z}')$
- 6: 若接受 θ' , 则令 $\theta^{(t+1)} = \theta'$, 否则令 $\theta^{(t+1)} = \theta^{(t)}$
- 7: $t \leftarrow t+1$

HMC 算法中采样候选点的产生有着一定的物理学背景, 这与传统 MH 系列的 MCMC 算法有所不同。但正如附录 A.4 节中所讨论的, 在一定条件下, HMC 算法仍然可以看作作为一种特殊的 MH 算法。如第 4 节所展示的, 相比于传统 MH 算法, HMC 算法有着更好的采样效果, 但由于其产生候选点的过程需

要多步迭代，计算复杂度相对较高。因此，在追求计算效率而对计算精度要求相对不高的情况下，可按照附录 A.5 节所讨论的，将其简化为朗之万蒙特卡洛(Langevin Monte Carlo, LMC)采样。基于这一考虑，如果进一步引入随机梯度的思想，就可得到目前在贝叶斯深度学习计算中广泛应用的 SGLD 算法[10]。

3.2. 教学设计与策略

虽然可以看作一种特殊的 MH 算法，但 HMC 采样算法构造的出发点与传统 MH 算法有较大不同，因此不宜直接放在 MH 算法框架下进行讲授。因此，笔者在教学实践中采用三步走的策略，即首先介绍算法构造过程并给出 HMC 采样算法，然后从理论上讨论其与 MH 算法的关系，最后介绍对其的拓展。

笔者会首先介绍用来刻画哈密顿动力系统的哈密顿方程组(见附录 A.1 节)及其相关性质，这对于有微分方程学习基础数学类高年级本科生而言是能够接收的。此外，由于哈密顿动力系统的物理学背景，笔者还会给出可以由哈密顿方程组刻画的物理过程的动画演示，以吸引学生的兴趣。在有了对哈密顿动力系统的初步认识后，再讲授基于微分方程组的解轨线产生采样候选点的思想及其具体方案，给出附录 A.2 节讨论的理想 HMC 采样，并指出其“完美”的采样效率(即候选点接收概率为 1)。进一步地，指出该采样过程事实上无法实现，需要对微分方程组进行离散化处理，并对比集中离散化策略(如欧拉法与改进欧拉法等)，给出效果最佳的 Leapfrog 方案，最终得到实际中最常使用的 HMC 采样算法(算法 1)。

在完整介绍 HMC 采样算法后，可基于附录 A.4 节的内容，分析 HMC 采样算法与 MH 算法的关系，得出在一定条件下可将 HMC 采样算法看作一种特殊的 MH 算法的结论。注意到这一部分理论性较强，且并不影响算法实现，因此若从实用角度出发，可以作为选讲内容，不作为对所有学生的普遍要求。

最后，可对 HMC 采样算法进行一些评论，指出其优缺点，并针对其计算复杂度较高的缺点讨论附录 A.5 节给出的拓展方向，即 LMC 采样以及更进一步的 SGLD 算法。这一部分的教学注意传递一种思想，即没有一种算法在所有方面都是最好的，在科学研究和工程实践中往往需要根据具体问题和实际需求进行权衡，寻求最适合的算法。

4. 应用实例

在本节中，我们将通过实例展示哈密顿蒙特卡洛采样的计算效果，并与几种典型的 MH 算法进行对比。

4.1. 问题与建模

本问题来自文献[8]的第 7.7 节，背景是使用 Logistic 模型对老年痴呆症数据进行分析。具体数据如表 2 所示，其中 x_i, y_i 分别表示第 i 个人的智力水平(取值范围为 0 到 20)和是否患有老年痴呆症(1 表示是，0 表示否)。据此可建立如下的 Logistic 模型：

Table 2. The intelligence test scores of 51 older adults [8]
表 2. 51 位老年人智力测验成绩[8]

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	9	1	19	11	0	37	9	0
2	13	1	20	14	0	38	11	0
3	6	1	21	15	0	39	14	0
4	8	1	22	18	0	40	10	0
5	10	1	23	7	0	41	16	0

续表

6	4	1	24	16	0	42	10	0
7	14	1	25	9	0	43	16	0
8	8	1	26	9	0	44	14	0
9	11	1	27	11	0	45	13	0
10	7	1	28	13	0	46	13	0
11	9	1	29	15	0	47	9	0
12	7	1	30	13	0	48	15	0
13	5	1	31	10	0	49	10	0
14	14	1	32	11	0	50	11	0
15	13	0	33	6	0	51	12	0
16	16	0	34	17	0	52	4	0
17	10	0	35	14	0	53	14	0
18	12	0	36	19	0	54	20	0

$$y_i \sim \text{Bernoulli}(\theta_i), \theta_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}, i = 1, 2, \dots, 54.$$

取 (β_0, β_1) 的先验为

$$\beta_j \sim \mathcal{N}(\mu_j, \sigma_j^2), j = 0, 1,$$

其中, 令 $\mu_j = 0$, σ_j^2 为很大的正数(如 10,000), 使 (β_0, β_1) 的先验接近平坦的无信息先验。

在上述建模下, 关于 (β_0, β_1) 的后验分布可表示为

$$\begin{aligned} \pi(\beta_0, \beta_1 | y) &\propto f(y | \beta_0, \beta_1) \pi(\beta_0, \beta_1) \\ &\propto \prod_{i=1}^{54} \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1 - y_i} \exp \left\{ -\frac{(\beta_0 - \mu_0)^2}{2\sigma_0^2} - \frac{(\beta_1 - \mu_1)^2}{2\sigma_1^2} \right\}. \end{aligned} \tag{2}$$

上述后验分布难以得到解析表达, 因此使用 MCMC 采样算法对其进行近似计算。

4.2. 计算效果

本文使用第 3 节讨论的 HMC 采样算法对式(2)中的后验分布进行计算, 其中每次采样中 Leapfrog 迭代的步长设为 $\varepsilon = 0.05$, 迭代次数设为 $L = 20$, 并与文献[8]第 7.7 节使用的 4 种 MH 采样算法进行了对比。所有方法均执行 60,000 次采样, 并取最后 12,000 个采样点用于对后验分布的近似。

图 1 和图 2 给出了不同方法的采样结果, 包括近似后验分布散点图以及直方图。可以看出, HMC 采样和以多元正态分布为建议分布的 MH 采样的后验分布直方图呈现出较好的单峰性质, 表明其良好的采样效果。对比这两种方法的后验分布散点图, 可以看出 HMC 采样得到的离群采样点相对更少。相较而言, 逐分量 MH 采样和切片 Gibbs 采样的效果稍差, 而随机游动 Metropolis 采样的效果最差, 这从相应的后验分布直方图可以看出。

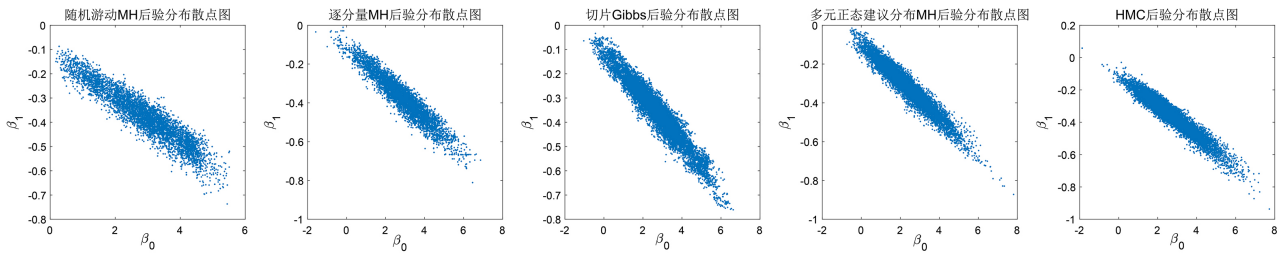


Figure 1. The scatter plots of the approximated posteriors by different MCMC sampling algorithms

图 1. 不同 MCMC 采样算法得到的近似后验分布散点图

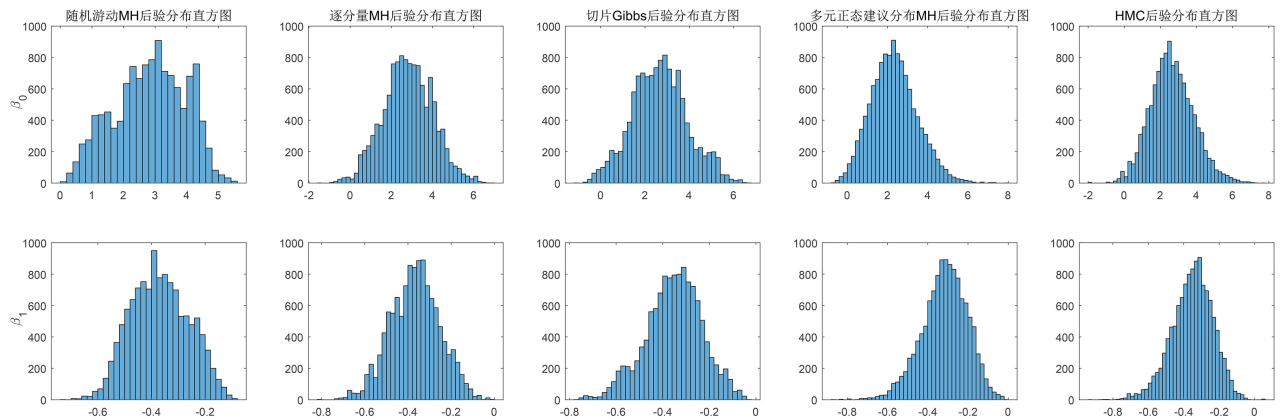


Figure 2. The histograms of the approximated posteriors by different MCMC sampling algorithms

图 2. 不同 MCMC 采样算法得到的近似后验分布直方图

图 3 和图 4 进一步通过样本路径和遍历均值对比不同方法的收敛特性。可以看出，在 5 种方法中，HMC 采样通过最少的采样次数就可以使遍历均值达到最为平稳的状态，表明其最佳的收敛性。相较而言，以多元正态分布为建议分布的 MH 采样和逐分量 MH 采样的收敛性质尚可，而随机游动 Metropolis 采样和切片 Gibbs 采样则较差。从样本路径图可以看出，随机游动 Metropolis 采样呈现出一定的趋势性，说明其采样点间的相关性较强，这是在对后验分布的近似计算中所不希望出现的现象。在这一点上，HMC 采样的表现最好，其样本路径没有呈现出任何的趋势性。

从上述结果可以看出，HMC 采样算法获得了最佳的采样结果，表明其优于传统 MCMC 算法的计算性能。

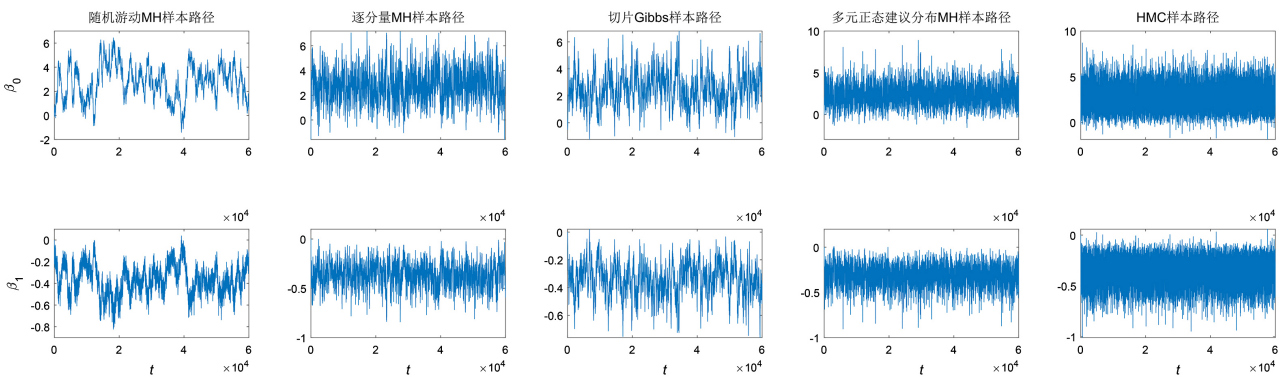


Figure 3. The sampling paths of different MCMC sampling algorithms

图 3. 不同 MCMC 采样算法的样本路径

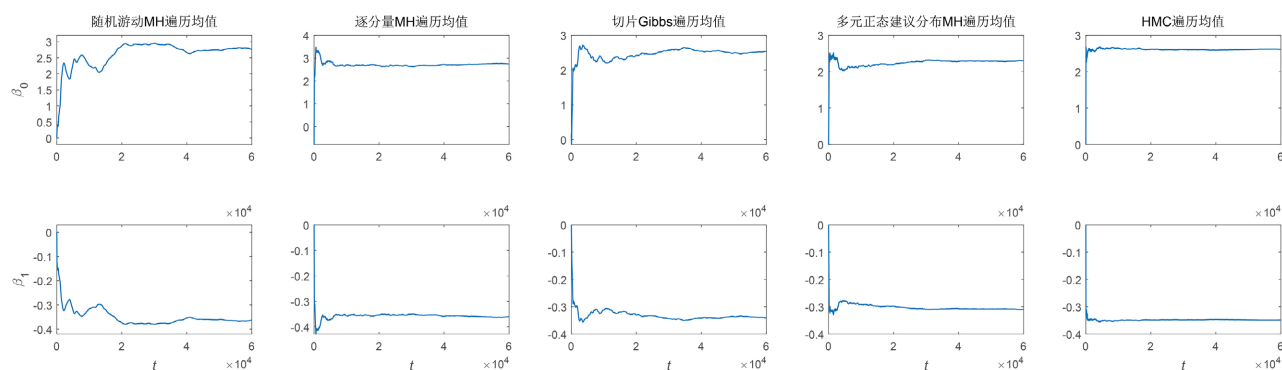


Figure 4. The ergodic means of samples by different MCMC sampling algorithms

图 4. 不同 MCMC 采样算法的样本遍历均值

5. 教学效果

笔者自 2021 年开始, 将哈 HMC 采样算法作为贝叶斯统计课程的教学内容, 目前已经 5 次开课。从授课过程看, 学生对该部分内容比较感兴趣, 并且能够较为充分地理解该算法的构造原理。在讲授算法原理后, 笔者会介绍第 4 节的应用实例, 并展示计算结果, 然后将其编程实现作为实践作业由学生完成。这样, 学生可以以第 4.2 节展示的计算结果作为“参考答案”来验证编程实现的正确与否。表 3 统计了自授课以来, 学生对于该算法的实现情况。可以看出, 大部分学生能够独立实现并成功运行算法, 而 2/5 以上的学生得到了较为准确的采样结果。图 5 给出了一些学生所得到的采样结果。可以看出, 这些采样结果与图 1~4 展示的结果类似。总体而言, 对于 HMC 采样算法的教学实践取得了良好的效果。

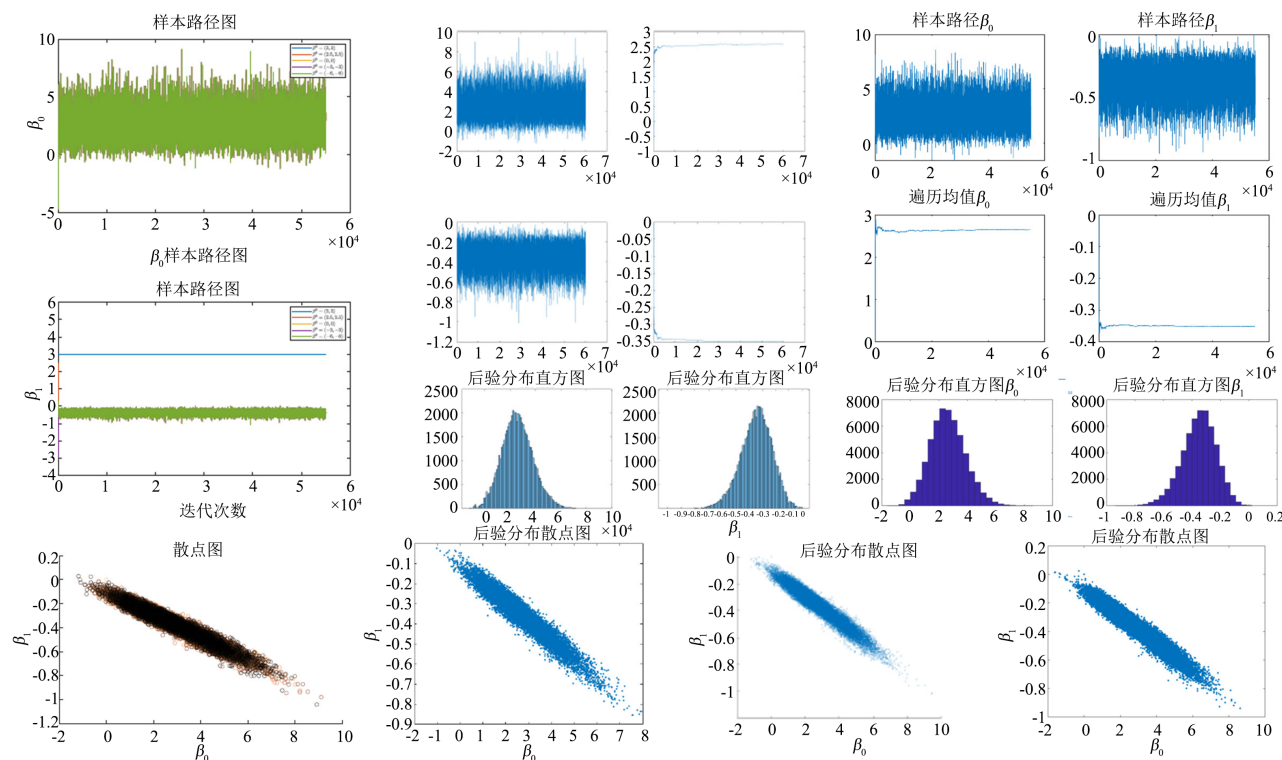


Figure 5. Example results obtained by students

图 5. 学生实现结果示例

Table 3. Overall student completion status of the HMC programming assignments since first teaching
表 3. 授课以来学生完成 HMC 编程实践作业的情况

授课年份	选课人数	独立实现人数	独立实现比例	结果准确人数	结果准确比例
2021	31	15	48.39%	9	29.03%
2022	27	18	66.67%	14	51.85%
2023	17	13	76.47%	9	52.94%
2024	22	13	59.09%	7	31.82%
2025	15	12	80.00%	8	53.33%
合计	112	71	63.39%	47	41.96%

6. 结语

本文探讨了在贝叶斯统计课程教学中引入哈密顿蒙特卡洛(Hamiltonian Monte Carlo, HMC)采样, 以实现后验分布近似计算的教学实践。相较于传统 MCMC 采样算法, HMC 采样得到的马氏链能够更快收敛到接近平稳的状态, 且其采样结果对于后验分布的近似也更为合理, 展现了在实际问题中良好的应用潜力。从教学效果看, 大部分学生能够从原理上理解并掌握该算法, 部分学生能够在编程实践中利用其获得良好的计算结果。笔者相信, 引入 HMC 采样可强化贝叶斯统计课程与学科发展前沿的联系, 提升教学内容的实用性, 同时有助于锻炼学生的实践动手能力, 更好地适应人工智能背景下的统计学人才培养需求。

基金项目

本研究得到国家自然科学基金项目(编号: 12471485)的资助。

参考文献

[1] Neal, R.M. (2011) MCMC Using Hamiltonian Dynamics. In: Brooks, S., Gelman, A., Jones, G.L. and Meng, X.L., Eds., *Handbook of Markov Chain Monte Carlo*, Chapman and Hall/CRC, 113-162. <https://doi.org/10.1201/b10905-6>

[2] Betancourt, M. (2018) A Conceptual Introduction to Hamiltonian Monte Carlo.

[3] 吴连伟, 张磊, 刘蒲胜. 统计计算语言学课程的思政元素设计与教学实践[J]. 计算机教育, 2023(1): 48-52.

[4] 陆靖, 张兆远. 数字化时代本科统计计算课程思政面临的挑战与对策[J]. 高教学刊, 2024, 10(11): 186-189.

[5] 刘柏森. 统计计算课程过程化考核模式的探索与实践[J]. 长春师范大学学报, 2024, 43(8): 136-140.

[6] 周影, 马维军, 佟良, 袁超凤, 谢婉秋. 新时代本科《统计计算》课程思政建设的思考与探索[J]. 统计与咨询, 2024(6): 32-35.

[7] 常振海, 刘薇, 王昱禾, 王丙参. 数据科学时代《统计计算》课程改革: 目标重构、教法创新与效果评价——以地方应用型本科院校为例[J]. 天水师范学院学报, 2025, 45(3): 109-114.

[8] 茆诗松, 汤银才. 贝叶斯统计[M]. 第 2 版. 北京: 中国统计出版社, 2012.

[9] 韦来生, 张伟平. 贝叶斯分析[M]. 第 2 版. 合肥: 中国科学技术大学出版社, 2021.

[10] Welling, M. and Teh, Y.M. (2011) Bayesian Learning via Stochastic Gradient Langevin Dynamics. In: Getoor, L. and Scheffer, T., Eds., *Proceedings of the 28th International Conference on International Conference on Machine Learning*, Omnipress, 681-668.

附录

本附录详细介绍哈密顿蒙特卡洛采样算法，作为正文的补充。

A.1. 哈密顿动力系统

哈密顿动力系统(Hamiltonian Dynamics 或 Hamiltonian System)来源于物理中的哈密顿力学，由如下微分方程组(哈密顿方程组)给出：

$$\begin{cases} \frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \end{cases} \quad (\text{A1})$$

其中， $q = (q_1, \dots, q_d)^T$ 为 d 维位置向量， $p = (p_1, \dots, p_d)^T$ 为 d 维动量向量， $H(q, p, t)$ 为哈密顿量。如果 H 不显式依赖于 t ，则有 $H(q, p, t) = H(q, p)$ 。此时，根据哈密顿方程组，容易验证 $\frac{dH}{dt} = 0$ ，即 H 关于 t 是不变的，在物理上可以理解为某封闭系统的总能量。此外，还可以证明，对于任意 s ，映射 $T_s : (q(t), p(t)) \mapsto (q(t+s), p(t+s))$ 为一一映射，也就是说系统状态关于 t 是可逆的。

一种典型的哈密顿系统由具有如下形式的哈密顿量给出：

$$H(q, p) = U(q) + K(p), \quad (\text{A2})$$

其中， $U(q)$ 仅由位置向量决定，可理解为系统在当前状态下的势能；而 $K(p)$ 仅由动量向量确定，可理解为系统在当前状态下的动能。在后文中，具有该形式的哈密顿量是哈密顿蒙特卡洛采样算法的构造基础。

A.2. 理想哈密顿蒙特卡洛采样

哈密顿动力系统看似与采样算法并无直接关系，但如果从统计物理的角度出发，将刻画系统状态的 (q, p) 看作随机变量，则可根据系统能量(即哈密顿量)定义系统状态的概率分布：

$$P(q, p) = \frac{1}{Z} \exp\left(-\frac{H(q, p)}{T}\right), \quad (\text{A3})$$

其中， Z 是使得上式右端为分布(概率密度)的归一化常数因子， T 为系统的温度参数。进一步取 H 为式(A2)的形式，并取 $T=1$ ，则有

$$P(q, p) = \frac{1}{Z} \exp(-U(q)) \exp(-K(p)), \quad (\text{A4})$$

并且可以看出关于 q 的边缘分布为

$$P(q) \propto \exp(-U(q)). \quad (\text{A5})$$

这一观察启发我们利用哈密顿动力系统的不同状态来产生来自目标分布，例如产生来自贝叶斯后验分布的样本。事实上，以所关注的参数 θ 作为位置向量代替 q ，另以辅助变量 z 作为动量向量代替 p ，若取

$$U(\theta) = -\ln[f(\mathbf{x}|\theta)\pi(\theta)], \quad (\text{A6})$$

就可保证该哈密顿动力系统的任一状态所对应的位置 θ 来自于目标后验分布 $\pi(\theta|\mathbf{x})$ 。由此，可构造哈密顿蒙特卡洛(Hamiltonian Monte Carlo, HMC)采样算法，其一步采样过程可描述为：给定上一步采样所获得的采样点 $\theta^{(r)}$ ，首先随机产生辅助变量 $z^{(r)}$ ；然后在 $H(\theta, z) = U(\theta) + K(z)$ (其中 $U(\theta)$ 由式(A6)给

出)确定的哈密顿动力系统下, 将映射 T_s 作用于 $(\theta^{(\tau)}, z^{(\tau)})$ 获得新的候选采样点 (θ', z') , 即由 $(\theta^{(\tau)}, z^{(\tau)})$ 出发沿哈密顿方程组所确定的轨线向前运动步长 s ; 最后计算接受概率, 依概率判断是否接受候选点 (θ', z') , 若接受则令新的采样点 $\theta^{(\tau+1)} = \theta'$, 否则令 $\theta^{(\tau+1)} = \theta^{(\tau)}$, 同时丢弃 z' 。注意到在上述过程中, $K(z)$ 的形式并不影响采样结果, 因此可将其取为如下二次型的形式:

$$K(z) = \frac{1}{2} z^T M^{-1} z, \quad (\text{A7})$$

这样可使得辅助变量 $z \sim \mathcal{N}(0, M)$ 。实践中, 可进一步将 M 取为对角矩阵甚至单位矩阵, 以方便算法实现。

上述 HMC 采样过程可总结为算法 A1。对于该算法, 有两点需要说明。首先, 理想情况下, 算法第 4 行的执行需要确定位置和动量的轨线, 即求解哈密顿方程组, 这在实践中一般是不可行的, 因此我们将算法 A1 称为“理想 HMC 采样”。而在实践中, 这一步骤需要进行数值离散化, 从而导致下一小节将要讨论的“离散化 HMC 采样”。其次, 理想情况下, 根据哈密顿量 H 的不变性, 第 5 行所定义的接受概率实际上恒为 1, 但在离散化操作后, 这一概率不再恒为 1。事实上, 在离散化情形下, 这一概率本质上即为 Metropolis-Hastings (MH) 算法中的接受概率, 这一点将在第 A.4 节进一步讨论(表 A1)。

Table A1. Ideal Hamiltonian Monte Carlo sampling

表 A1. 理想哈密顿蒙特卡洛采样

算法 A1: 理想哈密顿蒙特卡洛(HMC)采样

输入: 样本 x , $U(\theta) = -\ln f(x|\theta)\pi(\theta)$, $K(z) = \frac{1}{2} z^T M^{-1} z$

输出: 以 $\pi(\theta|x)$ 为平稳分布的马氏链采样 $\{\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(t)}, \dots\}$

- 1: 产生初始 $\theta^{(0)}$, $t \leftarrow 0$
- 2: 重复下面过程, 直至马氏链接近平稳状态
- 3: 从 $\mathcal{N}(0, M)$ 中产生 $z^{(t)}$
- 4: 从 $(\theta^{(t)}, z^{(t)})$ 出发运行哈密顿动力系统 $H(\theta, z) = U(\theta) + K(z)$ 至 (θ', z')
- 5: 以 $\alpha(\theta^{(t)}, z^{(t)}, \theta', z') = \min\{1, \exp(-H(\theta', z') + H(\theta^{(t)}, z^{(t)}))\}$ 的概率接受 (θ', z')
- 6: 若接受 θ' , 则令 $\theta^{(t+1)} = \theta'$, 否则令 $\theta^{(t+1)} = \theta^{(t)}$
- 7: $t \leftarrow t+1$

A.3. 离散化哈密顿蒙特卡洛采样

如前面所讨论的, 哈密顿方程组通常难以解析求解, 因此算法 1 中的第 4 行无法精确执行, 需要使用数值算法进行离散化近似。最直接的对微分方程组进行离散化的方法是欧拉法, 即使用一阶差分对微分进行近似, 从而给出方程组的数值解轨线。然而, 正如文献[1]中所讨论的, 欧拉法的数值精度有限, 无法满足采样需求。因此, 通常使用 Leapfrog 算法进行近似, 其计算格式如下:

$$\begin{cases} z\left(t + \frac{l\varepsilon}{2}\right) = z\left(t + (l-1)\varepsilon\right) - \frac{\varepsilon}{2} \nabla U(\theta) \Big|_{\theta=\theta(t+(l-1)\varepsilon)} \\ \theta(t+l\varepsilon) = \theta(t+(l-1)\varepsilon) + \varepsilon M^{-1} z\left(t + \frac{l\varepsilon}{2}\right) \\ z(t+l\varepsilon) = z\left(t + \frac{l\varepsilon}{2}\right) - \frac{\varepsilon}{2} \nabla U(\theta) \Big|_{\theta=\theta(t+l\varepsilon)} \end{cases}, \quad (\text{A8})$$

其中, ε 为算法的步长, $l=1,2,\dots,L$ 为迭代步数, $\theta(t)=\theta^{(l)}$, $z(t)=z^{(l)}$ 。

将上述 Leapfrog 近似应用于算法 A1, 即可得到实际中使用的 HMC 采样, 即正文 3.1 节中的算法 1。

A.4. 与 Metropolis-Hastings 算法的关系

上述哈密顿蒙特卡洛采样算法看似与传统的 MH 算法有较大区别, 但其本质上可以看作一种特殊的 MH 算法。事实上, 在 MH 算法的观点下, 若给定前一步采样点 $(\theta^{(r)}, z^{(r)})$, 则候选点的接受概率可以由下式计算:

$$\alpha(\theta^{(r)}, z^{(r)}, \theta', z') = \min \left\{ 1, \frac{f(\theta', z') r(\theta^{(r)}, z^{(r)} | \theta', z')}{f(\theta^{(r)}, z^{(r)}) r(\theta', z' | \theta^{(r)}, z^{(r)})} \right\}, \quad (\text{A9})$$

其中, $f(\cdot, \cdot)$ 为 θ 和 z 的联合分布的概率密度, $r(\cdot | \cdot)$ 为建议分布的概率密度。进一步地, 根据上一小节的讨论, 在哈密顿蒙特卡洛采样算法中, 我们有

$$f(\theta, z) \propto \exp(-H(\theta, z)). \quad (\text{A10})$$

而由于候选点 (θ', z') 是通过确定性的 Leapfrog 算法由 $(\theta^{(r)}, z^{(r)})$ 产生的(不妨设经过 L 步 Leapfrog 迭代后得到的点为 $(\theta_L^{(r)}, z_L^{(r)})$, 即 $\theta' = \theta_L^{(r)}$, $z' = z_L^{(r)}$), 因此可认为

$$r(\theta', z' | \theta^{(r)}, z^{(r)}) = \delta(\theta' - \theta_L^{(r)}) \delta(z' - z_L^{(r)}) = 1, \quad (\text{A11})$$

即 $q(\cdot | \cdot)$ 确定性地刻画了 L 步 Leapfrog 迭代。而在此观点下,

$$r(\theta^{(r)}, z^{(r)} | \theta', z') = \delta(\theta^{(r)} - \theta_{2L}^{(r)}) \delta(z^{(r)} - z_{2L}^{(r)}) = 0, \quad (\text{A12})$$

这是因为对 (θ', z') 继续执行 L 步 Leapfrog 迭代后将得到 $(\theta_L^{(r)}, z_L^{(r)})$, 而非 $(\theta^{(r)}, z^{(r)})$, 从而导致接受概率 $\alpha(\theta^{(r)}, z^{(r)}, \theta', z')$ 恒为 0。但若对算法 1 做一点修改, 在算法 2 第 4 行得到候选点 (θ', z') 后对 z' 进行反向操作, 即令 $z' = -z_L^{(r)}$, 那么此时, 由于产生 (θ', z') 的过程仍然是确定性的, 可得到

$$r(\theta', z' | \theta^{(r)}, z^{(r)}) = \delta(\theta' - \theta_L^{(r)}) \delta(z' + z_L^{(r)}) = 1. \quad (\text{A13})$$

同时, 由于哈密顿动力系统及其 Leapfrog 近似的可逆性, 由 (θ', z') 出发进行 L 步 Leapfrog 迭代后将得到 $(\theta^{(r)}, z^{(r)})$, 从而有

$$r(\theta^{(r)}, z^{(r)} | \theta', z') = \delta(\theta^{(r)} - \theta^{(r)}) \delta(z^{(r)} - z^{(r)}) = 1. \quad (\text{A14})$$

将式(A13)和(A14)带入式(A9)后可得到

$$\alpha(\theta^{(r)}, z^{(r)}, \theta', z') = \min \left\{ 1, \exp(-H(\theta', z') + H(\theta^{(r)}, z^{(r)})) \right\}, \quad (\text{A15})$$

与算法 2 中使用的接受概率一致。因此, 在上述观点下, 经过修改后的 HMC 采样算法就是一种特殊的 MH 算法。此外, 由于 z' 出现在 $H(\theta', z')$ 中的二次型部分, 是否对 z' 进行反向操作并不影响 $H(\theta', z')$ 的计算结果, 并且每次采样后只需保留对 θ 的采样结果, 故在算法执行过程中并不需要事实上实施该反向操作。

A.5. 拓展——朗之万蒙特卡洛采样

由于 HMC 采样算法的每步采样过程均要执行多步 Leapfrog 迭代, 因此相比于经典的 MCMC 算法, 其单步计算代价偏高。为减小计算代价, 适应大规模问题的应用需求, 可对 HMC 采样算法进行简化。具体而言, 可在每步采样过程中仅执行一步 Leapfrog 迭代, 这样式(A8)中的第三行不再必要, 从而候选点的产生过程可简化为

$$\theta' = \theta^{(\tau)} - \frac{\varepsilon^2}{2} M^{-1} \nabla U(\theta^{(\tau)}) + \varepsilon M^{-1} z, \quad z \sim \mathcal{N}(0, M). \quad (\text{A16})$$

进一步地，令 $M = I$ ，并直接接受候选点 θ' ，则可得到如下基于迭代的采样过程：

$$\theta^{(\tau+1)} = \theta^{(\tau)} - \frac{\varepsilon^2}{2} \nabla U(\theta^{(\tau)}) + \varepsilon z, \quad z \sim \mathcal{N}(0, I). \quad (\text{A17})$$

此即为朗之万蒙特卡洛(Langevin Monte Carlo, LMC)采样。虽然相比 HMC 采样，该采样算法的计算效果较差，但其计算效率则显著提升。事实上，该采样过程本质上可以理解为以 $\nabla U(\theta)$ 为极小化目标，对 θ 进行带随机扰动的梯度下降，属于复杂度较低的一阶算法，从而可适应当代大数据大模型条件下的计算需求。正是基于这一考虑，Welling 等人[10]进一步将随机梯度的思想引入 LMC 采样，给出了 SGLD 算法，该算法在机器学习中应用广泛，目前已成为贝叶斯深度学习计算中最常用的算法之一。