计算机科学
COMPUTER SCIENCE

贝叶斯推理与并行回火研究综述

湛进, 王雪飞, 成雨蓉, 袁野

引用本文

湛进，王雪飞，成雨蓉，袁野.贝叶斯推理与并行回火研究综述[J]．计算机科学, 2023, 50(2): 89-105.

ZHAN Jin, WANG Xuefei, CHENG Yurong, YUAN Ye. Overview of Research on Bayesian Inference and Parallel Tempering [J]. Computer Science, 2023, 50(2): 89-105.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100001

贝叶斯推理与并行回火研究综述

湛 进 王雪飞 成雨蓉 袁 野
北京理工大学计算机学院 北京 100081
(yukiho1124z@gmail.com)

摘 要 贝叶斯推理是统计学中的主要问题之一,旨在根据观测数据更新概率分布模型的先验知识。对于真实情况下常遇到的无法观测或难以直接计算的后验概率,贝叶斯推理可以对其进行近似,它是一种以贝叶斯定理为基础的重要方法。在许多机器学习问题中都涉及对包含各类特征数据的真实分布进行模拟和近似的过程,如分类模型、主题建模和数据挖掘等,因此贝叶斯推理在当今机器学习领域里具有重要而独特的研究价值。随着大数据时代的开始,研究者经由实际信息采集到海量的实验数据,导致需要模拟和计算的目标分布也非常复杂,如何在复杂数据下对目标分布进行结果精确和时间高效的近似推理,成为了当今贝叶斯推理问题的重难点。针对这一复杂分布模型下的推理问题,文中对近年来解决贝叶斯推理问题的两大主要方法——变分推理和采样方法,进行系统性地介绍和综述。首先,给出变分推理的问题定义与理论知识,详细介绍以坐标上升为基础的变分推理算法,并给出这一方法的已有应用与未来展望。然后,对国内外现有的采样方法的研究成果进行综述,给出各类主要采样方法的具体算法流程,并总结和对比这些方法的特性与优缺点。最后,引入并行回火技术,对其基本理论和方法进行概述,探讨并行回火与采样方法的结合与应用,为未来贝叶斯推理问题的发展探讨了新的研究方向。

关键词:变分推理;采样算法;并行回火;近似计算
中图法分类号 TP181

Overview of Research on Bayesian Inference and Parallel Tempering

ZHAN Jin,WANG Xuefei,CHENG Yurong and YUAN Ye
School of Computer,Beijing Institute of Technology,Beijing 100081,China

Abstract Bayesian inference is one of the main problems in statistics. It aims to update the prior knowledge of the probability distribution model based on the observation data. For the posterior probability that cannot be observed or is difficult to directly calculate,which is often encountered in real situations,Bayesian inference can obtain a good approximation. It is a kind of important method based on Bayesian theorem. Many machine learning problems involve the process of simulating and approximating the target distribution of various types of feature data,such as classification models,topic modeling,and data mining. Therefore,Bayesian inference has shown important and unique research value in the field of machine learning. With the beginning of the big data era,the experimental data collected by researchers through actual information is very large,resulting in the complex distribution of targets to be simulated and calculated. How to perform accurate and time-efficient approximation inferences on target distributions under complex data has become a major and difficult point in Bayesian inference problems today. Aiming at the inference problem under this complex distribution model,this paper systematically introduces and summarizes the two main methods for solving Bayesian inference problems in recent years,which are variational inference and sampling methods. Firsly,this paper gives the problem definition and theoretical knowledge of variational inference,introduces in detail the variational inference algorithm based on coordinate ascent,and gives the existing applications and future prospects of this method. Next,it reviews the research results of existing sampling methods at home and abroad,gives the specific algorithm procedure of various main sampling methods,as well as summarizes and compares the characteristics,advantages and disadvantages of these methods. Finally,this paper introduces parallel tempering technique,outlines its basic theories and methods,discusses the combination and application of parallel tempering and sampling methods,and explores new research directions for the future development of Bayesian inference problems.

Keywords Variational inference,Sampling methods,Parallel tempering,Approximate computation

到稿日期:2022-01-04 返修日期:2022-06-23
基金项目:国家自然科学基金(61902023,U1811262,U21B2007,61932004,61572119,61622202);中央高校基本科研业务费专项资金(N181605012)
This work was supported by the National Natural Science Foundation of China(61902023,U1811262,U21B2007,61932004,61572119,61622202) and Fundamental Research Funds for the Central Universities of Ministry of Education of China(N181605012).
通信作者:成雨蓉(yrcheng@bit.edu.cn)

1 引言

对复杂分布的推理是机器学习中的一个常见问题，许多贝叶斯方法都需要这种推理。该问题实质上是对真实分布的各类特征的模拟和近似。在得到或计算出这些信息后，就可以对分布进行分析、模拟推算或预测。在信息快速发展与更新的今天，世界已经进入大数据时代和智能时代，对数据的处理量和处理准确性、实时性都有了更高的要求，从不同需求和难点衍生出的各类算法也层出不穷。在过去的十年里，对于复杂分布的各种推理方法在生物、物理、金融、环境科学以及医学等领域都得到了广泛的应用。在实际情况中，实验者往往会采集到形式十分随机的数据，它本身就会成为某种形式的多峰分布，因此，对于实际多峰分布的求解和近似是机器学习中十分重要的方向之一。常见的情况是，实验者只能采集到大量样本，而不能得到样本的标签属性或样本有部分属性缺失，有时也会面临需要对高维度特征采样的情况。在这些情况下，使用传统的机器学习相关算法直接训练变得困难，而只能选择近似实际数据的分布。如何在复杂数据下对目标分布进行更高效的计算和推理，一直是整个机器学习领域研究的重难点，而这也是自然语言处理、图像识别、数据挖掘等许多领域中基础原理的重要部分。

对多峰分布的近似的各类算法都与概率论中贝叶斯统计的思想有着十分密切的关系。贝叶斯统计的思想起源于18世纪，在漫长的发展中出现过许多计算和近似算法，但其在实际问题上的应用常常受限，这主要是实际问题中维度和样本量的问题导致的。在实际应用中常常出现的高维度求积分、海量样本计算、稀疏分布近似等情况中，贝叶斯统计算法常常难以实现或计算困难，而这些步骤却是必须完成的。为了解决或缓解此类问题，现有的研究针对不同情境下的近似推理问题设计了不同的改良算法，主要分为确定性计算算法和非确定性计算算法两类。近年来也有将二者进行合并的混合算法。一般而言，确定性算法在近似质量上会受到与目标分布适配性的制约，即选择的近似分布和算法是否合适会对结果产生影响，但相比之下计算开销较低；非确定性算法适用范围广，近似效果普遍较好，但是通常计算代价相对较高。

贝叶斯推理是从贝叶斯的角度产生统计推断的过程，统计推断旨在根据可观察到的事物来了解不可观察到的事物，即统计推断是基于一个总体或一些样本中的某些观察变量得出结论的过程，例如关于总体或样本中某些潜在变量(通常是原因)的准时估计、置信区间或区间估计等。简而言之，贝叶斯推理是一种统计或是概率范式，在这种范式中，每次记录新的观测数据时就会更新由概率分布建模的先验知识，而观测数据的不确定性则由另一个概率分布建模。支配贝叶斯范式的整个思想嵌入在贝叶斯定理中，该定理表达了更新知识(后验)、已知知识(先验)以及来自观察的知识(可能性)之间的关系。

一个经典的例子是用贝叶斯推理进行参数估计。假设一个模型中数据 x 是根据未知参数 θ 的概率分布生成的，并且有关于参数 θ 的先验知识，可以用概率分布 $p(\theta)$ 来表示。那么，当观察到数据 x 时，我们可以使用贝叶斯定理来更新关于该参数的先验知识。根据贝叶斯定理，后验分布的计算需要3个条件：先验分布、可能性和证据。前两个条件很容易理解，因为它们是假设模型的一部分，在许多情况下，先验分布和可能性是显而易见的。然而，第三个条件，即归一化因子，则需要进行如下计算：

$$p(x)=\int_{\theta}p(x|\theta)\,p(\theta)\,\mathrm{d}\theta \tag{1}$$

虽然在低维数据中，这个积分可以较容易地计算出来，但在高维数据中它会变得很难处理。因此，在大多数实际问题中对后验分布进行精确计算是不可行的，必须使用一些近似技术来获得后验分布。

确定性算法是以贝叶斯推理的思想为基础的一类算法。它们一般从贝叶斯原理入手，用不同的方法对目标的似然函数进行最大化。其思想根据是否知道原始函数的分布又分为直接推理和近似推理两种。直接推理需要知道目标分布的详细形式，在实际情况下就是需要知道样本满足的具体的分布形式，通过计算得到各个参数的更新公式，再循环执行各个参数的更新步骤得到目标模型的参数；近似推理采用某种形式更为简单的分布族并间接近似目标分布，主要采用变分的思想，其总体目标是简化原始分布的形式用做近似，这个新分布被称为变分分布，算法通过对变分分布的参数进行更新，从而得到原始分布的近似。

直接推理最为经典的就是期望最大化算法，自从1996年期望最大化算法被提出以来，它的思想就被应用于信号处理、自然语言处理和计算生物学等各种领域，当需要处理的数据存在部分缺失，即存在隐变量时，该算法有非常不错的效果。但该算法也存在难以计算复杂分布的问题，且在更新时受初始值的影响很大，常常陷入局部最优。间接推理常见的算法一般都是从最经典的变分推理算法衍生而来。变分推理算法的思想是期望最大化算法的思想的推广，其应用场景也与期望最大化算法类似，且弥补了它在很多复杂模型上计算困难的缺点。尽管传统的近似推理算法有着计算方便、精度较高的优势，但也有自己的缺点。随着迭代次数的增加，近似推理会逐渐缩小搜索范围，这可能使其在搜索时困于局部最优值。

非确定性算法也被称为采样算法，其主要目的是当所需结果是样本时，避免对复杂的实际分布的求解，直接用某种方法近似并直接得到与目标分布相似的样本。在非确定性算法中，马尔可夫链蒙特卡洛算法是一个十分著名的经典算法，它的思想是在满足细致平稳条件的前提下，通过随机和模拟步骤，不断对目标分布进行采样和近似操作，一直到整个马尔可夫链趋于平稳状态，此时平稳状态之后的样本值就是与目标分布近似的值。非确定性算法往往在目标分布难以抽样且难以计算时避开计算困难的部分，通过对平稳分布的求解得到与原分布类似的样本，一般计算量较大，但稳定性较强，只要满足使用条件，通常都能得到最终的平稳分布。非确定性算法在马尔可夫链蒙特卡洛思想的基础上还有很多扩展模型，如MH采样、Gibbs采样等；也在很多领域的应用中取得了较为理想的效果，如金融上的模型模拟、地球物理模型的近似和文本词袋模型近似等方面。

随着大数据时代的到来与发展，贝叶斯推理作为统计学中的一个重要问题，也是许多机器学习方法中经常遇到的

问题,例如可以用于分类的高斯混合模型和用于主题建模的隐狄利克雷分配模型(Latent Dirichlet Allocation,LDA)等常见的概率图模型都需要在处理与拟合数据时应用贝叶斯推理。因此在当今的机器学习领域,贝叶斯推理越来越显露出其独特的重要性和研究价值。同时,由于模型的多样化以及假设或是维度等模型设置在不同具体问题下的差异,贝叶斯推理问题也有着具有针对性的不同解决方法。如上文所述,本文将重点讨论两种可用于解决贝叶斯推理问题的主要方法,即基于近似的变分推理方法(Variational Inference,VI)和基于采样的各类采样算法。本文还将引入对采样方法进行推广得到的并行回火方法(Parallel Tempering,PT)。并行回火的思想基于马尔可夫链蒙特卡洛采样(Markov Chain Monte Carlo,MCMC),同时对单一 MCMC 采样中可能导致局部最优解问题进行了改进与优化,从而给解决贝叶斯推理问题带来了新的技术。

本文主要对贝叶斯推理问题的变分推理和采样算法这两大类方法,以及并行回火这一采样算法的扩展技术的现有研究与应用进展进行综述。本文第 2 节详细介绍了变分推理的理论知识,具体阐述了主要的变分推理算法,并且介绍了变分推理在各个领域与学科中的应用,最后分析了变分推理的相关理论以及存在的开放性问题;第 3 节详细介绍了采样方法的理论知识,简要介绍了采样方法的研究发展与应用,并且给出了常见的采样算法的流程与实现,最后讨论并总结了各类采样方法的优缺点;第 4 节中详细介绍了并行回火的理论知识,重点分析了其优势和相关参数的设置,并且讨论了并行回火与采样方法的结合与应用。全文的技术树图如图 1 所示。

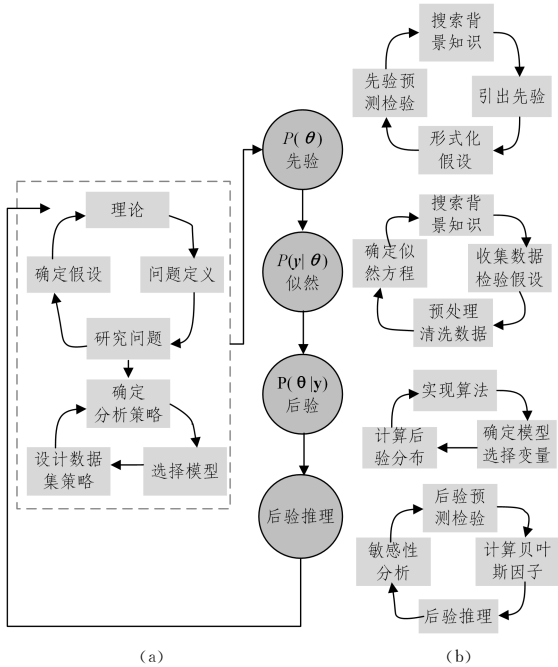

图 1　技术树

Fig. 1　Technology tree

2　变分推理研究综述

2.1　简介

变分推理(VI)是一种重要的贝叶斯近似推理方法,也是用于贝叶斯估计和机器学习领域中近似计算复杂积分的技术。

贝叶斯推理的变分技术的发展遵循了两个平行但独立的轨迹。1987 年,第一个变分程序出现[1],它结合统计力学的见解,引出了后来各种模型上的一系列变分推理过程[2-3]。同时,Hinton 等为一个类似的神经网络模型提出了一种变分算法[4],Neal 等将其与期望最大化(Expectation-Maximuzation)算法建立了重要联系[5],进一步引出了其他类型模型的各种变分推理算法[6]。

目前,关于变分推理的研究集中在以下几个方面:

(1)处理涉及大量数据的贝叶斯推理问题;

(2)使用改进的优化方法最小化 KL 散度;

(3)开发适用于广泛模型的变分推理;

(4)提高变分推理的准确性。

本节将从问题定义出发,介绍变分推理方法背后的基本思想:平均场推理和坐标上升优化。接着,将描述变分推理的模型的潜在变量和观察变量的指数族中的联合密度,其中包括现代贝叶斯统计中出现的许多棘手的模型,并揭示了变分推理和吉布斯采样器的深层关系[7]。然后,我们对该算法进行扩展以描述随机变分推理[8](Stochastic Variational Inference,SVI),该算法使用随机优化将变分推理扩展到海量数据情形。最后,本文将给出变分推理的应用与理论结果,以及下一步可以研究的问题。变分推理方法的总体技术树图如图 2 所示。

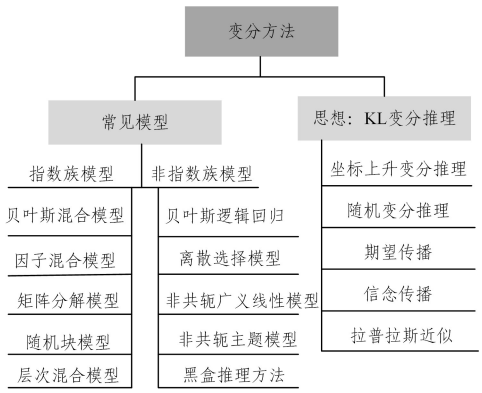

图 2　变分方法技术树

Fig. 2　Technology tree of variational inference

2.2　近似推理问题

首先定义广义的变分推理思想。对于潜变量 $\boldsymbol{z}=z_{1:m}$ 和观测值 $\boldsymbol{x}=x_{1:n}$,有:

$$p(\boldsymbol{z},\boldsymbol{x})=p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) \tag{2}$$

在贝叶斯模型中,潜在变量用于帮助调节数据的分布。贝叶斯模型从先验密度 $p(\boldsymbol{z})$ 中提取潜在变量,再通过可能性 $p(\boldsymbol{x}|\boldsymbol{z})$ 将它们与观测结果联系起来。在贝叶斯模型中的推理相当于调节数据和计算后验 $p(\boldsymbol{z}|\boldsymbol{x})$,这种计算通常需要近似推理。变分推理的目标是在给定观察变量的前提下近似潜在变量的条件密度,一般通过优化的思想来解决这个问题:在潜在变量上使用一系列密度,由自由"变分参数"参数化。优化找到这个族的成员,即参数的设置,在 KL 散度中最接近感兴趣的条件。简单来说,就是在已知目标分布数据的前提下用易于计算的的简单分布近似拟合难以计算或形式未知的真实

目标分布，如图 3 中对于一个未知的分布(灰色阴影)，可以用两个高斯分布(黑色和灰色虚线)逐渐近似。

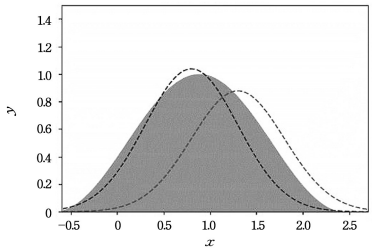

图 3 变分推理示例

Fig. 3 Example of variational inference

近似推理问题是在给定观测的条件下，计算潜在变量的条件密度，此条件密度可用于生成潜在变量的点或区间估计以及形成新数据的预测密度等。条件密度可以写为：

$$p(\boldsymbol{z}|\boldsymbol{x})=\frac{p(\boldsymbol{z},\boldsymbol{x})}{p(\boldsymbol{x})} \tag{3}$$

其中，分母包含观察的边际密度，也称为证据(Evidence)。从联合密度中边缘化潜在变量即可计算：

$$p(\boldsymbol{x})=\int p(\boldsymbol{z},\boldsymbol{x})\mathrm{d}\boldsymbol{z} \tag{4}$$

对于许多模型，这个证据积分不可计算或需要指数时间来计算。而证据正是我们从联合概率中计算的条件概率时需要的，这导致在这种模型中进行推理变得困难。

变分推理原理如下：

首先，引入一种衡量分布间距离的度量，即 KL 散度(Kullback-Leibler divergence)。

$$KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x}))=\mathbb{E}[\log q(\boldsymbol{z})]-\mathbb{E}[\log p(\boldsymbol{z}|\boldsymbol{x})] \tag{5}$$

式(5)即为对 $q(\boldsymbol{z})$ 和 $p(\boldsymbol{z}|\boldsymbol{x})$ 两个分布的距离度量。另外，由于 KL 散度通常无法计算，实际优化时一般用证据下界(Evidence Lower Bound，ELBO)来替代 KL 散度。

$$ELBO(q)=\mathbb{E}[\log p(\boldsymbol{z}|\boldsymbol{x})]-\mathbb{E}[\log q(\boldsymbol{z})] \tag{6}$$

寻找近似分布。假设有多个密度分布，其中每一个都含有潜在变量，由这一组密度相乘组成近似密度族 $\mathscr{D}$。然后找到该族某成员，使 KL 散度近似到精确后验分布。

$$q^*(\boldsymbol{z})=\arg\min_{q(\boldsymbol{z})\in\mathscr{D}} KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) \tag{7}$$

再用优化后的 $q^*(\cdot)$ 成员近似后验。这样，变分推理就将推理问题转化为了优化问题，如图 4 所示，通过逐渐迭代更新近似分布 $q(\boldsymbol{z};\boldsymbol{v})$ 中的参数 v 到 $q(\boldsymbol{z};\boldsymbol{v}^*)$，分布间的距离即 KL 值不断减小，近似分布逐渐接近真实后验 $p(\boldsymbol{z}|\boldsymbol{x})$。

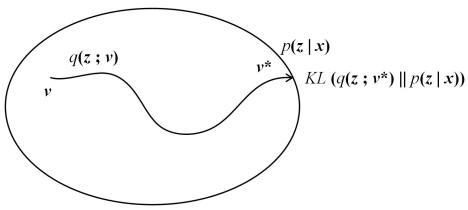

图 4 变分推理思想

Fig. 4 Basic thought of variational inference

其中，族 $\mathscr{D}$ 成员的类型和该类型分布的覆盖范围会影响优化的复杂性，在实际使用时需要结合具体情况考虑如何选择。变分推理的关键问题之一是要选择足够灵活的 $\mathscr{D}$ 以接近 $p(\boldsymbol{z}|\boldsymbol{x})$ 的密度，但优化时的复杂度也要尽量低。基于 $KL(\boldsymbol{q} \parallel \boldsymbol{p})$ 的优化，也称为 KL 变分推理(Barber，2012)，任何使用优化来近似密度的过程都可以被称为“变分推理”，包括期望传播[9]、信念传播[10]，甚至是拉普拉斯近似。

2.3 坐标上升平均场变分推理

第 2.2 节中提到了近似密度族 $\mathscr{D}$ 的概念，平均场族即是密度族中相乘的所有分布都具有同样的形式，且互相独立。利用 2.2 节中提到的证据下界 ELBO 和平均场族，可以将近似条件推理作为一个优化问题看待。本节将描述解决这一优化问题的最常用的算法之一：坐标上升变分推理(Coordinate ascent mean-field variational inference，CAVI)[11]。CAVI 迭代地优化平均场变分密度的每个因子，同时保持其他因子不变。它将证据下界 ELBO 上升到当地的最优值。CAVI 利用 ELBO 不断搜索，最终找到一个局部最优值。

CAVI 也可以被看作是一种“消息传递”算法[12]，其根据马尔可夫毯中变量的变分参数迭代更新每个随机变量的变分参数。这一观点使为大型模型设计自动化软件成为可能[9]。变分信息传递将变分推理与图形模型和概率推理的经典理论联系起来[13]，它已经被扩展到非共轭模型[14]，并通过因子图进行推广。

最后，CAVI 与近似推理的经典主力——吉布斯抽样密切相关[15]。吉布斯采样器保持了潜在变量的实现，并从每个变量的完整条件中迭代采样，本文将在第 3 节中详细介绍。CAVI 算法的伪代码如算法 1 所示。

算法 1 坐标上升变分推理(CAVI)算法

输入：模型 $p(\mathbf{x},\mathbf{z})$，数据集 $\mathbf{x}$

输出：变分密度 $q(\mathbf{z})=\prod_{j=1}^{m} q_j(z_j)$

1. 初始化因子 $q_j(z_j)$
2. while(ELBO 未收敛) do
3. for $j\in 1,\cdots,m$ do
4. 设置 $q_j(z_j)\propto\exp \mathbb{E}_{-j}[\log p(z_j|\mathbf{z}_{-j},\mathbf{x})]$
5. end
6. 计算 $\mathrm{ELBO}(q)=\mathbb{E}[\log p(\mathbf{z},\mathbf{x})]-\mathbb{E}[\log q(\mathbf{z})]$
7. end

2.4 变分推理与指数族

2.3 节描述了平均场变分推理和由之推导出的一种用于优化 ELBO 的坐标上升算法 CAVI，本节将介绍变分推理与指数族的关系。指数族模型包括许多广泛使用的模型，如指数族的贝叶斯混合模型、因子混合模型、矩阵分解模型、某些层次回归模型(如线性回归、概率回归)、网络的随机块模型、层次混合模型等。

2.4.1 条件共轭模型和贝叶斯模型

指数族模型的一个重要特例是具有局部变量和全局变量的条件共轭模型。像这样的模型在贝叶斯统计和统计机器学习中经常出现，其中全局变量是“参数”，局部变量是每个数据点的潜在变量。

首先介绍条件耦合模型。设 $\boldsymbol{\beta}$ 是一个全局潜在变量的向量，它可以潜在地控制任意数据。设 $\boldsymbol{z}$ 是局部潜在变量的向量，其第 i 个分量只支配第 i 个数据的上下文。联合密度为：

$$p(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{x})=p(\boldsymbol{\beta})\prod_{i=1}^{n} p(z_i,x_i|\boldsymbol{\beta}) \tag{8}$$

全局变量是混合成分，第 i 个局部变量是数据点 x_i 的聚类分配。为了确保每个完整的条件都在指数族中，首先假设每个 (x_i,z_i) 对的基于 $\boldsymbol{\beta}$ 的联合密度有一个指数族形式，其

中，$t(\cdot,\cdot)$ 为充分统计量。

$$p(z_i,x_i|\boldsymbol{\beta})=h(z_i,x_i)\exp\boldsymbol{\beta}^{\mathrm{T}}t(z_i,x_i)-a(\boldsymbol{\beta}) \tag{9}$$

接下来，将全局变量上的先验设为相应的共轭先验：

$$p(\boldsymbol{\beta})=h(\boldsymbol{\beta})\exp\alpha^{\mathrm{T}}[\boldsymbol{\beta},-a(\boldsymbol{\beta})]-a(\alpha) \tag{10}$$

用一个列向量以及足够的统计数据，将全局变量及其对数归一化器连接到局部变量的密度中。再利用共轭先验，使全局变量的完全条件在同一个族中。它的自然参数是：

$$\hat{\alpha}=[\alpha_1+\sum_{i=1}^{n}t(z_i,x_i),\alpha_2+n]^{\mathrm{T}} \tag{11}$$

再推导局部变量z_i的完整条件。给定$\boldsymbol{\beta}$和x_i，局部变量z_i有条件地独立于其他局部变量z_{-i}和其他数据x_{-i}。这来自于式(8)中的联合密度的形式。因此有：

$$p(z_i|x_i,\boldsymbol{\beta},\boldsymbol{z}_{-i},\boldsymbol{x}_{-i})=p(z_i|x_i,\boldsymbol{\beta}) \tag{12}$$

若这个密度是在一个指数族中，则可以进一步得到式(13)，这是式(9)中的局部似然项 $p(z_i,x_i|\boldsymbol{\beta})$的一个性质：

$$p(z_i|x_i,\boldsymbol{\beta})=h(z_i)\exp\eta(\boldsymbol{\beta},x_i)^{\mathrm{T}}z_i-a(\eta(\boldsymbol{\beta},x_i)) \tag{13}$$

接着，我们推导条件共轭模型中的变分推理，这是一般模型中的 CAVI。设$\boldsymbol{\beta}$上的变分后验近似为$q(\boldsymbol{\beta}|\lambda)$，它是一个与之前相同的指数族密度，其中$\lambda$为全局变分参数。类似地，将每个局部变量$z_i$上的与局部完全条件相同的指数族密度$q(z_i,\phi_i)$作为变分后验，由一个局部变分参数$\phi_i$控制。CAVI 即是在更新每个局部变分参数和更新全局变分参数之间进行迭代。其中，局部变分更新式为：

$$\varphi_i=\mathbb{E}_{\lambda}[\eta(\boldsymbol{\beta},x_i)] \tag{14}$$

式(14)是取式(12)中完全条件的自然参数的期望。相同地，利用式(11)中对自然参数的期望，全局变分更新式为：

$$\lambda=[\alpha_1+\sum_{i=1}^{n}\mathbb{E}_{\varphi_i}[t(z_i,x_i)],\alpha_2+n]^{\mathrm{T}} \tag{15}$$

CAVI 通过迭代每个本地参数的局部更新和全局参数的全局更新来优化 ELBO。为了评估收敛性，我们可以在每次迭代时计算 ELBO，直到得到一个不依赖于变分参数的常数，这是 ELBO 应用于式(8)中的联合概率密度和相应的平均场变分密度，并省略了不依赖于变分参数的项得到的结果，它的一种典型应用就是基于高斯混合模型的 CAVI。

$$ELBO=(\alpha_1+\sum_{i=1}^{n}\mathbb{E}_{\varphi_i}[t(z_i,x_i)])^{\mathrm{T}}\mathbb{E}_{\lambda}[\boldsymbol{\beta}]-(\alpha_2+n)\mathbb{E}_{\lambda}[a(\boldsymbol{\beta})]-\mathbb{E}[\log q(\boldsymbol{\beta},\boldsymbol{z})]$$

其中，

$$\mathbb{E}[\log q(\boldsymbol{\beta},\boldsymbol{z})]=\lambda^{\mathrm{T}}\mathbb{E}_{\lambda}[t(\boldsymbol{\beta})]-a(\lambda)+\sum_{i=1}^{n}\varphi_i^{\mathrm{T}}\mathbb{E}_{\varphi_i}[z_i]-a(\varphi_i) \tag{16}$$

2.4.2　随机变分推理

概率模型的现代应用通常需要分析大量的数据。然而，大多数后验推理算法并不容易进行缩放。CAVI 也不例外，该算法的坐标上升结构需要在每次迭代中迭代整个数据集。随着数据集的增长，每次迭代的计算成本都变得更加昂贵，继而出现了坐标上升的另一种选择：基于梯度的优化。它通过计算并在每次迭代中遵循其梯度来优化 ELBO。这一观点是使用随机变分推理(SVI)来扩大变分推理的关键[8]，也是一种结合了自然梯度[16]和随机优化[17]的方法。

SVI 着重于优化条件共轭模型的全局变分参数λ，计算的流程很简单。该算法保持了对全局变分参数的当前估计。算法过程如下：

(1)从完整数据集中对一个数据点进行子采样；

(2)使用当前的全局参数来计算子采样数据点的最佳局部参数；

(3)以适当的方式调整当前的全局参数。

SVI 的详细流程如算法 2 所示，它是一个优化 ELBO 的有效算法。

算法 2　条件共轭模型中的随机变分推理(SVI)

输入：模型 $p(\mathbf{x},\mathbf{z})$，数据集 $\mathbf{x}$，步长序列ϵ_t

输出：全局变分密度$q_\lambda(\boldsymbol{\beta})$

1. 初始化因子λ_0
2. while(True) do
3.　　从均匀分布中选择数据点 t，即令 t～Uni(1,…,n)
4.　　最优化局部变分参数$\varphi_t^*=\mathbb{E}_\lambda[\eta(\boldsymbol{\beta},x_t)]$
5.　在x_t的 n 次循环中不断计算坐标更新$\hat{\lambda}=\alpha+n\,\mathbb{E}_{\varphi_t^*}[f(z_t,x_t)]$
6.　　更新全局变分参数$\lambda_t=(1-\epsilon_t)\lambda_t+\epsilon_t\hat{\lambda}_t$
7. end

除了 CAVI 所需要的东西以外，SVI 并不需要新的推导。CAVI 的任何实现都可以立即扩展到一个随机算法，SVI 的主要优势是可以更高效地在较大数据集上进行计算。

2.5　非条件共轭的模型

前面描述的都是条件概率属于指数族的模型，但许多模型并没有这个特性，如贝叶斯逻辑回归：

$$\begin{aligned}&\beta_k\sim\mathcal{N}(0,1)\\&y_i|x_i,\beta\sim Bern(\sigma(\beta^{\mathrm{T}}x_i))\end{aligned} \tag{17}$$

其中，$\sigma(\cdot)$ 为逻辑函数。因为系数的后验密度不属于指数族，所以不能应用上面讨论的变分推理方法。具体来说，就是不能计算式(6)的 ELBO 第一项中的期望或 CAVI 中的坐标更新。

探索这种模型的变分方法一直是一项非常具有挑战性的研究工作。早在 1997 年，就有第一个针对逻辑回归的变分界被提出[18]。后来，Blei 等在 2007 年将他们的想法改编到非共轭主题模型[19]。在其他工作中，Braun 等推导出了离散选择模型的变分推理算法[20]，该算法也在条件共轭模型类之外。他们还开发了一种增量方法来近似难以计算的期望。Wand 等使用辅助变量方法、求积和混合近似来处理指数族之外的各种似然项[21]。

最近，研究人员推广了非共轭推理，寻找可以用于许多模型的配方。Wang 等为此采用了拉普拉斯近似和 delta 方法[22]，改进了非共轭广义线性模型和主题模型的推理；Bugbee 等也使用了这种方法进行半参数回归[23]。Knowles 等和 Jaakkola 等在消息传递算法中推广了早期算法的边界[24-25]，Wand 进一步简化和扩展了他们的方法[26]。Tan 等将这些消息传递方法应用于广义线性混合模型，并将它们与 SVI 结合起来[27]。Rohde 等统一了许多算法的发展，并为它们的数值实现提供了实际的见解[28]。

研究者们对利用蒙特卡罗(MC)估计梯度的困难变分目标也进行了一系列的研究。其想法是将 ELBO 的梯度作为期望，计算它的 MC 估计，然后使用重复 MC 梯度的随机优化，这种思想在多篇文献中都出现过[29-32]。最新的方法避免了任何特定于模型的推导，它被称为“黑盒”推理方法[33-36]。也有团队利用这些想法实现自动 VI 技术，该技术适用于

概率编程系统 Stan 中编写的任何模型，这是迈向无推导且易于使用的 VI 算法的重要一步。

2.6 应用

近年来，许多领域的研究人员已经使用变分推理来解决真实的问题。

在计算生物学中，概率模型为分析遗传数据提供了重要的构建模块。VI 已被用于全基因组关联研究[37]、调控网络分析[38]、基序检测[39]、系统发育隐马尔可夫模型[40]、群体遗传学[41]和基因表达分析[42]等方面。

在计算机视觉和机器人技术中，视觉研究人员经常分析较大的高维图像数据集。变分推理一直是该领域很重要的一种方法，它已经在推理非线性图像流形[43]和在视频中寻找图像层[44]、建立视频概率模型[45]、图像去噪[46]、图像识别追踪[47,48]、机器人位置定位和映射[49-50]，以及贝叶斯非参数图像分割[51]等领域得到应用。

同样地，计算神经科学也需要分析非常大的高维数据集，如高频时间序列数据或高分辨率功能磁成像数据。变分推理在神经科学中有很多应用，特别是对于自回归过程[52-54]。变分推理在神经科学中的其他应用包括多个主题的层次模型[55]、空间模型[56-59]、脑机接口[60]和因子模型[61-62]等。

在自然语言处理和语音识别领域，变分推理已被用于如语法分析[63]、语法归纳[64-66]、流文本模型[67]、主题建模[68]以及隐藏马尔可夫模型和词部标记[69]等问题。在语音识别中，变分推理已被用来拟合复杂耦合的隐藏马尔可夫模型[70]和开关动态系统[71]。

除此之外，变分推理还有许多其他的应用，包括市场营销[72]、最佳控制和强化学习[73-74]、统计网络分析[75-76]、天体物理学[77]，以及社会科学[78-79]等。

2.7 总结与展望

本节描述了变分推理(VI)——一种用优化来进行概率计算的方法。其目的是近似于给定观察变量 $\boldsymbol{x}$、求 $p(\boldsymbol{z}|\boldsymbol{x})$的潜在变量 $\boldsymbol{z}$ 的条件密度。其想法是假设一个密度族 Q，然后找到在 KL 散度中最接近目标分布的成员$Q^*(\cdot)$。然后，本节描述了使用平均场族使得变分推理可以适用于坐标上升优化(CAVI)，迭代地优化每个目标变量。此外，本节进一步讨论了指数族和条件共轭的特殊情况，并描述了随机变分推理[8](SVI)，它将平均场变分推理在大量数据下的计算变得可行。

变分推理领域还有很多值得钻研的问题。变分推理专注于优化 $KL(q(\boldsymbol{z}) \| p(\boldsymbol{z}|\boldsymbol{x}))$，并将其作为变分目标函数。其中，开发变分推理方法来优化其他度量就是一个很有前景的研究方向，如寻找比 ELBO 更接近真实值的下界[80-81]，或在保持优化能力的同时寻找更好的近似值。同时，探索如何权衡计算精度和速度，将变分推理和马尔可夫链蒙特卡洛算法良好地结合来进行近似推理也是未来研究的一个重要领域。此外，由于变分推理的统计特性还没有得到很好的理解，因此也可以研究变分推理与精确后验分布在统计学上的具体联系。

3 采样方法研究综述

3.1 简介

在统计学中有两种主要的采样类型。第一种是调查抽样，指从一组或总体中抽取样本；第二种是从概率分布中采样，在这一类采样问题中通常有一个概率密度函数或质量函数。本节将重点阐述从概率分布中采样的方法，主要解释了从累积分布函数中采样、蒙特卡洛近似、简单蒙特卡洛方法和马尔可夫链蒙特卡洛(MCMC)方法等采样算法。对于迭代独立的简单蒙特卡洛方法，本节具体概述了重要性采样和拒绝采样这两种方法。对于马尔可夫链蒙特卡洛方法，本节详细介绍了 Metropolis 算法、Metropolis-Hastings 算法、Gibbs 采样算法和分片采样算法等。然后解释了简单蒙特卡洛方法和更高效的蒙特卡洛方法的随机游走行为，包括哈密顿蒙特卡洛(Hamiltonian Monte Carlo)、阿德勒过松弛(Adler's overrelaxation)和有序过松弛等。最后，将上述采样方法进行对比，讨论并总结了不同方法的特性以及优缺点。

采样是统计学中的一项基础课题，这一概念主要分为两个不同的类别。第一类采样指从总体或是集合中选择实例的调查抽样，关于抽样调查的研究已经有许多文献和书籍资料，如文献[82-91]。调查抽样作为一个传统的统计学研究领域，在未来也有许多可能的发展领域[92]，特别是在分布式网络和图研究中[93-94]。调查抽样中被广泛使用的方法有简单随机抽样(SRS)[95]、自举法[96]、分层抽样、整群抽样[97]、多阶段抽样 、网络抽样[98]和滚雪球抽样[99]等。

另一类采样指从概率分布中采集样本。对于从概率分布中进行采样的问题而言，如果分布是简单分布，或者可以得到累积分布函数(CDF)，那么就可以轻松地从分布中采样。但是，如果分布很复杂就无法简单直接地从中采样。在实际应用中，数据分布通常十分复杂从而导致采样困难，例如数据分布可以是多种分布的混合[100]。现有研究表明可以通过从其他简单的样本分布中采样来近似复杂分布中的样本。使用这种采样近似思想的抽样方法被称为蒙特卡洛方法[11,101-104]。蒙特卡洛近似[102]可用于估计数据函数对数据分布的期望或是概率。蒙特卡洛方法可以分为两大类，即简单的蒙特卡洛方法和马尔可夫链蒙特卡洛(MCMC)[105]。需要注意的是蒙特卡洛方法是迭代的，在简单的蒙特卡洛方法中，每次迭代都独立于先前的迭代，并且每次的采样过程也是独立执行。简单的蒙特卡洛方法包括重要性采样[106]和拒绝采样(也被称为接受-拒绝采样)[107]。与上述简单的蒙特卡洛方法不同，在马尔可夫链蒙特卡洛方法中[108]，每次迭代都依赖于前一次迭代，因为这一方法具有马尔可夫特性[109]。马尔可夫链蒙特卡洛方法包括 Metropolis 算法[110]、Metropolis-Hastings 算法[111]、Gibbs 采样[112]和切片采样[113,114]等。由于算法中的随机游走行为，Metropolis 算法通常运行较慢，效率偏低[115]。为了提高采样方法的效率，使其能够更快地对大数据范围进行搜索，有效的改进方法有哈密顿蒙特卡洛方法[116]、阿德勒过松弛[117]和有序过松弛[118]等。蒙特卡洛方法最初是在计算物理学中发展起来的[119]，因此，这一方法在物理学中有着广泛的应用[120]。除此以外，蒙特卡洛方法在许多其他领域也有应用，如金融[121]和强化学习[122-124]等。本节简要介绍了调查抽样方法，详细介绍了基于分布的蒙特卡洛采样方法，重点阐述了不同蒙特卡洛方法的设计与实现。最后讨论并总结

了现有的采样方法的特性以及优缺点。采样方法的总体技术树图如图 5 所示。

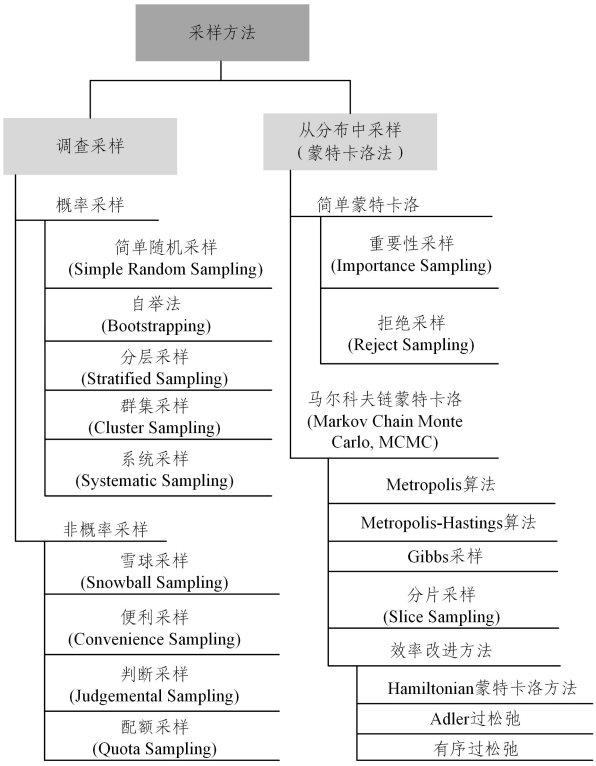

图 5　采样方法技术树图

Fig. 5　Technology tree of sampling methods

3.2　蒙特卡洛近似

假设我们正在考虑一些 d 维数据 $x\in Rd$。令 $f(x)$为数据的概率密度函数(Probability Density Function,PDF)。考虑 $h(x)$是数据 x 上的函数。根据定义,函数 $h(x)$对分布 $f(x)$的期望和函数 $h(x)$属于集合 A 的概率为:

$$E(h(x))=\int h(x)f(x)\mathrm{d}x \tag{18}$$

$$P(h(z)\in A)=\int_{h(x)\in A}f(x)\mathrm{d}x \tag{19}$$

蒙特卡洛近似的定义:使用分布 $f(x)$中大小为 n 的样本(即$x_1,\cdots,x_n\sim f(x)$),我们可以将上述两个式子近似得到方程:

$$E(h(x))\approx\frac{1}{n}\sum_{i=1}^{n}h(x_i) \tag{20}$$

$$P(h(z)\in A)\approx\frac{1}{n}\sum_{i=1}^{n}I(h(x_i)\in A) \tag{21}$$

其中,$I(\cdot)$表示条件指示函数,当条件满足时函数值为 1,当条件不满足时函数值为 0。

正如上述定义所述,蒙特卡洛近似旨在从分布中生成许多样本,以便通过这些样本的平均值来近似期望。显然,n 越大,近似效果越好。

一个经典的蒙特卡洛近似的示例就是使用 MC 方法来近似 π 的值[102]。如图 6 所示,考虑一个长度为 1 的正方形。1/4 圆存在于半径为 1 的正方形内。如果从正方形内部均匀地生成许多样本,可以看到落在 1/4 圆(圆点)内的样本占整个样本(圆点和正方形)的比例大约为 $\pi/4$,为预期的结果。生成的样本越多,这个比例就越接近 $\pi/4$。

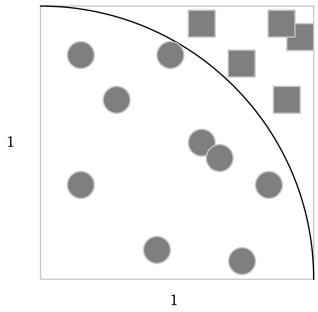

图 6　蒙特卡洛示例

Fig. 6　Example of Monte Carlo method

3.3　简单的蒙特卡洛方法

蒙特卡洛方法是从分布中生成样本的迭代方法,简单的蒙特卡洛方法会独立地从概率分布中抽取样本,因为每一个步骤或是每一次迭代都不依赖于前一次迭代。因此,在简单的蒙特卡洛方法中,迭代的过程是在数据或分布空间中独立执行的[105]。这一类简单的蒙特卡洛方法中的重要方法有重要性采样和拒绝采样(也称为接受-拒绝采样),接下来本节将详细阐述这两种方法。

3.3.1　重要性采样

首先我们考虑一个可能十分复杂的分布,其概率密度函数或是概率质量函数可以表示为:

$$f(X)=\frac{P(X)}{Z} \tag{22}$$

其中,Z 是边缘分布,也被称为归一化因子。因为难以在全数据域上进行积分,所以我们无法直接计算 Z 的值。

在重要性采样中,我们需要计算目标分布 $h(X)$在数据分布为 $f(X)$或是 $P(X)$上的期望值。由于分布过于复杂,无法直接计算,因此我们考虑使用另一个简单分布 $Q(X)$作为辅助来估计需要求得的期望值。这一引入的简单分布可以是能够轻松地对其进行采样的任意分布,如均匀分布和高斯分布。重要性采样的核心思想就是从简单分布 $Q(X)$中采集样本,而不是从复杂分布 $P(X)$中。首先算法从简单分布 $Q(X)$中采样得到 n 个样本$\{x_i\}_{i-1}^{n}$,接下来我们考虑目标函数 $j(x)$在这 n 个样本上的平均值$\frac{1}{n}\sum_{i=1}^{n}h(x_i)$。由于所采集的样本并不是从数据分布 $P(X)$中得到的,因此这一平均值也并不是我们所需要的目标期望值。为了使用这一平均值来估计目标期望值,需要加入权重因子并求和计算式(23)(由该式可以看出,我们所采集的样本数量 n 越大,最终的估计期望值也会越准确):

$$\mathbb{E}_{x\sim f(X)(h(x))}\approx\frac{1}{\sum_{j=1}^{n}\frac{P(x_j)}{Q(x_j)}}\sum_{i=1}^{n}\frac{P(x_i)}{Q(x_i)}h(x_i) \tag{23}$$

值得一提的是,Neal 于 2001 年针对重要性采样提出了一个改良的算法[125],他将退火的思想与重要性采样相结合,从而进行了良好的优化。

3.3.2　拒绝采样

回到求解目标分布 $h(X)$在复杂且难以直接计算的数据分布为 $f(X)$或是 $P(X)$上的期望值这一问题,与 3.3.1 节介绍的重要性采样类似,拒绝采样(或者称之为接受-拒绝采样)

也将使用一个简单分布 $Q(X)$作为建议分布来采集样本，但不同的是，拒绝采样的核心思想为使用这些从简单分布中得到的样本来生成复杂分布 $P(X)$上的样本。

在拒绝采样的过程中，我们考虑一个易于采样的简单分布 $Q(X)$，对于正整数常数值 c，有：

$$c * Q(x) \geqslant P(x), \forall x \in dom(X) \tag{24}$$

其中，$dom(X)$表示分布的定义域或数据 X 的取值范围。为了从复杂分布 $P(X)$中采样，我们首先从简单分布 $Q(X)$采集样本，即 $x_i \sim Q(X)$。接下来，我们从均匀分布 $U(0, c * Q(x_i))$采样得到数值u_i，如果u_i小于 $Q(x_i)$，那么就接受这一次采样结果，即将本次采集的样本值x_i作为复杂分布 $P(x)$的样本；否则，拒绝这次采样结果，并且重复执行上述过程。拒绝采样的详细算法流程和示意图分别如算法 3 和图 7 所示。

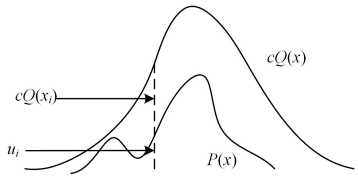

图 7 拒绝采样

Fig. 7 Rejection sampling

算法 3 拒绝采样(Rejection Sampling)

输入：复杂分布 P(X)，建议分布 Q(X)，常数 c

输出：样本集合 S={x_i}～P(X)

1. 初始化常数 c
2. while index i from 1 to n do
3. 从建议分布 Q(X)中抽样得到样本值x_i，即令x_i～Q(X)
4. 从均匀分布 U(0，c * Q(x_i))中抽样得到数值u_i，即令u_i～U(0，c * Q(x_i))
5. if u_i<P(x_i) then
6. Accept x_i：S←S∪{x_i}
7. else
8. Reject x_i：i←i−1
9. end

现今的研究中已有许多拒绝采样方法的衍生与改进算法，如自适应拒绝采样[126-128]、集合拒绝采样[129]、鉴别器拒绝采样[130]、变分拒绝采样[131]等。

3.4 马尔可夫链蒙特卡洛方法

除了 3.3 节中介绍的简单蒙特卡洛方法以外，另一类蒙特卡洛方法被称为马尔可夫链蒙特卡洛(MCMC)方法[105,132]。与简单的蒙特卡洛方法相比，马尔可夫链蒙特卡洛方法的特点在于迭代和采样不是独立的，每次蒙特卡洛迭代步骤都依赖于之前的迭代步骤，这个特点被称为马尔可夫特性。本节首先简要介绍马尔可夫链的理论与相关性质，再详细阐述基于马尔可夫链的几种 MCMC 采样方法。

3.4.1 马尔可夫链与马尔可夫特性

马尔可夫链蒙特卡洛(Markov Chain Monte-Carlo，MCMC)采样是一种在概率空间内通过随机采样来估计参数的后验分布的技术，这一方法在机器学习、深度学习以及自然语言处理等方面都有着广泛的应用。对于复杂贝叶斯模型中的后验分布，用马尔可夫链蒙特卡洛算法做近似求解通常是有效的，特别是当概率密度函数难以积分，或是概率密度函数的积分没有反函数导致难以对根据所需概率值抽样时，或者当维数过高造成“维数灾难”时，都可以考虑使用 MCMC 来进行对复杂分布的近似和抽样。

首先，本文引入马尔可夫链的定义与细致平稳条件。马尔可夫链假设某一时刻状态转移的概率只依赖于前一个状态，而与其他任何状态无关，例如，对于状态序列 $X_1, X_2, X_3, \cdots, X_{t-1}, X_t, X_{t+1}, \cdots$来说，其在$X_{t+1}$时刻的状态只与$X_t$时刻的状态有关，即 $P(X_{t+1} | X_1, X_2, \cdots, X_{t-1}, X_t) = P(X_{t+1} | X_t)$。

这一模型忽略了多个时刻之间的互相影响，在设计上大大简化了模型的复杂度。在实际使用中，许多时间序列模型如循环神经网络、隐马尔可夫模型和马尔可夫链蒙特卡洛采样中，这种“最近的状态影响最大”的思想被大量的研究与应用证明确实有效。由于某一时刻状态转移只依赖于前一个状态，那么只要求出每两个状态间的转移概率，就可以确定一个马尔可夫链模型。

其具体例子如图 8 所示。这个马尔可夫链模型就可以表示心情的转移，图中圆圈代表 3 个状态：开心、平静和沮丧。箭头代表从一个状态转移到另一个状态，箭头上的数字代表转移概率。图中马尔可夫链模型的状态转移矩阵为：$\boldsymbol{P}=\begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{pmatrix}$。

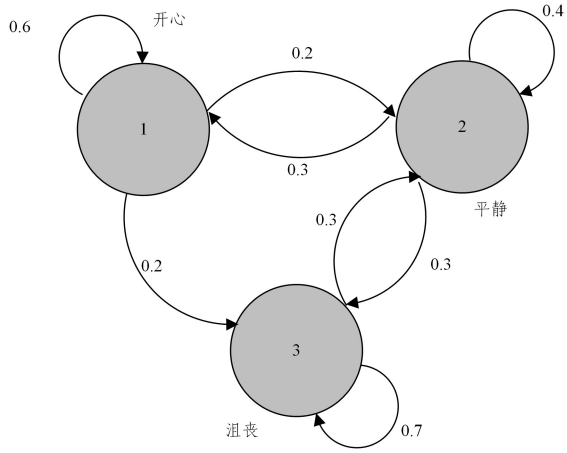

图 8 马尔可夫链

Fig. 8 Markov chain

马尔可夫链蒙特卡洛算法通过一系列连续的随机选择形成一个马尔可夫链，该马尔可夫链的平稳分布是目标分布。在马尔可夫链中，需要从平稳分布 π 中找到马尔可夫链的状态转移矩阵 $\boldsymbol{P}$。同时，由于平稳分布的定义，算法的实施要求概率分布 $\pi(x)$ 和非周期马尔可夫链状态转移矩阵 $\boldsymbol{P}$ 能在迭代的最后满足细致平稳条件，即满足 $\pi\boldsymbol{P}=\pi$。以二维的变量举例，马尔可夫链的细致平稳条件为对所有的 i 和 j，满足：

$$\pi(i)\boldsymbol{P}(i,j)=\pi(j)\boldsymbol{P}(j,i) \tag{25}$$

则称概率分布 $\pi(x)$是状态转移矩阵 $\boldsymbol{P}$ 的平稳分布。同时，如果此时这个状态转移矩阵还满足不可约、非周期和正常返，则这个状态分布 π 就是马尔可夫链唯一的平稳分布。

由于一般情况下,目标平稳分布和任意一个马尔可夫链状态转移矩阵 $\boldsymbol{Q}$ 不满足细致平稳条件,因此为了求得目标平稳分布,MCMC 算法引入了一个概率值 $\alpha(i,j)$,用以满足细致平稳条件,即:

$$\pi(i)\boldsymbol{Q}(i,j)\alpha(i,j)=\pi(j)\boldsymbol{Q}(j,i)\alpha(j,i) \tag{26}$$

那么按照对称性,有:

$$\alpha(i,j)=\pi(j)\boldsymbol{Q}(j,i),\alpha(j,i)=\pi(i)\boldsymbol{Q}(i,j) \tag{27}$$

由此就得到了目标分布 $\pi(x)$ 满足细致平稳条件的马尔可夫链状态转移矩阵 $\boldsymbol{P}$:

$$\boldsymbol{P}(i,j)=\boldsymbol{Q}(i,j)\alpha(i,j) \tag{28}$$

其中,$\alpha(i,j)$一般被称为接受率,取值在(0,1)之间,可以理解为一个概率值,这很像拒绝接受采样。拒绝接受采样是以一个常用分布通过一定的拒绝或接受概率得到一个非常见分布,马尔可夫链蒙特卡洛采样是以一个常见的马尔可夫链状态转移矩阵 $\boldsymbol{Q}$ 通过引入一定的接受率得到目标转移矩阵 $\boldsymbol{P}$,两者解决问题的思路是类似的。该算法的目标是求平稳分布后得到的近似原分布的采样。

3.4.2　Metropolis 采样

使用 Metropolis 等[110]提出的 Metropolis 算法可以从复杂的分布中采样,与上文类似,用 $f(\boldsymbol{X})$ 或 $P(\boldsymbol{X})$ 表示复杂分布,有:

$$f(\boldsymbol{X})=\frac{P(\boldsymbol{X})}{Z} \tag{29}$$

使用一个简单分布 Q 作为建议分布,建议分布通常会使用高斯分布,因为高斯分布易于采样和计算。接下来,从数据范围内的一个随机数或是随机向量开始整个算法过程。然后,根据当前样本的位置来采集下一个样本,使用简单的条件分布 $Q(x_{i+1};x_i)$作为建议分布。这个建议分布是对称的,即:

$$Q(x_{i+1};x_i)=Q(x_i;x_{i+1}) \tag{30}$$

接受率的计算式为:

$$P_{\text{accept}}=\min\left(\frac{P(x_i)}{P(x_{i-1})},1\right) \tag{31}$$

如果计算得到的接受率符合条件,那么我们就接受本次从建议分布中采样得到的样本 x_i;否则拒绝。重复上述采样过程,直到得到我们所需要的 n 个样本为止。具体的 Metropolis 算法流程和示意图如算法 4 和图 9 所示。

算法 4　Metropolis 采样(Metropolis Sampling)

输入:复杂分布 P(**X**),建议分布 Q(x_{i+1};x_i),常数 c

输出:样本集合 S=$\{x_i\}\sim$P(**X**)

1. 初始化常数 c
2. while sample index i from 1 to n do
3. 　从建议分布 Q(**X**;x_{i-1})中抽样得到样本值 x_i,即令 $x_i\sim$Q(**X**;x_{i-1})
4. 　计算接受率 $P_{\text{accept}}=\min\left(\frac{P(x_i)}{P(x_{i-1})},1\right)$
5. 　从均匀分布 U(0,1)中抽样得到数值 u_i,即令 $u_i\sim$U(0,1)
6. 　if $u_i<P_{\text{accept}}$ then
7. 　　Accept x_i:S←S∪$\{x_i\}$
8. 　else
9. 　　Reject x_i:i←i−1
10. 　end

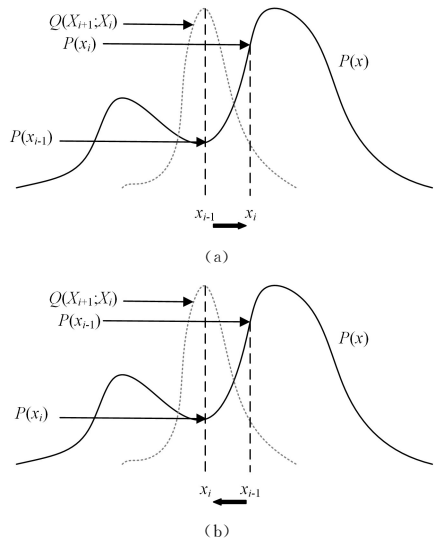

图 9　Metropolis 采样

Fig. 9　Metropolis sampling

3.4.3　Metropolis-Hastings 采样

Hastings 将 Metropolis 算法推广到不一定对称的建议分布上,我们称之为 Metropolis-Hastings 算法[111]。Metropolis-Hastings 算法与 Metropolis 算法的区别在于从建议分布中采样得到的样本的接受率的计算上,Metropolis-Hastings 算法的接受率计算式为:

$$P_{\text{accept}}=\min\left(\frac{P(x_i)Q(x_{i-1};x_i)}{P(x_{i-1})Q(x_i;x_{i-1})},1\right) \tag{32}$$

这个计算式可以看作是两项的乘法:

$$P_{\text{accept}}=\min\left(\frac{P(x_i)}{Q(x_i;x_{i-1})}\times\frac{Q(x_{i-1};x_i)}{P(x_{i-1})},1\right) \tag{33}$$

观察上述接受率计算公式可以看出,Metropolis-Hastings 算法的优化与前文中介绍的重要性采样中权重参数的引入有异曲同工之妙。

3.4.4　Gibbs 采样

Gibbs 采样方法最初由 Geman 等[15]提出,这一方法使用 d 维条件分布来从 d 维多元复杂分布 $P(\boldsymbol{X})$ 中抽取样本[11,105,133]。Gibbs 采样方法是以物理学家 Josiah Willard Gibbs 的名字命名的。这种方法最直观的思想类似于优化方法中的坐标下降技术[134-135],它假设以其余坐标为条件的数据的每个坐标特征维度的条件分布是一个容易采样的分布。

在 Gibbs 采样的问题中,我们想要从一个多变量分布 $P(\boldsymbol{X})$ 中采集样本,其中有 $\boldsymbol{X}\in\mathbb{R}^d$,即:

$$\mathbb{R}^d\ni x_i:=[x_i^{(1)},x_i^{(2)},\cdots,x_i^{(d)}]^{\mathrm{T}} \tag{34}$$

首先,Gibbs 采样算法从数据范围内的任意一个 d 维随机向量开始,然后,我们从第一维度相对于其他维度的条件分布中采集第一个样本的第一个维度数值。对于第一个样本的其他所有的维度都重复上述的采样过程,其中第 j 维的采样如式(35)所示:

$$x_i^{(j)}\sim P(\boldsymbol{X}^{(j)}\mid\boldsymbol{X}^{(1)},\cdots,\boldsymbol{X}^{(j-1)},\boldsymbol{X}^{(j+1)},\cdots,\boldsymbol{X}^{(d)}) \tag{35}$$

我们对所有的维度都执行这一采样过程,直到采集到第一个样本的每个维度数值。接下来,由第一个样本作为基础开始,对第二个样本的每个维度也重复上述采样过程,即迭代地运行这一采样过程直到生成所有的样本。然而我们注意

到，由于算法是从一个随机向量开始生成样本的，这个随机向量并不一定是符合条件的有效向量，因此算法所计算出的最初的若干样本也并不一定是合法的。我们将这样一段初始的迭代过程称为 burn-in 过程，以t_{burnIn}来表示，在经过了一定的 burn-in 过程之后，我们就能够接受算法所得到的所有样本。可以看出，Gibbs 采样其实是 Metropolis-Hastings 采样在建议样本的接受率$P_{\text{accept}}=1$时的一种特殊情况。Gibbs 采样的详细流程如算法 5 所示。

算法 5 Gibbs 采样(Gibbs Sampling)

输入：复杂分布 $P(\mathbf{X})$，建议分布 $Q(x_{i+1}|x_i)$，常数 c

输出：样本集合 $S=\{x_i\}\sim P(\mathbf{X})$

1. 初始化$\mathbf{x}_0$为 dom($\mathbf{X}$)内的一个随机 d 维向量
2. while sample index i from 1 to $n+t_{\text{burnIn}}$ do
3. $\quad x_i \leftarrow x_{i-1}$
4. $\quad$ while dimension j from 1 to d do
5. $\quad x_i^{(j)} \sim P(\mathbf{X}^{(j)} \mid \mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(j-1)}, \mathbf{X}^{(j+1)}, \cdots, \mathbf{X}^{(d)})$
6. $\quad\quad x_i \leftarrow [x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(d)}]^T$
7. $\quad\quad$ if $i \geqslant t_{\text{burnIn}}$ then
8. $\quad\quad\quad$ Accept x_i: $S \leftarrow S \cup \{x_i\}$
9. end

根据前文的具体描述可知，Gibbs 采样的挑战之一就是我们无法确切地知道最合适的进行 burn-in 迭代过程的次数。较多次的 burn-in 迭代会导致更多无用的计算，从而带来更多的时间开销；而较少次的 burn-in 迭代可能会导致算法给出的集合中存在一些无效的样本，使得最终得到的有效样本集合数量不够。不过，现有研究已经表明，Gibbs 采样以及 Metropolis 算法运行速度非常快，即使是处于较少次的 burn-in 迭代过程的条件下，也能得到较为理想的采样结果[136]。

3.4.5 分片采样

由前文的阐述可知，Metropolis 和 Metropolis-Hastings 算法的问题之一是难以确定从建议分布中采样时的最佳步长，而 Neal[113]和 Skilling 等[114]提出的分片采样则通过自适应地调整步长，使步长具有鲁棒性来处理这个问题。分片采样同样用于从复杂分布 $P(\boldsymbol{X})$中采集样本。分片采样的算法如图 10 所示。首先，我们在 $dom(\boldsymbol{X})$中随机采集一个初始样本点。然后，考虑在 d 维数据空间中的任意方向(线)来处理一维分布。对于这一思想，需要注意的是，在 Gibbs 采样中我们只考虑沿着条件分布中某一个维度的方向；然而，分片采样为研究者提供了选择数据空间中任何方向的自由。与拒绝采样的做法类似，我们从均匀分布中抽取一个随机数，即采样 $u_{i-1}\sim U(0, P(x_{i-1}))$。接着围绕这个采样点$u_{i-1}$考虑一个带有长度的分片(Slice)，或者称之为步长 δ。在这一步骤中采样点u_{i-1}可以处于分片中的任意位置，并不一定需要处在其中间。这一方法对分片的长度或步长也十分具备鲁棒性，只要现有分片的末端能够落在分布 $P(\boldsymbol{X})$之内，我们就能继续向分片端点的两侧延展并连接新的分片，如图 10 所示。

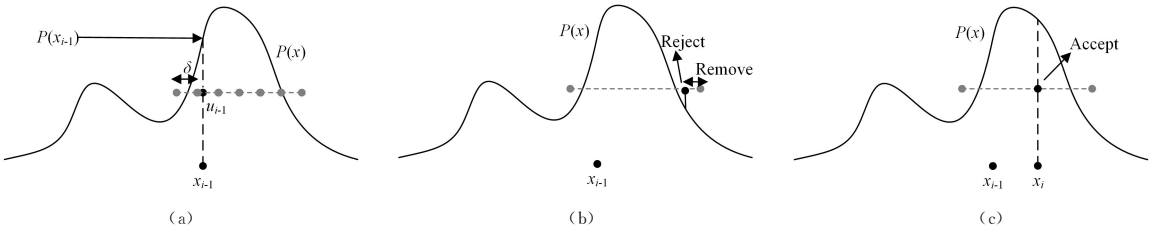

图 10 分片采样

Fig. 10 Fragment sampling

然后，我们在所有连接分片的范围内选择一个随机点，如果这个随机点落在了分布 $P(\boldsymbol{X})$的上方，那么就选择拒绝本次随机采样，并且根据这个随机点相对于x_{i-1}的位置关系，将随机点到这一侧的分片端点的这段线段从整个分片中移除(即以随机点为基准将分片分为两段，保留包含x_{i-1}的这一段而删除剩下的)。这一操作可以确保对整个连接分片的范围进行细化和精炼。如果本次选择的随机点被拒绝，那么就在细化后的分片范围内再次选择一个随机点，如果这个新的随机点依然落在了分布 $P(\boldsymbol{X})$的上方，则重复上述调整分片范围的操作；但如果这个新的随机点落在了分布 $P(\boldsymbol{X})$的下方，那么就选择接受其作为下一个采集的样本x_i。分片采样算法将重复执行上述过程直到我们获得所需的全部 n 个样本集合 $S=\{x_i\}_{i=1}^n$。

3.5 采样方法的讨论与总结

本节简要总结了现有的采样方法并回顾了这些的方法优缺点。采样方法分为两大类，即调查抽样和基于概率分布的蒙特卡洛采样。调查抽样中通常有数据点或向量的集合，研究人员从这些点中进行抽样。而蒙特卡洛采样则是从数据的分布函数中进行采样，用于近似分布上数据函数的期望或概率。现有的蒙特卡洛方法可以分为简单的蒙特卡罗方法和马尔可夫链蒙特卡洛方法。

在简单的蒙特卡洛方法中，每次迭代都独立于前一次迭代，因为每次迭代是使用简单样本分布独立执行的。简单的蒙特卡罗方法包括重要性采样和拒绝采样：重要性采样是使用另一种易于采样的分布来近似复杂分布上的数据函数的期望；拒绝采样是使用易于采样的上限分布从而实现在复杂分布中采样。

在马尔可夫链蒙特卡洛方法中，每次迭代都依赖于前一次迭代，因此采样不是盲目独立的，而是具有马尔可夫链的特性。现今研究中重要的 MCMC 方法包括 Metropolis 采样、Metropolis-Hastings 采样、Gibbs 采样和分片采样。Metropolis 采样算法使用易于采样的简单分布来抽取下一次的样本，同时这样的简单分布需要满足其均值与前一个样本相同，这一作为辅助的建议分布函数在 Metropolis 算法中是对称的。

Metropolis-Hastings 算法通过改良 Metropolis 算法中建议分布函数的接受率的大小,实现了放松建议分布函数需求的对称限制,并且放大了 Metropolis 算法的接受率,同时加快了马尔可夫链的收敛,使得 Metropolis 算法的应用范围更广。Gibbs 采样使用以其余坐标为条件的每个坐标的条件分布来抽取样本,Gibbs 采样可以被看作是概率为 1 的 Metropolis-Hastings 算法的一个特例。在 Metropolis 算法中,一个十分重要的问题就是如何选择合适的步长。与这一算法不同,分片采样是一种对步长具有鲁棒性的 MCMC 方法,其特点是考虑前一个样本两端的分片,并在这些分片的范围内抽取样本。蒙特卡洛方法的另一个问题就是随机游走行为通常比较缓慢,在数据范围很大或是特征维度很高时无法快速得到结果,导致算法效率不高。针对这一问题提出的哈密顿蒙特卡洛方法可以做到更快速地在数据范围内进行搜索。此外,也可以使用一些过松弛方法,例如 Adler 过松弛和有序过松弛等技术来加速 Gibbs 采样对数据范围的探索,尤其是在数据维度之间高度相关的情况下。

4　并行回火研究综述

4.1　简介

并行回火(Parallel Tempering,PT),或称之为副本交换,这一模拟技术的起源可以追溯到 1986 年 Swendsen 等的研究[137]。这篇论文中介绍了一种被称为副本蒙特卡洛的方法,这一方法会在一系列的温度下模拟目标分布的副本,并且相邻温度下的副本会进行部分配置信息的交换。如今,我们更为熟悉的具有完整配置信息交换特点的并行回火方法,则是由 Geyer 于 1991 年提出的,该文献中[138]给出了具体的阐述与公式。在并行回火方法被提出的最初,这一新方法的应用仅限于统计物理学中的问题。然而,当 Hansmann 在生物分子的蒙特卡罗模拟[139]中使用了该方法之后,Falcioni 等使用了并行回火进行 X 射线结构测定[140],Okamoto 等提出了并行回火的分子动力学版本[141]等。由此,并行回火这一方法在物理、化学、生物学、工程和材料科学等领域的应用得到了迅速的发展。

并行回火又名副本交换的马尔可夫链蒙特卡洛采样,是 MCMC 的一种推广,该算法的一般思想是模拟所要求取的目标系统的 M 个副本系统,每个副本与原系统完全相同,以使它们处于不同温度下对目标分布进行近似和采样。高温下的系统通常能对更大的范围区域进行采样,但其采样精度通常较低;而低温系统通常能够在空间局部区域进行较为精确的采样,但通常在计算机模拟的时间内可能会陷入局部的最小值。针对在高温和低温条件下进行采样的不同特点,并行回火方法通过允许不同温度下的系统交换完整配置的行为来实现良好的采样。因此,并行回火方法引入较高温度的系统从而确保较低温度系统可以访问采样空间内的每一组具有代表性的低温区域,即具有了跳出局部最优解的手段,这也是单一马尔可夫链采样方法中通常无法访问的区域,以此来实现较好的全局最优性。这一过程如图 11 所示。

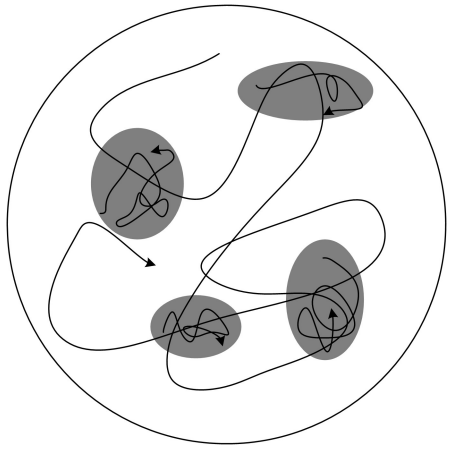

图 11　并行回火

Fig. 11　Parallel tempering

由定义可知,并行回火方法模拟 M 个而不是一个副本,这一需求与传统的 MCMC 方法相比会导致增大 M 倍的计算量,这种额外的消耗正是该方法在提出之初无法被迅速采用的原因之一。但随着应用的逐渐增加,最终我们发现并行回火模拟的效率比标准的单温度蒙特卡罗模拟要高出 $1/M$ 倍以上,效率提高的原因在于并行回火允许较低温度系统对搜索空间的更多区域进行采样。在传统 MCMC 方法中,如果只是对 M 倍长的单一温度模拟进行常规采样,那么较低温度系统也依然无法访问这些区域。除此之外,由于并行回火在各个温度的副本之间都能保持良好的独立性,因此这一方法也可以有效地利用大型 CPU 集群,使得其中不同的副本能够实现真正在硬件层面上的并行运行,从而大大缩短运行时间,提高算法效率。并行回火方法的另一个好处是可以生成处于一系列不同温度之下的模拟结果,这些结果具有特定的研究价值。综上所述,现今众多学者普遍认识到并行回火是一种十分有用且强大的计算方法。

在并行回火方法的实现中,一个容易引发争议的问题就是关于副本之间进行配置交换的细节。与此相关的问题包括:使用多少个不同的副本并且每个副本分别在什么温度下使用?相邻温度的副本之间应该尝试交换的频率是多少?以及在不同副本上花费的相对计算工作量有多大等。此外,还有一个值得思考的问题是,由于模拟一个大小为 N 的系统需要以 $\sqrt{N}$ 的规模增长的副本数量,为了克服这一复杂度的局限性,如何做到仅交换系统的一部分但依然能够合理保证模拟与运算的正确性。本文也将提出并阐述这些存在的争议点。

如今,由于并行回火方法在模拟领域的广泛应用,也出现了一些值得讨论的新议题。首先需要思考的一点就是温度可能并不总是最好的回火参数,并行回火也可以使用温度以外的顺序参数进行,最重要的是如何选择在交换行为下能够提供最有效的平衡的顺序参数。此外,研究表明并行回火方法也可以扩展到高维空间,这就引出了在多维空间下的多个有序参数中,如何保证同一模拟中的许多参数之间进行交换的正确性这一问题。

并行回火可以与大多数其他的模拟方法相结合,因为

如果正确地执行了交换步骤，就可以保持基础模拟系统的详细平衡或是平衡条件。由此可见，并行回火方法在计算机模拟领域中的适用范围是非常广泛的。这一特点更是带来了许多具有广阔前景的方向，比如并行回火未来也许能够与量子计算方法相结合。

4.2 基于蒙特卡洛采样的并行回火理论

在典型的并行回火模拟系统中，通常会有 M 个副本，每个副本都有一个不同的温度参数 T_i。通常地，有 $T_1<T_2<\cdots<T_m$，并且 T_1 即为目标副本的温度值。并行回火方法的精髓在于，当处于不同的单一温度值下的副本在经过固定步的蒙特卡洛移动之后，算法会随机选择两个相邻温度的副本，根据计算得到的接受率的值来决定是否选择这两个副本进行交换。需要注意的是，两个副本之间的交换会交换所有的参数信息。

以通用的并行回火模拟系统为例，在每个副本的温度值设置好之后，由于副本之间没有能量交互，这个大型副本整体的分配函数将由式(36)给出：

$$Q=\prod_{i=1}^{M}\frac{q_i}{N!}\int \mathrm{d}r_i^N \exp[-\beta_i U(r_i^N)] \tag{36}$$

其中 $q_i=\prod_{j=1}^{N}(2\pi m_j k_B T_i)^{3/2}$ 来自于对动量的积分；m_j 是原子 j 的质量；r_i^N 表示了系统 i 中的 N 个粒子的位置；$\beta_i=1/(k_B T_i)$ 是温度值的倒数；U 代表势能，或称之为不涉及动量的哈密顿量的一部分。如果在所有条件下执行交换行为的概率都相等，则副本 i 和 j 之间的交换将被接受的概率为：

$$A=\min\{1,\exp[+(\beta_i-\beta_j)(U(r_i^N)-U(r_j^N))]\} \tag{37}$$

交换行为通常发生在处于相邻温度下的系统之间，即在式(37)中有 $j=i+1$。

并行回火是统计力学中的一种精确方法，因为这一方法能够满足细致平稳或平稳条件[142]，其具体的满足形式取决于实现的方式。并行回火的这一特点是其与模拟退火相比的一个重要优势，因为在后一种方法中无法定义整体的平均值。并行回火是对单一温度下的系统中的任何一组蒙特卡洛运动的补充，并且这类单系统运动是通过在每一次尝试相邻温度下的交换来执行的。为了满足细致平稳条件，尽管在固定步数的单温度蒙特卡洛移动之后执行交换能够满足平稳的充分条件，交换操作也必须以一定的概率执行。一种交换和单温度蒙特卡洛移动的典型序列如图 12 所示。

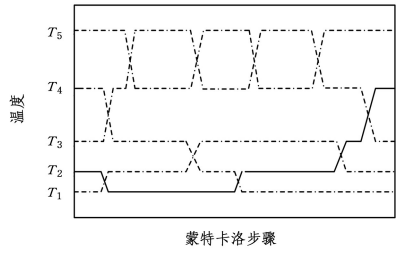

图 12 温度交换

Fig. 12 Temperature exchange

Kofke 对交换实验的平均接受率〈A〉进行了分析，并认为这个量应该与不同温度区域之间的熵差有关[143]。假设系统内部的能量遵循高斯分布，使用计算机模拟大量的这种类型的系统并进行仔细研究，如图 13 所示，Kofke 发现平均接受率〈A〉由以下式子给出：

$$\langle A\rangle=erfc\left[\left(\frac{1}{2}C_v\right)^{1/2}\frac{1-\beta_j/\beta_i}{(1+(\beta_j/\beta_i)^2)^{1/2}}\right] \tag{38}$$

其中，C_v 是恒定体积下的比热容，并且此数值在 β_i 和 β_j 之间的温度范围内可以假设为是一个常量。简而言之，实验的接受率取决于对较高温度进行采样的系统恰好位于对较低温度进行采样的系统来说比较重要的搜索空间区域的可能性。当考虑并行回火系统模拟的最佳温度选择时，关于接受率的这种理论分析十分有用，这一点将在下文进行详细阐述。

4.3 并行回火的最佳温度选择

我们希望能够以最少的计算工作量实现最好的可能采样结果，因此如何选择并行回火模拟中使用的副本数量和每个副本的温度非常重要。副本的最高温度必须足够高，以确保不会有副本陷入局部能量最小值，也就是局部最优解；同时使用的副本数量必须足够多，以确保所有相邻副本之间都能发生交换操作。现今的研究已经给出了一些关于副本的数量和温度的建议。从图 13 和式(32)可以清楚地看出，能量直方图必须重叠才能使得交换步骤被接受。Sugita 等和 Kofke 提出可以使所有不同副本的接受率保持一致，从而尝试确保每个副本在每个温度下能够经历相同的模拟时间[141,143]。Kofke 的实验表明，对于 C_v 在整个温度范围内恒定的系统，温度的几何级数(T_i/T_j = 常数)会导致相同的接受率。Sanbonmatsu 等[144]提出可以通过迭代求解以下式子来获得目标接受率：

$$A_{\text{target}}=\exp[\Delta\beta\Delta E] \tag{39}$$

其中，ΔE 是处于相邻温度下的系统平均能量的差值[144]。在自适应温度控制方案中，Schug 等已经提出并使用了用于调整不同系统的温度以确保接受率能够维持在可接受范围内的迭代方法[145]。Predescu 等扩展了这些方法，并且提出了一种将温度分配给副本的最佳方案[146]，该方案也基于迭代地调整系统的温度。在他们的方案中，系统的最低温度是固定的，而其他系统的温度则通过迭代求解由式(40)来确定：

$$\frac{\Delta E}{\sigma_m}\Big|_{T_j}=\left[\frac{\Delta E}{\sigma}\right]_{\text{target}} \tag{40}$$

对于每个温度 T_j 而言，有 $\sigma_m=[\sigma(T_j)+\sigma(T_j)]/2$，表示两个系统能量的平均偏差，我们可以通过选择目标值来达到所需要的接受率。

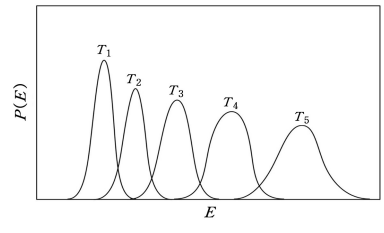

图 13 温度选择

Fig. 13 Temperature selection

Predescu 等还考虑了并行回火模拟中的最佳接受率和副本数[146]。他们通过研究工作中使用的案例发现，已实现的 20% 的接受率并没有提高模拟的性能。Kone 等对系统中温度区间的选择进行了分析，其中假设 C_v 在每个温度区间内为分段常数[147]。他们认为，尽管情况并非总是如此，但这种假设是合理的，其并不需要一个迭代方案，该方案会消耗宝贵的

CPU 时间并违反详细平衡。他们的分析基于最大化系统的均方位移σ^2，因为它在温度范围内执行随机游走。σ^2的值与接受的掉期数量成正比。通过在接受概率方面最大化σ^2，他们发现 23%的接受概率是最佳的。该值与 Predescu 等的经验确定的 20%惊人地相似。Kone 等建议调整温度间隔，以在模拟的初始平衡期间实现 23%的接受概率。这种方法似乎是在高效混合的平行回火模拟中选择温度间隔的有效方法。最近，Huse 等提出了一种类似的温度选择方案，该方案使用自适应反馈优化算法来最小化最低和最高温度之间的往返时间[148]。这种方法更直接地表征了高温和低温之间的混合温度系统。在复杂的情况下，配置交换的概率存在细微的瓶颈，往返时间可能比平均接受概率能更好地表征并行回火的整体效率。Huse 等的方法对于这种复杂的情况来说是一种很有效的方法，但一个相关的问题是应该在每个副本上花费多少模拟工作。例如，似乎低温复制品将受益于额外的模拟工作，因为低温下的相关时间更长。这个问题在文献中没有涉及。

4.4　并行回火与采样方法的结合与应用

并行回火的一般思想并不局限于不同温度下系统之间的交换，已有大量研究人员提出了许多基于交换可替代参数的方法，用以最大限度地减少影响正确采样的障碍。此外，并行回火可以与大多数的采样方法相结合，这一技术的使用使许多现有计算方法的采样有了很大的改进。

Fukunishi 等开发了一种哈密顿并行回火方法，并将其应用于生物分子系统[149]。在这种方法中，只有部分粒子之间的相互作用能量在不同的副本之间进行了调整。使用多个交换变量的并行回火首先由 Yan 等提出和开发[150-151]。他们没有考虑不同温度下的一维副本数组，而是建议使用 n 维数组，其中每个维度代表在不同副本之间存在差异的一个参数。他们的方案允许相同维度之内和不同维度之间的参数交换。与之相似，Sugita 等在分子动力学研究中利用了多维交换的思想[152]。Faller 等也在多规则系统中研究和实现了并行回火技术[153]。如今，并行回火在多规则模拟、自由能计算和伞状采样等领域中也得到了广泛的研究与应用[152-156]。并行回火与现有采样方法最有效的组合之一是使用基于 Wang-Landau 采样的状态密度方法，这一方法被提出后也取得了许多扩展和成果[157-163]。在长足的实验与研究中，并行回火已经与许多其他计算方法相结合，而且在几乎所有的情况下，这一方法的使用都带来了更好的采样结果和更高的计算准确性。

并行回火已经被成功地应用于许多一般性优化问题。Habeck 等提出了一种采样算法，用于搜索贝叶斯数据分析中出现的概率密度[164]。该方法利用了 Tsallis 统计数据，并通过解释折叠蛋白质的 NMR 实验数据证明了引入并行回火的有效性。在图像分析中，与模拟退火方法相比，并行回火已被证明能够做到将成功率提高两倍，同时将平均位置误差降低 1/2。并行回火还被应用于在如风险分析中经常出现的复杂和崎岖不平的势能面上，计算和定位全局的最小值。这一应用理念与使用并行回火技术来实现贝叶斯推理中存在的计算复杂后验分布函数的全局最优解有异曲同工之妙，这也正是本文在讨论贝叶斯推理时引入了并行回火方法的重要原因。

结束语　本文针对复杂分布模型下的推理问题的研究进行回顾，总结了包括变分推理和采样的两种主要解决方法。首先，本文对近似推理的问题定义和基本理论进行介绍，并总结了变分推理中的基础算法 CAVI、指数族下的变分推理和非条件共轭模型中的变分法；然后，介绍了包括蒙特卡洛、MCMC、MH 及其衍生算法在内的多种采样算法；随后结合先前的论述介绍了并行回火算法的思想。本文对不同算法的应用场景也分别进行了总结。

根据文中的论述，我们认为，未来复杂分布模型的相关算法可以在如何增加变分算法的更优下界、寻找精度和速度的平衡边界、探索变分法的统计学原理等方面更进一步，也可以在加快采样算法的游走速度、设计更优的过松弛方法，以及选取更好的并行回火温度序列等方面予以改进，并且可以在更多实际情境下尝试应用。

参 考 文 献

[1] PETERSON C, ANDERSON J R. A mean field learning algorithm for neural networks [J]. Complex Systems, 1987, 1(5): 995-1019.

[2] SAUL L K, JORDAN M I. Exploiting Tractable Substructures in Intractable Networks [J]. Advances in Neural Information Processing Systems, 1998, 8: 486-492.

[3] SAUL L K, JAAKKOLA T, JORDAN M I. Mean field theory for sigmoid belief networks [J]. Mean Field Theory for Sigmoid Belief Networks, 1996, 4(1): 61-76.

[4] HINTON G E, VAN CAMP D. Keeping the neural networks simple by minimizing the description length of the weights[C]// Proceedings of the Sixth Annual Conference on Computational Learning Theory. 1993: 5-13.

[5] NEAL R M, HINTON G E. A view of the EM algorithm that justifies incremental, sparse, and other variants[M]// Learning in Graphical Models. Dordrecht: Springer, 1998: 355-368.

[6] BISHOP C M, SVENSÉN M. Bayesian hierarchical mixtures of experts[J]. arXiv: 1212.2447, 2012.

[7] GELFAND A E, SMITH A F M. Sampling-based approaches to calculating marginal densities[J]. Journal of the American Statistical Association, 1990, 85(410): 398-409.

[8] HOFFMAN M D, BLEI D, WANG C, et al. Stochastic Variational Inference[J]. Journal of Machine Learning Research, 2013, 14: 1303-1347.

[9] MINKA T P. Expectation propagation for approximate Bayesian inference[J]. arXiv: 1301.2294, 2013.

[10] YEDIDIA J S, FREEMAN W T, WEISS Y. Bethe free energy, Kikuchi approximations, and belief propagation algorithms[J]. Advances in Neural Information Processing Systems, 2001, 13: 689.

[11] BISHOP C M, NASRABADI N M. Pattern recognition and machine learning[M]. New York: Springer, 2006.

[12] WINN J, BISHOP C M, JAAKKOLA T. Variational message passing[J]. Journal of Machine Learning Research, 2005, 6(4): 661-694.

[13] PEARL J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference(Judea Pearl)[J]. Artificial Intelligence, 1990, 48(8): 117-124.

[14] KNOWLES D, MINKA T. Non-conjugate variational message passing for multinomial and binary regression[J]. Advances in

Neural Information Processing Systems, 2011, 24: 110.
[15] GEMAN S, GEMAN D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984(6): 721-741.
[16] AMARI S I. Natural gradient works efficiently in learning[J]. Neural Computation, 1998, 10(2): 251-276.
[17] ROBBINS H, MONRO S. A Stochastic Approximation Method [J]. Annals of Mathematical Statistics, 1951, 22(3): 400-407.
[18] JAAKKOLA T S, JORDAN M I. A variational approach to Bayesian logistic regression models and their extensions[C]// Sixth International Workshop on Artificial Intelligence and Statistics. PMLR, 1997: 283-294.
[19] BLEI D M, LAFFERTY J D. A correlated topic model of Science [J]. Annals of Applied Statistics, 2007, 1(1): 17-35.
[20] BRAUN M, MCAULIFFE J. Variational inference for large-scale models of discrete choice [J]. Publications of the American Statistical Association, 2010, 105(489): 324-335.
[21] WAND M P, ORMEROD J T, PADOAN S A, et al. Mean field variational Bayes for elaborate distributions [J]. Bayesian Analysis, 2011, 6(4): 847-900.
[22] WANG C, BLEI D M. Variational Inference in Nonconjugate Models [J]. Journal of Machine Learning Research, 2012, 14(1): 1005-1031.
[23] BUGBEE B D, BREIDT F J, VAN DER WOERD M J. Laplace variational approximation for semiparametric regression in the presence of heteroscedastic errors[J]. Journal of Computational and Graphical Statistics, 2016, 25(1): 225-245.
[24] KNOWLES D, MINKA T. Non-conjugate variational message passing for multinomial and binary regression[J]. Advances in Neural Information Processing Systems, 2011, 24: 110.
[25] JAAKKOLA T S, JORDAN M I. Bayesian parameter estimation via variational methods [J]. Statistics & Computing, 2000, 10(1): 25-37.
[26] WAND M P. Fully simplified multivariate normal updates in non-conjugate variational message passing[J]. Journal of Machine Learning Research, 2014, 15: 1351-1369.
[27] TAN L, NOTT D J. A stochastic variational framework for fitting and diagnosing generalized linear mixed models[J]. Bayesian Analysis, 2012, 9(4): 963-1004.
[28] ROHDE D, WAND M P. Semiparametric mean field variational Bayes: General principles and numerical issues[J]. The Journal of Machine Learning Research, 2016, 17(1): 5975-6021.
[29] BERNARDO J M, BAYARRI M J, BERGER J O, et al. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures[J]. Bayesian Statistics, 2003, 7: 453-464.
[30] NOTT D J, KOHN R. Regression Density Estimation With Variational Methods and Stochastic Approximation[J]. Journal of Computational & Graphical Statistics, 2012, 21(3): 797-820.
[31] PAISLEY J, BLEI D, JORDAN M. Variational Bayesian inference with stochastic search[J]. arXiv: 1206.6430, 2012.
[32] WINGATE D, WEBER T. Automated variational inference in probabilistic programming[J]. arXiv: 1301.1299, 2013.
[33] KINGMA D P, WELLING M. Auto-encoding variational bayes [J]. arXiv: 1312.6114, 2013.
[34] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models[C]// International Conference on Machine Learning. PMLR, 2014: 1278-1286.
[35] RANGANATH R, TRAN D, BLEI D. Hierarchical variational models[C]// International Conference on Machine Learning. PMLR, 2016: 324-333.
[36] TRAN D, RANGANATH R, BLEI D M. The variational Gaussian process[J]. arXiv: 1511.06499, 2015.
[37] LOGSDON B A, HOFFMAN G E, MEZEY J G. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis[J]. BMC Bioinformatics, 2010, 11(1): 12-23.
[38] SANGUINETTI G, LAWRENCE N, RATTRAY M. Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities[J]. Bioinformatics, 2006, 22(22): 2775-2781.
[39] XING E P, WEI W U, JORDAN M I, et al. LOGOS: a modular Bayesian model for de novo motif detection[J]. Journal of Bioinformatics & Computational Biology, 2003, 2(1): 127-154.
[40] VLADIMIR J, NEBOJSA J, CHRIS M, et al. Efficient approximations for learning phylogenetic HMM models from data[J]. Bioinformatics, 2004, 20(suppl_1): i161-i168.
[41] RAJ A, STEPHENS M, PRITCHARD J K. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets[J]. Genetics, 2014, 197(2): 573-589.
[42] STEGLE O, PARTS L, DURBIN R, et al. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies[J]. PLOS Computational Biology, 2010, 6(5): e1000770.
[43] BISHOP C M, WINN J M. Non-linear Bayesian image modelling [C]// European Conference on Computer Vision. Berlin: Springer, 2000: 3-17.
[44] PAWAN KUMAR M, TORR P H S, ZISSERMAN A. Learning layered motion segmentations of video[J]. International Journal of Computer Vision, 2008, 76(3): 301-319.
[45] CHAN A B, VASCONCELOS N. Layered Dynamic Textures [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2009, 31(10): 1862-1879.
[46] LIKAS A C, GALATSANOS N P. A Variational Approach for Bayesian Blind Image Deconvolution[J]. IEEE Transactions on Signal Processing, 2004, 52(8): 2222-2233.
[47] YU T, WU Y. Decentralized multiple target tracking using netted collaborative autonomous trackers[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR'05). IEEE, 2005: 939-946.
[48] VERMAAK J, LAWRENCE N D, PEREZ P. Variational inference for visual tracking[C]// 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2003: 33-43.
[49] CUMMINS M, NEWMAN P. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance[J]. The International Journal of Robotics Research, 2008, 27(6): 647-665.
[50] RAMOS F, UPCROFT B, KUMAR S, et al. A Bayesian approach for place recognition[J]. Robotics & Autonomous Systems, 2012, 60(4): 487-497.
[51] SUDDERTH E, JORDAN M. Shared segmentation of natural scenes using dependent Pitman-Yor processes[J]. Advances in

Neural Information Processing Systems,2008,21:1585-1592.
[52] ROBERTS S J,PENNY W D. Variational Bayes for generalized autoregressive models[J]. IEEE Transactions on Signal Processing,2002,50(9):2245-2257.
[53] FLANDIN G,PENNY W D. Bayesian fMRI data analysis with sparse spatial basis function priors[J]. NeuroImage, 2007, 34(3):1108-1125.
[54] HARRISON L M,GREEN G. A Bayesian spatiotemporal model for very large data sets[J]. Neuroimage, 2010, 50(3): 1126-1141.
[55] WOOLRICH M W,BEHRENS T,BECKMANN C F,et al. Multilevel linear modelling for FMRI group analysis using Bayesian inference[J]. NeuroImage,2004,21(4):1732-1747.
[56] SATO M A,YOSHIOKA T,KAJIHARA S,et al. Hierarchical Bayesian estimation for MEG inverse problem[J]. Neuroimage, 2004,23(3):806-826.
[57] ZUMER J M,ATTIAS H T,SEKIHARA K,et al. A probabilistic algorithm integrating source localization and noise suppression for MEG and EEG data[J]. Neuroimage,2007,37(1):102-115.
[58] LASHKARI D,SRIDHARAN R,VUL E,et al. Search for patterns of functional specificity in the brain:A nonparametric hierarchical Bayesian model for group fMRI data[J]. Neuroimage, 2012,59(2):1348-1368.
[59] NATHOO F S,BABUL A,MOISEEV A,et al. A variational Bayes spatiotemporal model for electromagnetic brain mapping [J]. Biometrics,2014,70(1):132-143.
[60] SYKACEK P,ROBERTS S J,STOKES M. Adaptive BCI based on variational Bayesian Kalman filtering:an empirical evaluation [J]. IEEE Transactions on Biomedical Engineering, 2004, 51(5):719-727.
[61] MANNING J R,RANGANATH R,NORMAN K A,et al. Topographic Factor Analysis:A Bayesian Model for Inferring Brain Networks from Neural Data[J]. Plos One,2014,9(5):e94914.
[62] GERSHMAN S J,BLEI D M,NORMAN K A,et al. Decomposing spatiotemporal brain patterns into topographic latent sources[J]. Neuroimage,2014,98:91-102.
[63] LIANG P,JORDAN M I,KLEIN D. Probabilistic grammars and hierarchical Dirichlet processes[J]. The Handbook of Applied Bayesian Analysis,2009,104:776-819.
[64] KURIHARA K,SATO T. Variational Bayesian grammar induction for natural language[C]// International Colloquium on Grammatical Inference. Berlin:Springer,2006:84-96.
[65] NASEEM T,CHEN H,BARZILAY R,et al. Using universal linguistic knowledge to guide grammar induction[J]. Empirical Methods in Natural Language Processing,2010,1:1234-1244.
[66] COHEN S B,SMITH N A. Covariance in Unsupervised Learning of Probabilistic Grammars[J]. Journal of Machine Learning Research,2010,11(6):3017-3051.
[67] YOGATAMA D,WANG C,ROUTLEDGE B R,et al. Dynamic language models for streaming text[J]. Transactions of the Association for Computational Linguistics,2014,2:181-192.
[68] BERNARDO J M,BAYARRI M J,BERGER J O,et al. The variational Bayesian EM algorithm for incomplete data:with application to scoring graphical model structures[J]. Bayesian Statistics,2003,7:453-464.
[69] WANG P,BLUNSOM P. Collapsed variational Bayesian inference for hidden Markov models[C]// Artificial Intelligence and Statistics. PMLR,2013:599-607.
[70] REYES-GOMEZ M J,ELLIS D P W,JOJIC N. Multiband audio modeling for single-channel acoustic source separation[C]// 2004 IEEE International Conference on Acoustics,Speech,and Signal Processing. IEEE,2004,5:641-644.
[71] DENG L. Switching dynamic system models for speech articulation and acoustics[M]// Mathematical Foundations of Speech and Language Processing. New York:Springer,2004:115-133.
[72] BRAUN M,MCAULIFFE J. Variational inference for large-scale models of discrete choice[J]. Publications of the American Statistical Association,2010,105(489):324-335.
[73] BROEK B V D,WIEGERINCK W,KAPPEN B. Graphical model inference in optimal control of stochastic multi-agent systems.[J]. Journal of Artificial Intelligence Research,2008,32: 95-122.
[74] FURMSTON T,BARBER D. Variational methods for Reinforcement Learning[J]. Journal of Machine Learning Research, 2009,9:241-248.
[75] HOFMAN J M,WIGGINS C H. A Bayesian Approach to Network Modularity[J]. Physical Review Letters,2007,100(25): 258701.
[76] AIROLDI E M,BLEI D M,FIENBERG S E,et al. Mixed Membership Stochastic Blockmodels[J]. Journal of Machine Learning Research,2008,9:1981-2014.
[77] REGIER J,MILLER A,MCAULIFFE J,et al. Celeste:Variational inference for a generative model of astronomical images [C]// International Conference on Machine Learning. PMLR, 2015:2095-2103.
[78] EROSHEVA E A,FIENBERG S E,JOUTARD C. Describing disability through individual-level mixture models for multivariate binary data[J]. The Annals of Applied Statistics,2007, 1(2):502-537.
[79] GRIMMER J. An Introduction to Bayesian Inference via Variational Approximations[J]. Political Analysis, 2011,19(1): 32-47.
[80] BARBER D,DE V L P. Variational Cumulant Expansions for Intractable Distributions[J]. Computer Science,1999,10(1): 435-455.
[81] LEISINK M,KAPPEN H. A tighter bound for graphical models [J]. Advances in Neural Information Processing Systems,2000, 13(9):2149-2171.
[82] WATSON-GANDY A. Elements of Sampling Theory[J]. Journal of the Operational Research Society,1976,27(1):138-139.
[83] SMITH T. The Foundations of Survey Sampling:A Review[J]. Journal of the Royal Statistical Society,1976,139(2):183-204.
[84] HANSEN M H. Some history and reminiscences on survey sampling[J]. Statistical Science,1987,2(2):180-190.
[85] CHAUDHURI A,STENGER H. Survey Sampling:Theory and Methods[M]. CRC Press,1992:241-242.
[86] UEBE G. Yves Tillé:Sampling algorithms[J]. AStA Advances in Statistical Analysis,2009,93(1):117-118.
[87] CHAUDHURI A,STENGER H. Survey Sampling:Theory and Methods[M]. CRC Press,1992:241-242.
[88] FULLER W A. Sampling statistics[M]. John Wiley & Sons. 2011.
[89] HIBBERTS M,BURKE JOHNSON R,HUDSON K. Common

survey sampling techniques[M]//Handbook of survey methodology for the social sciences. Springer, New York, NY, 2012: 53-74.
[90] SINGH R, MANGAT N S. Stratified sampling[M]// Elements of Survey Sampling. Dordrecht: Springer, 1996: 102-144.
[91] BREWER K, GREGOIRE T G. Introduction to Survey Sampling [J]. Handbook of Statistics, 2009, 29: 9-37.
[92] MICHAEL B J. The Future of Survey Sampling [J]. Public Opinion Quarterly, 2011(5): 872-888.
[93] HANNEKE S, XING E P. Network completion and survey sampling[C]// Artificial Intelligence and Statistics. PMLR, 2009: 209-215.
[94] HECKATHORN D D, CAMERON C J. Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling [J]. Annual Review of Sociology, 2017, 43(1): 101-119.
[95] WATSON-GANDY A. Elements of Sampling Theory[J]. Journal of the Operational Research Society, 1976, 27(1): 138-139.
[96] EFRON B, TIBSHIRANI R J. An introduction to the bootstrap [M]. FLORIDA: CRC press, 1994.
[97] WATSON-GANDY A. Elements of Sampling Theory[J]. Journal of the Operational Research Society, 1976, 27(1): 138-139.
[98] ERICKSON B H, NOSANCHUK T A. Applied network sampling[J]. Social Networks, 1983, 5(4): 367-382.
[99] GOODMAN L A. Snowball Sampling[J]. Annals of Mathematical Statistics, 1961, 32(1): 148-170.
[100] GHOJOGH B, CROWLEY M. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial [J]. arXiv: 1905. 12787, 2019.
[101] MACKAY D J C. Introduction to monte carlo methods[M]// Learning in Graphical Models. Dordrecht: Springer, 1998: 175-204.
[102] CAFLISCH R E. Monte carlo and quasi-monte carlo methods [J]. Acta Numerica, 1998, 7: 1-49.
[103] HAMMERSLEY J. Monte carlo methods [M]. Berlin: Springer Science & Business Media, 2013.
[104] KROESE D P, TAIMRE T, BOTEV Z I. Handbook of monte carlo methods[M]. Berlin: John Wiley & Sons, 2013.
[105] MACKAY D J C. Information Theory, Inference, and Learning Algorithms[J]. IEEE Transactions on Information Theory, 2003, 50(10): 1461-1462.
[106] GLYNN P W, IGLEHART D L. Importance Sampling for Stochastic Simulations[J]. Management Science, 1989, 35(11): 1367-1392.
[107] WELLS M T, CASELLA G, ROBERT C P. Generalized accept-reject sampling schemes[M]// A Festschrift for Herman Rubin. Institute of Mathematical Statistics, 2004: 342-348.
[108] MURRAY I. Advances in Markov chain Monte Carlo methods [M]. London: University of London, 2007.
[109] KOLLER D, FRIEDMAN N. Probabilistic graphical models: principles and techniques[M]. Massachusetts: MIT Press, 2009.
[110] METROPOLIS N, ROSENBLUTH A W, ROSENBLUTH M N, et al. Equation of state calculations by fast computing machines[J]. Journal of Chemical Physics, 1953, 21(6): 1087-1092.
[111] HASTINGS W K. Monte Carlo sampling methods using Markov chains and their applications[J]. Biometrika, 1970, 57: 97-109.
[112] GEMAN S, GEMAN D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984(6): 721-741.
[113] NEAL R M. Slice sampling[J]. Annals of Statistics, 2003, 31(3): 705-767.
[114] SKILLING J, MACKAY D J C. [Slice Sampling]: Discussion [J]. Annals of Statistics, 2003: 31(3): 753-755.
[115] SPITZER F L. Principles of Random Walk[M]. Berlin: Principles of Random Walk, 1975.
[116] DUANE S, KENNEDY A D, PENDLETON B J, et al. Hybrid monte carlo[J]. Physics Letters B, 1987, 195(2): 216-222.
[117] ADLER S L. Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadratic actions[J]. Physical Review D Particles & Fields, 1981, 23(12): 2901-2904.
[118] NEAL R M. Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation[M]// Learning in Graphical Models. Dordrecht: Springer, 1998: 205-228.
[119] NEWMAN M E J, BARKEMA G T. Monte Carlo methods in statistical physics[M]. Clarendon Press, 1999.
[120] KURT B. Monte Carlo methods in statistical physics[J]. Topics in Current Physics, 1990, 46(2): 252-253.
[121] GLASSERMAN P. Monte Carlo methods in financial engineering[M]. New York: Springer, 2004.
[122] BARTO A, DUFF M. Monte Carlo Matrix Inversion and Reinforcement Learning[J]. Advances in Neural Information Processing Systems, 1994, 6(6): 687-694.
[123] WANG Y, WON K S, HSU D, et al. Monte carlo bayesian reinforcement learning[J]. arXiv: 1206. 6449, 2012.
[124] MONTAGUE P R. Reinforcement learning: an introduction, by Sutton, RS and Barto, AG[J]. Trends in Cognitive Sciences, 1999, 3(9): 360.
[125] NEAL R M. Annealed importance sampling[J]. Statistics and Computing, 2001, 11(2): 125-139.
[126] GILKS W R, WILD P. Adaptive rejection sampling for Gibbs sampling[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1992, 41(2): 337-348.
[127] TEH D. Concave-Convex Adaptive Rejection Sampling[J]. Journal of Computational & Graphical Statistics, 2011, 20(3): 670-691.
[128] MARTINO L, MÍGUEZ J. A generalization of the adaptive rejection sampling algorithm[J]. Statistics and Computing, 2011, 21(4): 633-647.
[129] DELIGIANNIDIS G, DOUCET A, RUBENTHALER S. Ensemble rejection sampling[J]. arXiv: 2001. 09188, 2020.
[130] AZADI S, OLSSON C, DARRELL T, et al. Discriminator rejection sampling[J]. arXiv: 1810. 06758, 2018.
[131] GROVER A, GUMMADI R, LAZARO-GREDILLA M, et al. Variational rejection sampling[C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2018: 823-832.
[132] SCOLLNIK D P M. An introduction to Markov Chain Monte Carlo methods and their actuarial applications[C]// Proceedings of the Casualty Actuarial Society. 1996, 83: 114-165.
[133] GELFAND A E. Gibbs sampling[J]. Journal of the American statistical Association, 2000, 95(452): 1300-1304..
[134] WU T T, LANGE K. Coordinat descent algorithms for lasso penalized regression[J]. Annals of Applied Stats, 2008, 2(1): 224-244.
[135] WRIGHT S J. Coordinate descent algorithms[J]. Mathematical

Programming,2015,151(1):3-34.
[136]DWIVEDI R,CHEN Y,WAINWRIGHT M J,et al. Log-concave sampling:Metropolis-Hastings algorithms are fast! [C]// Conference on learning theory. PMLR,2018:793-797.
[137]SWENDSEN R H,WANG J S. Replica Monte Carlo Simulation of Spin-Glasses[J]. Physical Review Letters,1986,57(21):2607-2609.
[138]GEYER C J. Markov chain Monte Carlo Maximum Likelihood [J]. Computing Science & Statistics,1992,91(8):133-169.
[139]HANSMANN U. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules[J]. Chemical Physics Letters,1997,281(1/2/3):140-150.
[140]FALCIONI M,DEEM M W. A Biased Monte Carlo Scheme for Zeolite Structure Solution[J]. The Journal of Chemical Physics,1998,110(3):1754-1766.
[141]SUGITA Y,OKAMOTO Y. Replica-exchange molecular dynamics method for protein folding[J]. Chemical Physics Letters,1999,314(1/2):141-151.
[142]MANOUSIOUTHAKIS V I,DEEM M W. Strict Detailed Balance is Unnecessary in Monte Carlo Simulation[J]. The Journal of Chemical Physics,1999,110(6):2753-2756.
[143]KOFKE D A. On the acceptance probability of replica-exchange Monte Carlo trials[J]. Journal of Chemical Physics,2002,117(15):6911-6914.
[144]SANBONMATSU K Y,GARCÍA A E. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics[J]. Proteins Structure Function and Bioinformatics,2002,46(2):225-234.
[145]SCHUG A,HERGES T,WENZEL W. All-atom folding of the three-helix HIV accessory protein with an adaptive parallel tempering method[J]. Proteins Structure Function & Bioinformatics,2004,57(4):792-798.
[146]PREDESCU C,PREDESCU M,CIOBANU C V. On the efficiency of exchange in parallel tempering Monte Carlo simulations[J]. The Journal of Physical Chemistry B,2005,109(9):4189-4196.
[147]KONE A,KOFKE D A. Selection of temperature intervals for parallel-tempering simulations[J]. Journal of Chemical Physics,2005,122(20):2607.
[148]TREBST S,HUSE D A,TROYER M. Optimizing the ensemble for equilibration in broad-histogram Monte Carlo simulations [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2004,70(4):46701.
[149]FUKUNISHI H,WATANABE O,TAKADA S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems:Application to protein structure prediction[J]. Journal of Chemical Physics,2002,116(20):9058-9067.
[150]YAN Q,DE PABLO J J. Hyper-parallel tempering Monte Carlo:Application to the Lennard-Jones fluid and the restricted primitive model[J]. Journal of Chemical Physics,1999,111(21):9509-9516.
[151]YAN Q,PABLO J D. Hyperparallel tempering Monte Carlo simulation of polymeric systems[J]. Journal of Chemical Physics,2000,113(3):1276-1282.
[152]SUGITA Y,OKAMOTO Y. Replica-exchange molecular dynamics method for protein folding[J]. Chemical Physics Letters,1999,314(1/2):141-151.
[153]FALLER R,YAN Q,PABLO J D. Multicanonical parallel tempering[J]. Journal of Chemical Physics,2002,116(13):5419-5423.
[154]MITSUTAKE A,SUGITA Y,OKAMOTO Y. Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. I. Formulation and benchmark test[J]. Journal of Chemical Physics,2003,118(14):6664-6675.
[155]MURATA K,SUGITA Y,OKAMOTO Y. Free energy calculations for DNA base stacking by replica-exchange umbrella sampling[J]. Chemical Physics Letters,2004,385(1/2):1-7.
[156]OKAMOTO Y. Generalized-Ensemble Algorithms:Enhanced Sampling Techniques for Monte Carlo and Molecular Dynamics Simulations[J]. Journal of Molecular Graphics & Modelling,2004,22(5):425-439.
[157]WANG F,LANDAU D P. Efficient,Multiple-Range Random Walk Algorithm to Calculate the Density of States[J]. Physical Review Letters,2001,86(10):2050-2053.
[158]YAN Q,PABLO J D. Fast Calculation of the Density of States of a Fluid by Monte Carlo Simulations[J]. Physical Review Letters,2003,90(3):35701.
[159]FASNACHT M,SWENDSEN R H,ROSENBERG J M. Adaptive integration method for Monte Carlo simulations[J]. Physical Review E,2004,69(5 Pt 2):56704.
[160]KIM E B,FALLER R,YAN Q,et al. Potential of Mean Force between a Spherical Particle Suspended in a Nematic Liquid Crystal and a Substrate[J]. The Journal of Chemical Physics,2002,117(16):7781-7787.
[161]RATHORE N,KNOTTS IV T A,DE PABLO J J. Density of states simulations of proteins[J]. Journal of Chemical Physics,2003,118(9):4285-4290.
[162]RATHORE N,YAN Q,PABLO J D. Molecular simulation of the reversible mechanical unfolding of proteins[J]. Journal of Chemical Physics,2004,120(12):5781.
[163]MASTNY E A,PABLO J D. Direct calculation of solid-liquid equilibria from density-of-states Monte Carlo simulations[J]. Journal of Chemical Physics,2005,122(12):9352.
[164]HABECK M,NILGES M,RIEPING W. Replica-Exchange Monte Carlo Scheme for Bayesian Data Analysis[J]. Physical Review Letters,2005,94(1):18105.

ZHAN Jin, born in 1998, postgraduate, is a member of China Computer Federation. Her main research interests include machine learning, Bayesian learning and artificial intelligence.

CHENG Yurong, born in 1989, Ph.D, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include queries and analysis over uncertain graphs, knowledge bases, social networks, and spatio-temporal databases.

(责任编辑:何杨)