马氏链蒙特卡洛方法

孙望涛

清华大学电子工程系

2019 年 11 月

摘　要

本文主要探讨了马尔可夫链蒙特卡洛方法（Markov Chain Monte Carlo, MCMC）的仿真应用：针对二维高斯分布与 Potts 模型，分别使用 Metropolis-Hasting(MH)方法，Gibbs/Swendsen–Wang(SW)方法对其进行采样，分析其统计特征。根据仿真结果对比分析了各算法的优缺点及适用范围。

关键词：MCMC 方法，Potts 模型，MH 算法，Gibbs 算法，SW 算法

Abstract

This paper mainly discusses the simulation application of Markov Chain Monte Carlo (MCMC): for the two-dimensional Gaussian distribution and Potts model, using the Metropolis-Hasting (MH) method, Gibbs/Swendsen–Wang (SW) The method samples it and analyzes its statistical characteristics. According to the simulation results, the advantages and disadvantages of each algorithm and its application range are compared and analyzed.

Keywords: MCMC method, Potts model, MH algorithm, Gibbs algorithm, SW algorithm

1.介绍

马尔可夫链(Markov Chain) 是一组具有马尔可夫性质的离散随机变量的集合：在给定当前状态的条件下，未来的状态与过去的状态独立。具体可表示为：$p(X_{t+1}|X_t,\ldots,X_1)=p(X_{t+1}|X_t)$

有关马尔可夫链的性质在此不再赘述。

蒙特卡罗方法(Monte Carlo method)是一种著名的数值计算方法，通过使用随机数（或更常见的伪随机数）来解决计算问题。本文将要介绍的是两者的结合：马氏链蒙特卡洛方法(Markov Chain Monte Carlo method)。

对于给定分布π的随机变量，我们希望生成一组样本来模拟其分布；然而，在许多问题中无法直接获得给定分布的独立样本。MCMC 方法的基本技巧是通过生成一个相关的马氏链（一串相关的随机变量），使得其平稳分布为给定的π。

Metropolis 算法是马尔科夫链蒙特卡罗的基石。它是由 Metropolos 等人在 1953 年的一篇仅 4 页的文章中提出。Metropolis 算法用一个对称的建议分布 T(x,y)来产生一个潜在的转移点，然后根据特定的接受拒绝方法来决定是否转移到该潜在点。最初的 Metropolis 算法将 T 取为对称的函数，而 Metropolis-Hasting 方法将之推广到非对称的 T。

然而在 Metropolis 算法在许多应用中取的都是对称的局部均匀移动，在高维是计算量较大，迭代次数较少时难以迅速收敛到相应平稳分布。Gibbs 采样(Geman&Geman,1984)在 MH 算法的基础上重新寻找了细致平稳条件，将其改进为在多个条件分布上轮换跳转，产生一种新的迭代方式。

对于 Potts 模型，由于其体系状态数随着维数增大急剧增大，跳转概率极小，因此传统的 MH 方法迭代次数较大。对此 Swendsen, R. H.和 Wang, J.-S.在 1987 年提出 SW 算法，在原模型的基础上增加了一维 bond 变量，在增广的模型上使用 Gibbs 抽样方法估计了 Potts 模型的归一化常数。

2.模型

本文主要研究 MCMC 方法在以下两个模型上的应用：

（1）二维高斯分布

对服从二维高斯分布

$$\mathcal{N}\left\{\begin{pmatrix}x_1\\x_2\end{pmatrix}\middle|\begin{pmatrix}5\\10\end{pmatrix},\begin{pmatrix}1&1\\1&4\end{pmatrix}\right\}$$

的随机变量 X，我们可计算其两维的均值分别为 5，10，方差分别为 1，2，相关系数为 0.5。其概率密度函数(Probabily Density Function, PDF)为

$$f(x) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

其中$\mu = \begin{pmatrix} 5 \\ 10 \end{pmatrix}$，$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$。

（2）Potts 模型

Potts 模型源于统计物理，是一种重要的概率模型。对 Potts 模型的归一化常数的估计，代表了一大类科学计算问题，至今仍是非常活跃的一个研究课题。模型的具体表述如下：

定义一个无向图 G，其中的点集为一系列离散随机变量(s1, . . . ,sn)，每个点的状态取自 q 个不同的状态中；定义边集为一些(si，sj)的二元组。Kronecker 积$\delta_{si,sj}$在 si=sj 时为 1，否则为 0。在此基础上，定义该状态的能量 u 及概率 p 如下：

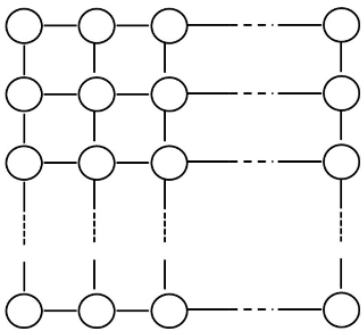
$$p(x) = \frac{1}{Z(T)} exp[-\frac{u(x)}{T}],$$

$$u(x) = \sum_{\substack{i \leftrightarrow j \\ i,j=1,\ldots,K}} 1(x_i = x_j),$$

其中 T 为体系温度，i←→j 表示 i，j 有边相连。归一化常数 Z(T)表示为：

$$Z(T) = \sum_x exp[-\frac{u(x)}{T}]$$

本文以 q = 10, K = 20，G 为 K*K 的矩阵模型为例研究如何通过采样来估计 Potts 模型的归一化常数 Z(T)。



3.基本理论

（1）Metropolis-Hasting 方法

对于任意给定的概率分布π，需要构造马氏链的单步概率转移矩阵 P，使得该马氏链的平稳分布为π，即$\pi = P\pi$。考虑该条件的加强，即细致平稳条件：

$$\pi_i P_{i,j} = \pi_j P_{j,i}$$

对该等式两侧关于 j 求和，并根据 P 的行归一性，即可得平稳分布等式，说明此等式为平稳分布的充分条件。

在此基础上，设计易于实现的条件概率转移矩阵 T(i, j)，依据以下步骤生成符合要求的马氏链：

初始化，随机取 x(0)作为第一个样本
for n = 0:N-1
取 xp ~ T(x(n), •)
计算 p = min{1, $\pi(xp)T(xp,x(n))/\pi(x(n))T(x(n),xp)$}
抽取 u ~ uniform(0,1)
若 u < p，则令 x(n+1) = xp，否则 x(n+1)=x(n)
end for

接下来证明此算法产生的 x 序列满足细致平稳条件：

$$\pi(i)P(i,j) = \pi(i)T(i,j)p(i,j) = \pi(i)T(i,j)\min\left\{1,\frac{\pi(j)T(j,i)}{\pi(i)T(i,j)}\right\}$$

$$= \min\{\pi(i)T(i,j),\pi(j)T(j,i)\} = \pi(j)T(j,i)\min\left\{1,\frac{\pi(i)T(i,j)}{\pi(j)T(j,i)}\right\}$$

$$= \pi(j)T(j,i)\mathrm{p(j,i)} = \pi(j)P(j,i)$$

（2）Gibbs 采样

以二维吉布斯采样（即给定二维随机变量的分布）为例，按以下步骤生成符合要求的马氏链：

初始化，随机取 X(0)=[x_0 y_0]’作为第一个样本
for n = 0:N-1
取 xp ~ $\pi(x|y_n)$
计算 p = min{1, $\pi([xp\ y_n]')/\pi(X(n))$}
抽取 u ~ uniform(0,1)
若 u < p，则令 x_{n+1} = xp，否则 x_{n+1} = x_n
取 yp ~ $\pi(y|x_n)$

计算 p = min{1, $\pi([yp\ x_n]')/\pi(X(n))$}

抽取 u～uniform(0,1)

若 u < p，则令 y_{n+1} = yp，否则 y_{n+1} = y_n

end for

同样证明此算法产生的 x 序列满足细致平稳条件：

由：$p(x_1,y_1)P(y_2|x_1)=p(x_1)P(y_1|x_1)P(y_2|x_1)$

$p(x_1,y_2)P(y_1|x_1)=p(x_1)P(y_2|x_1)P(y_1|x_1)$

可得$p(x_1,y_1)\cdot P(y_2|x_1)=p(x_1,y_2)\cdot P(y_1|x_1)$

4.算法设计与分析

（1）二维高斯分布的采样

由于模型极为简单，故无需增加操作，直接应用采样算法即可。

1）MH 算法

对于已知的高斯分布，我们需要设计一个推荐分布 Q，使得生成序列在二维平面上游走。朴素的想法是选择对称的推荐方案：

方案 1：随机推荐，即无视上一时刻的 xn，在均值附近一定范围内均匀抽取推荐矢量 xp，由于 Q 矩阵对称，故仿真实现时转移概率直接统一设为 1。

方案 2：均匀 xy 转移推荐，即将分布的中心移到 xn，在 xn 附近矩形区间内均匀抽取推荐矢量 xp，同样由于 Q 矩阵对称，故仿真实现时转移概率直接统一设为 1。

但对称推荐方案等于忽视了已知分布的先验知识，于是有以下设计：

方案 3：偏向均值的均匀 xy 转移推荐，在方案 2 的基础上，推荐矢量 xp = xn+随机数+k(u-xn)，其中第三项为偏向均值的转移项，这是由于先验知识告诉我们马氏链最终会收敛到均值附近，因此加这一项可以加快收敛速度。由于此时 Q 不对称，故需计算转移概率，但由于转移分布在有限区间内均匀，超出此区间则为 0，故经实验设定落在有限区间内为 1，否则为 0.1（即若 xp 转移到比 xn 更靠近均值的位置，需补偿 10 倍的接收概率）

方案 4：偏向均值的均匀法向/切向转移推荐，在方案 3 的基础上，推荐矢量 xp = xn+随机数+k(u-xn)，其中随机数向不再沿 x,y 方向均匀分布，而是沿法向（指向均值），切向（垂直法向）均匀分布。由于此时 Q 不对称，故也需计算转移概率，同 3 转移分布在有限区间内均匀，超出此区间则为 0，故经实验设定落在有限区间内为 1，否则为 0.1。

对上述四种方案进行实验仿真，各采 50000 个样本，计算接受率，舍弃前 n 个后剩余的样本计算其相关系数ρ，绘制$\rho-n$曲线如下。

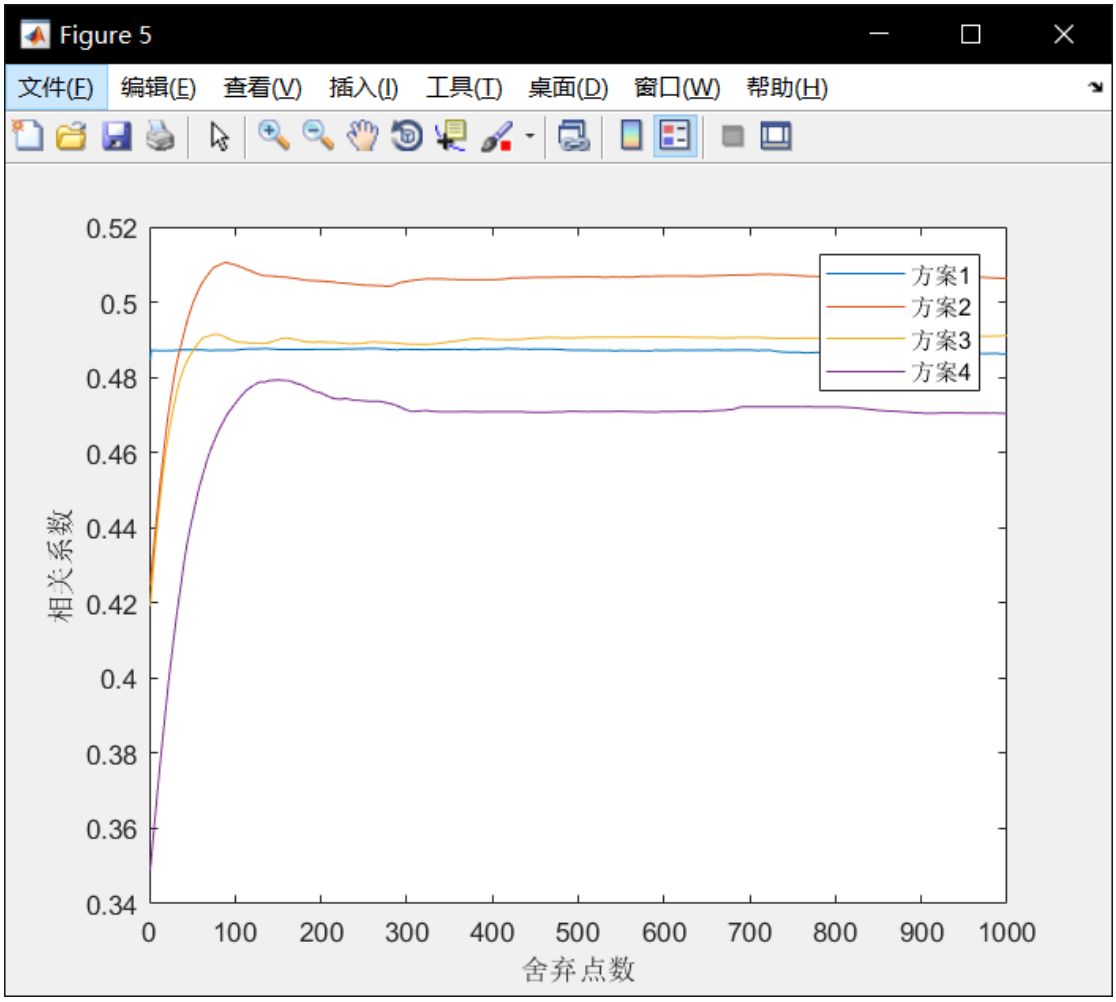
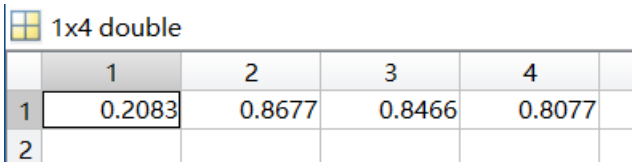


图 4-1-1：MH 算法四种推荐方案的相关系数关于舍弃点数曲线



1x4 double

	1	2	3	4
1	0.2083	0.8677	0.8466	0.8077
2				

图 4-1-2：MH 算法四种推荐方案接受率

可以看到，方案 2-4 相对方案 1 显著提高了接受率，同时方案 2、3 具有较高的收敛速度。但由于方案 3、4 并没有给出精确的转移概率，故其采样的相关系数与 0.5 相比有一定误差。

2）Gibss 算法

采样值在两个维度上来回跳转更新，每个维度上的更新相当于一次无 Q 矩阵的均匀推荐。实验仿真结果如下：

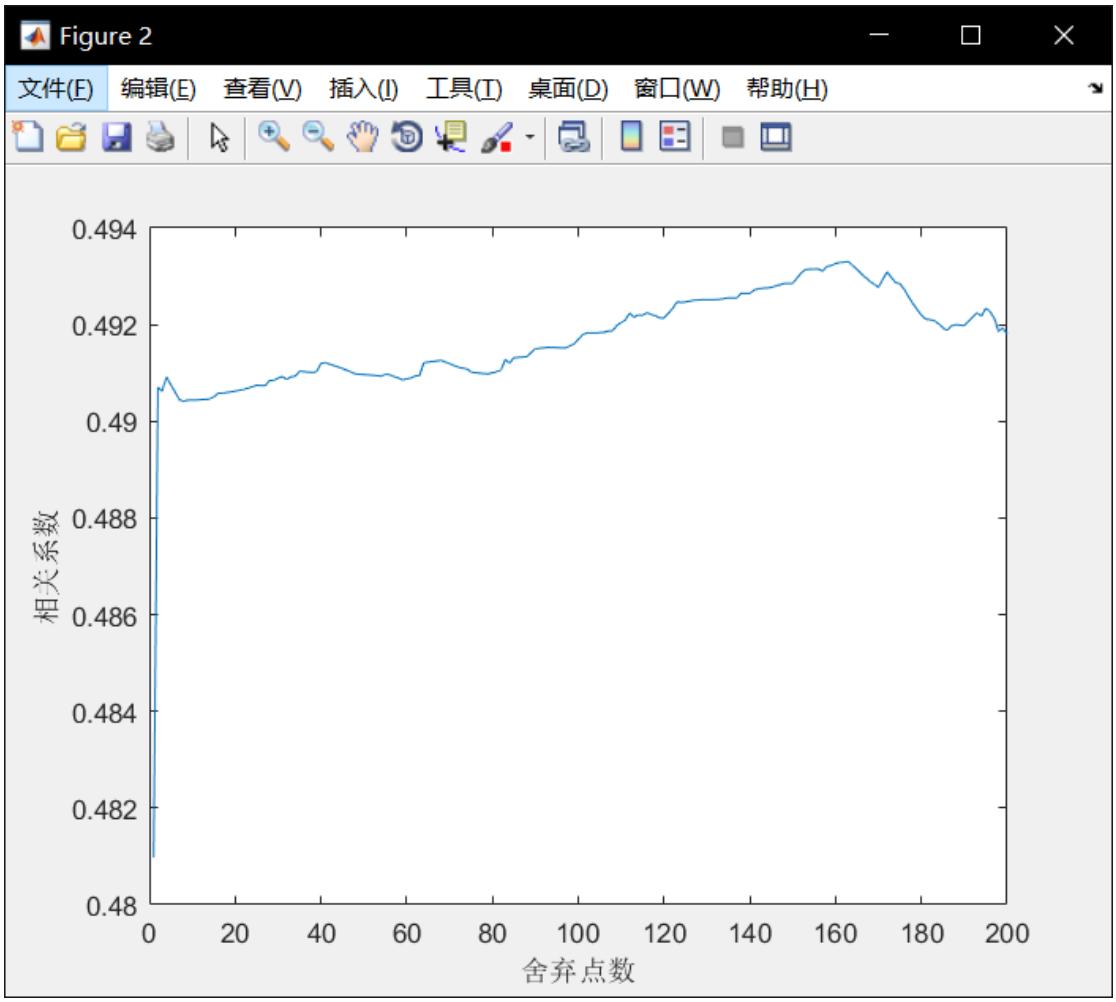


图 4-1-3：Gibbs 算法的相关系数关于舍弃点数曲线

	1	2
1	0.4110	
2		

图 4-1-4：Gibbs 算法的接受率

可以看到，相对于 MH 算法，Gibbs 采样能以非常高的速度收敛（几乎在一个采样点内），且获得的相关系数更为准确，但其接受率较低。

（2）Potts 模型的采样及归一化常数估计

对于 Potts 模型

$$p(x) = \frac{1}{Z(T)} exp[-\frac{u(x)}{T}],$$

$$u(x) = \sum_{\substack{i \leftrightarrow j \\ i,j=1,\ldots,K}} 1(x_i = x_j),$$

$$Z(T) = \sum_x exp[-\frac{u(x)}{T}]$$

显然，直接暴力枚举体系的每个状态 x，并计算其未归一化概率再求和，是不现实的（本例 q=10,K=20 的情形下，x 的可能状态有 10^400 种）。

由 Z 的指数形式，我们想到可以两侧取对数

$$lnZ = ln\left(\sum_x exp[-\beta u(x)]\right)$$

其中$\beta = 1/T$

由于求和的存在，ln 无法化简 e 指数，于是想到两侧对β求导

$$-\frac{d}{d\beta} lnZ = -\frac{1}{\sum_x exp[-\beta u(x)]} \sum_x -u(x) exp[-\beta u(x)]$$

$$= E[u(x)]$$

巧妙的用 u(x)期望的估计代替了 Z 的估计，两侧积分得到

$$lnZ = -\int_a^b E[u(x)]d\beta + C$$

在$\beta = 0$时可直接求出$Z(T) = 10^{400} = C$，故取$a = 0$，b 分别取 1/T = 1.4,1.4065,1.413,1.4195,1.426。

以下分别用 MH 算法与 SW 算法估计给定β下的$E[u(x)]$，积分得到 lnZ 的估计值。（实验仿真结果均以 J = 1.4 为例，设定积分间隔 delta = 0.0065，每个 β 下迭代次数 iterations = 100）

1）MH 算法

我们需要对输入的 xn(20*20 矩阵)，推荐可能跳转的状态 xp

方案 1：随机推荐，即无视上一时刻的 xn，x 中每个单元均独立均匀选择 q 个状态中的一个。由于对称选择，故转移概率直接返回 1。

方案 2：小幅变化推荐，对 xn 中的每个单元，1/4 的概率跳转到 q-1，1/4 的概率跳转到 q+1（此处 q 取值[1,10]中整数，0 认为是 10，11 认为是 1），1/2 概率不变。由于对称选择，故转移概率直接返回 1。

方案 3：平滑推荐，对 xn 中的每个单元，取周围 4 个单元的平均值，加上一个[-2,2]均匀的随机整数。由于推荐不对称，故用 x 的方差 var(x)来估计跳转概率，对二参输入 x,y，返回 var(x)/var(y)。

实验仿真结果如下：

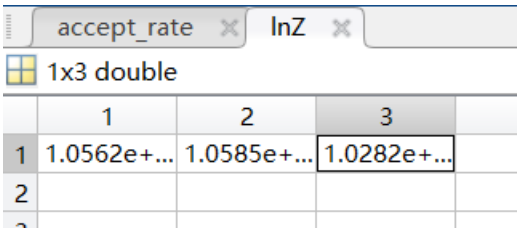


图 4-2-1：MH 算法三种方案的 lnZ 估计值（未显示出的部分为 e+3）

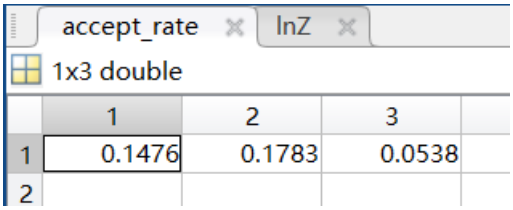


图 4-2-2：MH 算法三种方案的接受率

可以看到，MH 算法对 lnZ 的估计值在 1050 左右，三种方案的接受率都比较低，方案 3 由于没有准确描述转移概率导致结果相对其余两者有偏差。

在此高维模型下，可以发现 MH 算法计算得到的跳转概率极小，因此效率不高，且在有限的样本数下，收敛程度低，导致对 lnZ 的估计非常不精确。

2）SW 算法

Swendsen－Wang 算法的核心思想是：将所有单元的状态取值集合 x 看作一个随机变量，主动引入另一维随机变量“bond”，对 x 与 bond 的联合分布做 Gibbs 采样，再求 x 的边缘（即向 x 投影，去掉 bond）得到 x 的样本序列。算法的具体描述如下：

随机初始状态 x(0)，对所有相邻（有边相连）的单元 i,j，初始化定义在此边集上的 bond 变量。

for n = 0:N-1

定义概率 p = 1-exp(-β)

首先更新 bond，对所有相连的 i,j，若 si != sj，则设置 bond(i,j)为禁止的（在实现中简单的赋值为-1）；若 si == sj，则以 p 的概率设置 bond(i,j) = 1，1-p 的概率 bond(i,j) = 0。

再更新单元状态，对所有以 bond = 1 相连的状态，将其划分为一个聚簇，对图中所有聚簇，每个聚簇以均匀的概率选取 q 个状态中的一个（每个聚簇内的所有单元状态相同）。

依以上步骤得到的状态即为 x(n+1)。

end for

实验仿真中，将图以 bond = 1 为判据划分成聚簇的算法采用深度优先搜索实现，时间复杂度为 O(K^2)，空间复杂度为 O(K^2)。仿真结果如下：

图 4-2-2：SW 算法 J=1.4 的 lnZ 估计值

结果相对于 MH 算法有较大的提升，主要由于是在给定的 100 次迭代内更快的收敛，预热产生的样本个数相对较少，逼近 lnZ 的真实值。

为了得到精确的估计，我们细化积分间隔 delta，令 delta = 0.001 进行仿真，得到

图 4-2-3：细化积分间隔后 SW 算法 J=1.4 的 lnZ 估计值

图 4-2-4：细化积分间隔后 SW 算法 J=1.4065 的 lnZ 估计值

图 4-2-5：细化积分间隔后 SW 算法 J=1.413 的 lnZ 估计值

图 4-2-6：细化积分间隔后 SW 算法 J=1.4195 的 lnZ 估计值

图 4-2-7：细化积分间隔后 SW 算法 J=1.426 的 lnZ 估计值

可以发现，温度倒数 J 每提高 0.0065，lnZ 的值增加 2 左右。这与 J 与 lnZ 在表达式中是近似的线性关系相印证。

使用 SW 算法，绘制 J = 1.4,1.4065,1.413,1.4195,1.426 的能量直方图如下：

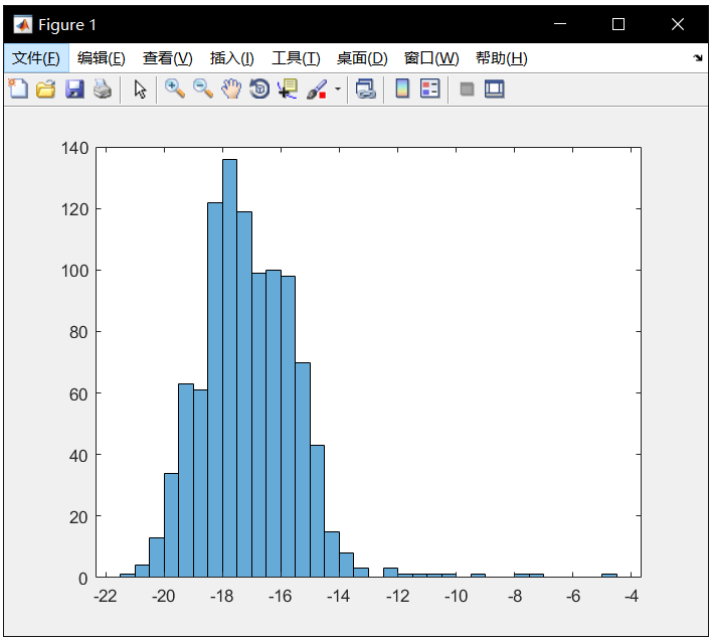


图 4-2-8：J = 1.4 时能量直方图

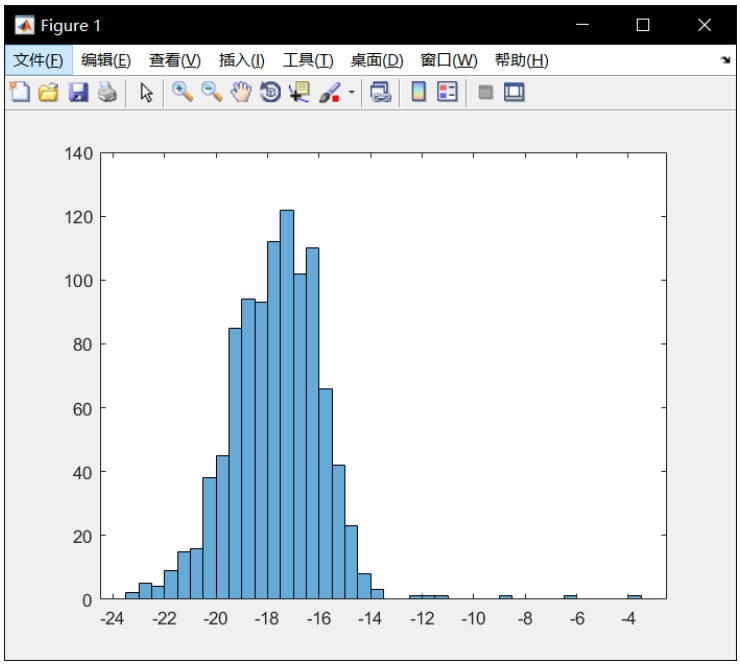


图 4-2-9：J = 1.4065 时能量直方图

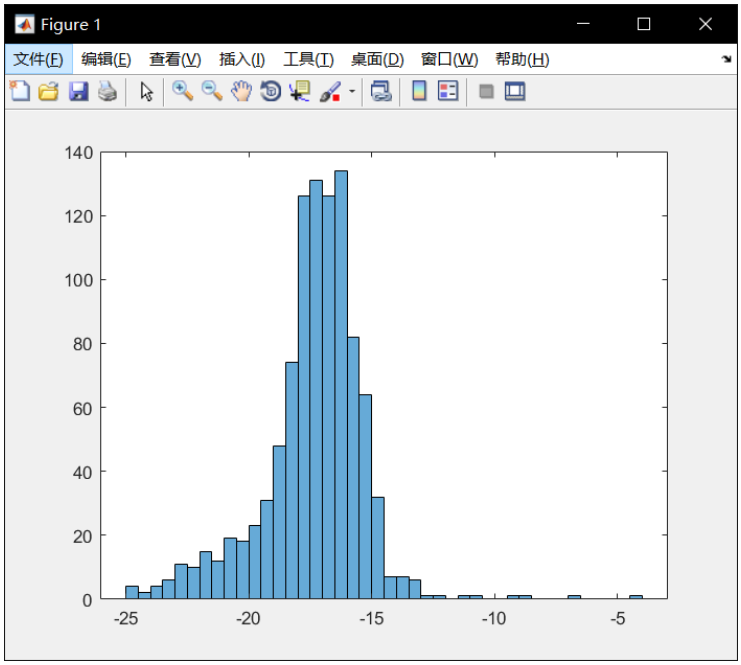


图 4-2-10：J = 1.413 时能量直方图

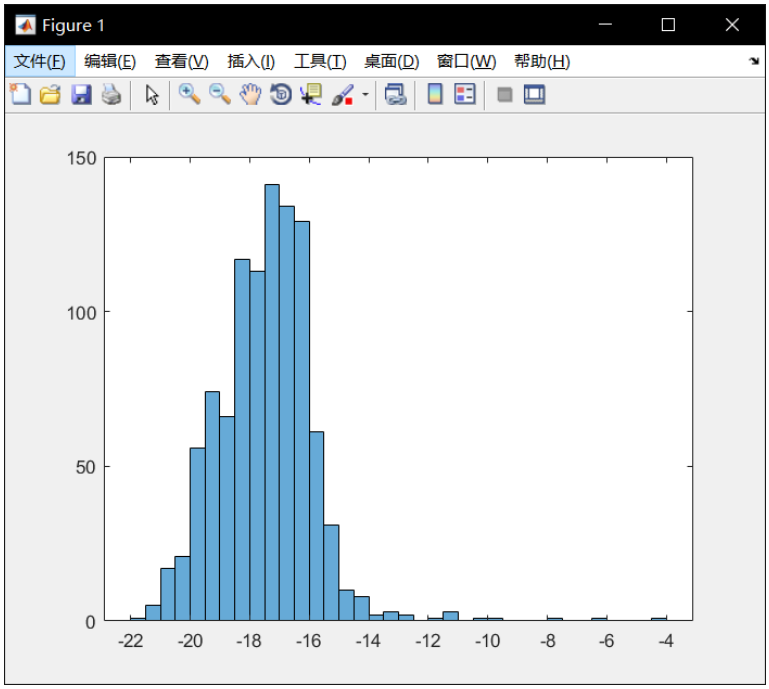


图 4-2-11：J = 1.4195 时能量直方图

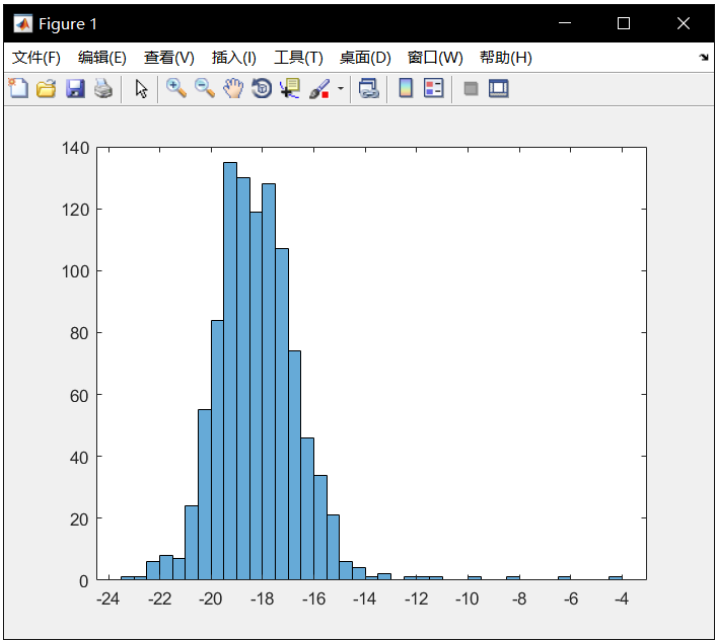


图 4-2-12：J = 1.426 时能量直方图

对上述仿真结果，抽取最终分布中出现频率最大的状态 x 作为典型样本，以 20*20 图片格式显示如下（以不同的灰度值表示 q 个不同的状态）：

图 4-2-13：J = 1.4 时典型样本

图 4-2-14：J = 1.4065 时典型样本

图 4-2-15：J = 1.413 时典型样本

图 4-2-16：J = 1.4195 时典型样本

图 4-2-17：J = 1.426 时典型样本

5.结论

传统的 MH 算法，适用于给定的分布维度较低，转移概率较大的情形。此时设定推荐分布 T 为均匀分布可以获得较精确的结果，但接受率较低；将 T 优化为均匀游走推荐等方法可有效提高接受率，但没有利用给定分布的先验信息；将 T 设计为不对称的，倾向于分布π的推荐方式可以获得较高的收敛速度，但若不能严格计算/实现 T 的条件概率将会导致采样的统计量不足够精确。

Gibbs 算法在给定分布的条件分布易求的情况下，相对于 MH 算法能大幅提高跳转效率及收敛速度，得到更精确的统计结果。

SW 算法极富创意的对难以处理的 Potts 模型增添了一维变量，从而应用 Gibbs 采样的方法快速迭代，相对于用传统的 MH 处理 Potts 模型更为高效。

6.致谢

感谢高帆同学与我讨论 Potts 模型的细节，以及对 SW 算法的理解。

感谢余书涵同学推荐编辑论文的工具，以及给予我极大精神上的关怀。

7.参考文献

[1]林元烈,应用随机过程.清华大学出版社, 2002.

[2]MacKay and C. Davidj.,Information Theory, Inference, and Learning Algorithms. Cam-bridge University Press, 2003.

[3]J. S. Liu,Monte Carlo strategies in scientific computing. Springer Science & BusinessMedia, 2008.

[4]I. Murray, D. J. C. Mackay, Z. Ghahramani, and J. Skilling, “Nested sampling for pottsmodels,”Advances in Neural Information Processing Systems, pp. 947–954, 2006.

[5]Z. Tan, “Optimally adjusted mixture sampling and locally weighted histogram analysis,”Journal of Computational and Graphical Statistics, vol. 26, no. 1, pp. 54–65, 2017.

[6]朱新玲,中南财经政法大学信息学院 武汉科技大学管理学院,”马尔科夫链蒙特卡罗方法研究综述”,2009