Lecture Notes for STAT546: Computational Statistics

—Lecture 4: Monte Carlo

Faming Liang

Purdue University

August 28, 2024

Difficulties with MH algorithm

The MH algorithm suffers from two difficulties:

- The local-trap problem;
- Inability to sample from distributions with intractable integrals.

The local-trap problem refers to that in simulations of a complex system whose energy landscape is rugged, the sampler gets trapped in a local energy minimum indefinitely, rendering the simulation ineffective.

Let $f(x) \propto c(x)\psi(x)$ denote a distribution of interest, where $c(x)$ denotes an intractable integral. Clearly, the MH algorithm cannot be applied to sample from $f(x)$, as the acceptance probability would involve an unknown ratio $c(x')/c(x)$, where x' denotes the proposed value. This difficulty naturally arises in Bayesian inference for many statistical models, such as spatial statistical models, random effects models, and exponential random graph models. To alleviate or overcome these two difficulties, various advanced MCMC methods have been proposed in the literature, including the auxiliary variable-based methods, the population-based methods, importance weight-based methods, and stochastic approximation-based methods, which will be described in the followed lectures.

Auxiliary variable MCMC Methods

Consider the problem of sampling from a multivariate distribution with density function $f(x)$. It is known that *Rao-Blackwellization* (Bickel and Doksum, 2000) is the first principle of Monte Carlo simulation: In order to achieve better convergence of the simulation, one should try to integrate out as many components of x as possible. However, sometimes, one can include one or more additional variables in simulations to accelerate or facilitate the simulations. This often occurs in two scenarios:

- *The target distribution $f(x)$ is multimodal*: An auxiliary variable, such as temperature or some unobservable measurement, is included in simulations to help the system to escape from local-traps.
- *The target distribution $f(x)$ includes an intractable normalizing constant*: An auxiliary realization of X is included in simulations to have the normalizing constants canceled in simulations.

The MH algorithm involves two basic components, the target distribution and the proposal distribution. Accordingly, the auxiliary variable methods can also be performed in two ways, namely augmenting auxiliary variables to the target distribution and augmenting auxiliary variables to the proposal distribution. We refer to the former as the method of target distribution augmentation and to the latter as the method of proposal distribution augmentation.

The method of target distribution augmentation:

- Specify an auxiliary variable u and the conditional distribution $f(u|x)$ to form the joint distribution $f(x,u) = f(u|x)f(x)$.
- Update (x,u) using the MH algorithm or the Gibbs sampler.

The samples of $f(x)$ can then be obtained from the realizations, $(x_1,u_1), \ldots, (x_N,u_N)$, of (X,U) through marginalization or projection.

The method of proposal distribution augmentation:

- Specify a proposal distribution $T(x', u|x)$ and its reversible version $T(x, u|x')$ such that $\int T(x', u|x)du = T(x'|x)$ and $\int T(x, u|x')du = T(x|x')$.
- Generate a candidate sample x' from the proposal $T(x', u|x)$, and accept it with probability $\min\{1, r(x, x', u)\}$, where

$$r(x, x', u) = \frac{f(x')}{f(x)} \frac{T(x, u|x')}{T(x', u|x)}.$$

Repeat this step to generate realizations $x_1, \ldots, x_N$, which will be approximately distributed as $f(x)$ when N becomes large.

The validity of this method can be shown as follows. Let

$$K(x'|x) = \int_{\mathcal{U}} s(x, x', u)g(u)du + I(x = x')[1 - \int_{\mathcal{X}} \int_{\mathcal{U}} s(x, x^*, u)g(u)dudx^* \quad (1)$$

denote the integrated transitional kernel from x to x', where $s(x, x', u) = T(x', u|x)\ r(x, x', u)$, and $I(\cdot)$ is the indicator function. Then

$$f(x) \int_{\mathcal{U}} s(x, x', u)g(u)du = \int_{\mathcal{U}} \min\{f(x')T(x, u|x'), f(x)T(x', u|x)\}g(u) \quad (2)$$

which is symmetric about x and x'. This implies $f(x)K(x'|x) = f(x')\ K(x|x')$.

Simulated Annealing

Suppose that one aims to find the global minimum of an objective function $H(x)$, which is also called the energy function in the standard terms of simulated annealing. By augmenting to the system an auxiliary variable, the so-called temperature T, minimizing $H(x)$ is equivalent to sampling from the Boltzmann distribution $f(x, T) \propto \exp(-H(x)/T)$ at a very small value (closing to 0) of T. In this sense, we say that *sampling is more basic than optimization.*

In order to sample successfully from $f(x, T)$ at a very small value of T, Kirkpatrick and co-authors suggested to simulate from a sequence of Boltzmann distributions, $f(x, T_1), \ldots, f(x, T_m)$, in a sequential manner, where the temperatures form a decreasing ladder $T_1 > T_2 > \cdots > T_m$ with $T_m \approx 0$ and T_1 being reasonably large such that most uphill MH moves at that level can be accepted. The simulation at high temperature levels aims to provide a good initial sample, hopefully a point in the attraction basin of the global minimum of $H(x)$, for the simulation at low temperature levels.

Simulated Annealing Algorithm

- Initialize the simulation at temperature T_1 and an arbitrary sample x_0.
- At each temperature T_i, simulate of the distribution $f(x, T_i)$ for N_i iterations using a MCMC sampler. Pass the final sample to the next lower temperature level as the initial sample.

Simulated Tempering

Suppose that it is of interest to sample from the distribution $f(x) \propto \exp(-H(x))$, $x \in \mathcal{X}$. As in simulated annealing, simulated tempering (Marinari and Parisi, 1992; Geyer and Thompson, 1995) augments the target distribution to $f(x, T) \propto \exp(-H(x)/T)$ by including an auxiliary variable T, called temperature, which takes values from a finite set pre-specified by the user.

Simulated Tempering

Simulated Tempering

- Draw a random number $U \sim \text{Uniform}[0,1]$, and determine the value of j according to the proposal transition matrix (q_{ij}).
- If $j = i_t$, let $i_{t+1} = i_t$ and let x_{t+1} be drawn from a MH kernel $K_{i_t}(x, y)$ which admits $f(x, T_{i_t})$ as the invariant distribution.
- If $j \neq i_t$, let $x_{t+1} = x_t$ and accept the proposal with probability

$$\min\left\{1, \frac{\widehat{Z}_j}{\widehat{Z}_{i_t}} \exp\left\{-H(x)\left(\frac{1}{T_j} - \frac{1}{T_{i_t}}\right)\right\} \frac{q_{j,i_t}}{q_{i_t,j}}\right\},$$

where $\widehat{Z}_i$ denotes an estimate of Z_i. If it is accepted, set $i_{t+1} = j$. Otherwise, set $i_{t+1} = i_t$.

Issues on Simulated Tempering

- *Choice of the temperature ladder*: The highest temperature T_1 should be set such that most of the uphill moves can be accepted at that level. The intermediate temperatures can be set in a sequential manner: Start with T_1, and sequentially set the next lower temperature such that

$$\mathrm{Var}_i(H(x))\delta^2 = O(1), \tag{3}$$

where $\delta = 1/T_{i+1} - 1/T_i$, and $\mathrm{Var}_i(\cdot)$ denotes the variance of $H(x)$ taken with respect to $f(x, T_i)$. This condition is equivalent to requiring that the distributions $f(x, T_i)$ and $f(x, T_{i+1})$ have considerable overlap. In practice, $\mathrm{Var}_i(H(x))$ can be estimated roughly through a preliminary run of the sampler at level T_i.

- *Estimation of Z_i's*: This is the key to the efficiency of simulated tempering. If the Z_i's are well estimated, simulated tempering will perform like a "symmetric random walk" along the temperature ladder (if ignoring the x-updating steps). Otherwise, it may get stuck at a certain temperature level, rendering a failure of the simulation. In practice, the Z_i's can be estimated using the stochastic approximation Monte Carlo method (Liang *et al.*, 2007; Liang, 2005). Alternatively, the Z_i's can be estimated using the reverse logistic regression method as suggested by Geyer and Thompson (1995).