Advanced computational methods-Lecture 11

Other MCMC algorithms.

1 Gibbs sampler

This was introduced by brothers of Stuart and Geman in 1984. Gibbs sampler is commonly used for Bayesian inference. Compared with the traditional EM (expectation-maximization) algorithm, it is a stochastic one.

This is applicable when the joint distribution is not known or hard to sample but the conditional distribution (not marginal) is easier to sample from.

Suppose we want to sample a multivariate distribution $\rho(Z) = \rho(Z_1, Z_2, \dots, Z_M)$, we know all the marginal distributions. Suppose we have already sampled Z^n, we do the following to get Z^{n+1}:

- $Z_1^{n+1} \sim p(z_1 | Z_2^n, Z_3^n, \dots, Z_M^n)$.
- $Z_2^{n+1} \sim p(z_2 | Z_1^{n+1}, Z_3^n, \dots, Z_M^n)$.
- $Z_3^{n+1} \sim p(z_3 | Z_1^{n+1}, Z_2^{n+1}, \dots, Z_M^n)$.
- $\dots$
- $Z_M^{n+1} \sim p(z_M | Z_1^{n+1}, \dots, Z_{M-1}^{n+1})$.

We first of all claim that the joint distribution is invariant. Consider the jump from (Z_j^m, Z_i^n) to (Z_j^m, Z_i^{n+1}). The transition probability is given by

$$q(x, y | Z_j^m) = p(y | Z_j^m).$$

The desired distribution is the joint distribution $\rho(Z)$. The detailed balance hods for this ρ

$$\rho(Z_j^m, x) q(x, y | Z_j^m) = \rho(Z_j^m, y) q(y, x | Z_j^m).$$

This is found easily by the definition of the conditional distribution $p(y | Z_j^m)$.

With the detailed balance, we find that this basic version can be viewed as a special case of Metropolis-Hastings. In fact, the function $a(x, y | Z_j^m)$ is always 1 and the samples are always accepted. This can therefore be regarded as an MH algorithm. However, its extensions can be more general.

When n is large enough, the distribution of Z^n is close to the joint distribution. In practice, one may disregard some initial samples. Also, we may only want every n_1th sample. The first is because the distribution will be close to the stationary distribution only after some steps. The second is considered because the adjacent data are usually not independent.

1.1 Some extensions

The obvious extension is to consider a stochastic version of the Gibbs sampler:

- We choose an index i randomly from $\{1, 2, \dots, M\}$
- We sample $\xi \sim p(z|Z_{-i})$, where Z_{-i} means the vector by deleting Z_i.
- Set $Z_i = \xi$. In other words, the new random variable is

$$Z = (Z_1, \dots, Z_{i-1}, \xi, Z_{i+1}, \dots, Z_M).$$

Other variants include blocked Gibbs sampling, in which we may sample several Z_i's by conditioning on others.

Another variant is the collapsed Gibbs sampling. Imagine that we want to sample from $p(x, y|z)$. In the Gibbs sampler above, we can do $x^{n+1} \sim p(x|y^n, z)$ and $y^{n+1} \sim p(y|x^{n+1}, z)$. In the collapsed version, we do the following

$$y^{n+1} \sim p(y|z), \quad x^{n+1} \sim p(x|y^{n+1}, z).$$

Here, $p(y|z) = \int p(x, y|z)\,dx$. Note that the one with one variable integrated must be sample first. Otherwise, the distribution we are sampling from is not correct.

2 Hamiltonian MCMC

Clearly, in the running average, the two adjacent data are not independent. If we run Metropolis-Hastings, we need the time to be longer than the correlation time (the correlation between X^n and X^{n+m} decays exponentially with m. Hence, we need m bigger than some time that ensures the correlation to be smaller than a given value.)

The Hamiltonian MCMC is designed by Duane, Kennedy, et al in 1987 to reduce the correlation using a Hamiltonian evolution. It also allows a higher acceptance rate. Interestingly, the original name by the authors is 'Hybrid Monte Carlo', which is 'HMC' in short. Nowadays, it is better known as 'Hamiltonian Monte Carlo', which is also 'HMC' in short.

For more details, see review articles "MCMC using Hamiltonian dynamics", "A conceptual introduction to Hamiltonian Monte Carlo".

2.1 Motivation and idea

In statistical mechanics, it is well-known that a system with Hamiltonian $H(p, q)$ that interacts with the surrounding heat bath could result in the the canonical distribution:

$$\rho(p, q) \propto \exp(-\beta H(p, q)),$$

in the thermoequilibrium. Often, the Hamiltonian is given by

$$H(p, q) = \frac{p^2}{2m} + U(q),$$

where the first term is the kinetic energy while the second is the potential energy. The marginal distribution on q is therefore

$$\exp(-\beta U(q)).$$

Hence, if we choose

$$U(q) = -\frac{1}{\beta} \log \rho(q),$$

we can run the dynamics of the system to obtain the desired distribution $\rho(q)$.

This motivates the HMC and Langevin MC. In HMC, the system is closed and evolves according to Hamilton ODEs while in Langevin dynamics, we have whilte noise and friction that describe the interaction with heat bath.

2.2 The Hamiltonian system and its discretization

In the Hamilton mechnics, $H(p, q)$ is the energy functional. The dynamics is given by

$$\begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i}, \\ \dot{p}_i &= -\frac{\partial H}{\partial q_i}. \end{aligned}$$

It is convenient to introduce the matrix

$$J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}.$$

The dynamics is written as

$$\frac{dz}{dt} = J\nabla_z H,$$

where $z = [q, p]^T$.

Often, the Hamiltonian is given by

$$H(p, q) = U(q) + K(p),$$

and

$$K(p) = \frac{1}{2}p^T M^{-1} p,$$

and $M = diag(m_1, \ldots, m_d)$. There are several properties

- Reversible. According to the dynamics, if the state goes from (q_0, p_0) to (q_1, p_1), we negate p and then have $(q_1, -p_1)$. From this state, after the same time, it will arrive at $(q_0, -p_0)$.
- The Hamiltonian is conserved. $\frac{d}{dt}H(p(t), q(t)) = 0$.
- Volume-preserving. The dynamics is $\frac{dz}{dt} = V(z) := J\nabla_z H$. We can verify easily that
$$\nabla_z \cdot V(z) = 0.$$
- Symplectic. Let the flow map be T_t and the corresponding Jacobi matrix be B_t. Then,
$$B_t^T J^{-1} B_t = J^{-1}.$$

The symplecticity is in fact the two-form (signed areas).

To discretize, the naive idea is to use forward Euler,

$$\begin{aligned} p_{n+1} &= p_n - \tau \nabla_q U(q), \\ q_{n+1} &= q_n + \tau \frac{p_n}{m} \end{aligned}$$

This scheme, however, behaves poorly in practice. The reason is that the volume in phase space is not preserved, and the Hamiltonian is not preserved.

A simple modified scheme

$$\begin{aligned} p_{n+1} &= p_n - \tau \nabla_q U(q_n), \\ q_{n+1} &= q_n + \tau \frac{p_{n+1}}{m_i} \end{aligned}$$

behaves much better. The reason is that it preserves the volume. In fact, this is related to a large class of methods, called **symplectic methods** for Hamiltonian discretizations. The symplectic method preserves the two forms. *The energy is not conserved accurately, but the energy can be preserved for a long time.*

The following leapfrog method, which shares some flavor of Strang splitting, is as follows

$$\begin{aligned}
p_{n+1/2} &= p_n - (\tau/2)\nabla_q U(q_n),\\
q_{n+1} &= q_n + \tau\frac{p_{n+1/2}}{m_i},\\
p_{n+1} &= p_{n+1/2} - (\tau/2)\nabla_q U(q_{n+1}).
\end{aligned}$$

This is a symplectic method.

2.3 Hamiltonian MC

The motivation comes from the canonical distribution in statistical mechanics. In the thermo-equilibrium, the probability that a configuration appears is

$$p(q,p) = \frac{1}{Z}\exp(-H(q,p)).$$

Of course, the canonical distribution is not for a Hamiltonian system, because for a closed system, the energy is conserved. This is for a system in a heat bath. The system has its own Hamitonian, but it also has interaction with the enviroment so that its energy can be changed. Anyway, this gives us the motivation to use the Hamilton system for sampling.

Suppose we want to sample from $\rho(q)$. Then, we construct a Hamiltonian

$$H(q,p) = -\log\rho(q) + K(p) =: U(q) + K(p).$$

Then, the marginal distribution of the canonical distribution is exactly what we want. Hence, we have the following HMC

- Sample $p_n \sim p(p|q_n) \propto \exp(-\frac{|p|^2}{2})$.
- Here, we apply Metropolis update. Run the Hamiltonian dynamics from state (q_n, p_n) to (q^*_{n+1}, p^*) using the leapfrog scheme, with step size τ for L steps. (It other words, we have 'time' consumed to be $L\tau$). Then, we negate p so that the proposed state is (q^*_{n+1}, p^*_{n+1}). With probability

$$\min(1, \exp(-H(q^*_{n+1}, p^*_{n+1}) + H(q_n, p_n)))$$

we accept the state and then $q_{n+1} = q^*_{n+1}$. Otherwise, $q_{n+1} = q_n$.

Remark 1. - *In the Metropolis update, since the dynamics is reversible, so $q(x, y) = q(y, x)$ and hence, the function a is determined totally by the desired canonical distribution (Gibbs distribution).*

- *Since the Hamiltonian is symmetric about the momentum, we do not really need to negate the momentum p^*.*

- *The resampling of p at the beginning is very important. Otherwise, the distribution of q will not range over the desired distribution* $\exp(-U(q))$

To verify that the method works, one should check that the canonical distribution is invariant under the process. This is straightforward to see.

A second thing is to gurantee the ergodicity. This is not very clear for HMC. In fact, in some cases, the chain is not aperiodic so that it is not ergodic. However, this is very rare to happen. However, to make sure the ergodicity, one may make the running time random. We will not go into details for these issues.

3 Langevin MCMC

3.1 Motivation and the algorithm

In Hamiltonian MCMC, there is no noise. If we consider the dynamical system with white noise, we can have the following system of equation

$$\begin{aligned} &dx = v\,dt, \\ &mdv = -\gamma v\,dt - \nabla_x U(x)\,dt + \sqrt{2\gamma/\beta} dW, \end{aligned}$$

where $v = p/m$. The distribution can be proven to converge to the Gibbs measure using hypocoercivity

$$\propto \exp(-\beta(U(x) + \frac{m|v|^2}{2}))$$

The marginal distribution is what we want. Note that this measure is invariant under the SDE. To see this, we can define $X = [x, v]^T$, can write it as

$$dX = b(X)\,dt + [0, \sqrt{2\gamma/\beta}]^T dW.$$

Using the Fokker-Planck equation for X, we obtain the Fokker-Planck for the system.

We can also consider the overdamped regime:

$$dx = -\frac{1}{\gamma}\nabla_x U\,dt + \sqrt{\frac{2}{\gamma\beta}}dW$$

The invaraint measure of this SDE is

$$\propto \exp(-\beta U(x))$$

Hence, we can use both to do the MCMC.

For the exponential convergence of the overdampled case for sampling and the Metropolis-adjusted discretization, you may read ”Exponential convergence of Langevin distributions and their discrete approximations” and “Sampling from a strongly log-concave distribution with the unadjusted Langevin algorithm”. For the underdamped one (with momentum), the reference is “Underdamped Langevin MCMC: a nonasymptotic analysis”.

The overdamped Langevin MCMC

We take the temperature parameter to be $\beta = 1$, and consider the following SDE

$$dX = -\nabla U\,dt + \sqrt{2}\,dB.$$

Recall that

$$U = -\log\rho.$$

We hope U to be strongly convex so that we need ρ to be log concave.

A version of simple Langevin MC is therefore given by the Euler-Maruyama scheme

$$X^{n+1} = X^n - \gamma_{n+1}\nabla U(X^n) + \sqrt{2\gamma_{n+1}}Z_{n+1}$$

where $Z_{n+1} \sim \mathcal{N}(0,1)$.

3.2 A second order scheme for stationary distribution

In the work “Rational construction of stochastic numerical methods for molecular sampling”, Leimkuhler and Matthews obtained a modification of the Euler-Maruyama scheme

$$X^{n+1} = X^n - k\nabla V(x) + \frac{\sqrt{2\beta}}{2}(\Delta W_n + \Delta W_{n+1})$$

Note that $\{X^n\}$ is not a Markov chain. This scheme solves the SDE still with first order but for sampling the stationary distribution π using averages, it is second order. Hence, it is good for MCMC sampling.