Hamiltonian Monte Carlo

Yuzhe Ou

Aug 2020

1 Background

In Bayesian inference, for most probabilistic models of practical interest, exact inference is intractable (e.g. due to difficulty of get analytical form for integral). Thus approximation methods such as variational Bayes, expectation propagation. Beside these sampling methods have been studied for decades as a popular tool to help with approximation inference.

We focus on the general problem in Bayesian inference which involves a probability distribution $p(z)$, in a lot of cases $p(z) = q(\theta|D)$ where $z = \theta$ and posterior $q(\theta|D)$ is of interest. All well-posed Bayesian computations reduce to expectations, i.e. the expectation of some function $f(z)$ with respect to $p(z)$. And consequently, it is an integral in continuous case:

$$E[f] = \int f(z)p(z)dz \tag{1}$$

The direct computation is usually difficult as $f(z)p(z)$ has complex form in practical.

The idea of using sampling method is that we can approximation this using a finite sum

$$\hat{f} = \frac{1}{L}\sum_{l=1}^{L} f(z^{(l)}) \tag{2}$$

where samples $\{z^{(l)}\}_{l=1,2,...,L}$ are drawn from distribution $p(z)$.

2 Monte Carlo Markov Chain

The idea of Monte Carlo Markov Chain (MCMC) is to construct a first order Markov Chain of which the stationary/invariant distribution is our target distribution $p(z)$. In other word, such Markov Chain, $z^{(1)}, z^{(2)}, ..., z^{(t)}, z^(t+1), ...$, which is a series of random variables provide us samples $z^{(t+1),z^{(t+2),...}}$ from $p(z)$ if the Markov Chain reaches stationary state at time t (i.e. $p(z^{(m)}) = p(z),\ m > t$).

Let's denote the transition probability as $T_m(z^{(m)}, z^{(m+1)}) \equiv p(z^{(m+1)}|z^{(m)})$. A sufficient condition for ensuring the required distibution $p(z)$ being invariant with respect to the constructed Markov Chain is to satisfy the property of detailed balance, which is defined as

$$p(z)T(z, z^{'}) = p(z^{'})T(z^{'}, z) \tag{3}$$

In fact in order to get samples from $p(z)$, we only need to know $\tilde{p}(z)$ which is proportional to $p(z)$ through an normalization constant Z_p in the form: $p(z) = \frac{1}{Z_p}\tilde{p}(z)$. We assume that $\tilde{p}(z)$ is known and can be readily evaluated for all the discussions behind.

The general framework of MCMC, Metropolis-Hasting algorithm (Hastings, 1970)(abbreviated as MH), achieves the goal by repeating a two-step process: Let $z^{(\tau)}$ be the current state of Markov Chain,

1. Generate $z^* \sim Q(z|z^{(\tau)})$
2. Set $z^{(\tau+1)} = z$ with probability $A(z^*, z^{(\tau)}) = min(1, \frac{\tilde{p}(z^*)Q(z|z^{(\tau)})}{\tilde{p}(z^\tau)Q(z^{(\tau)}|z^*)})$,
otherwise set $z^{(\tau+1)} = z^{(\tau)}$.

Where $Q(\cdot \mid \cdot)$ is a probability density function (p.d.f) called proposal distribution and $A(\cdot\ ,\ \cdot)$ is acceptance rate/probability that has value between 0 and 1. Thus we get an first order Markov Chain with transition probability $T(z^{(\tau)}, z^{(\tau+1)}) = Q(z^{(\tau+1)}|z^{(\tau)})A(z^{(\tau+1)}, z^{(\tau)})$. Further it is easy to prove that this satisfies detailed balance, thus the samples given by this Markov Chain after invariant state are exactly from our target distribution $p(z)$.

3 Limitation of common MH algorithms

As we introduced in section 2, we need to defined some proposal distribution Q in order to define transition process on the Markov Chain. It seems that Q can be any valid probability distribution, but poor choice of Q will result in slow convergence or even no convergence, and this situation is even more serious when the dimension of z is high. For example a popular choice for Q is

$$Q(z|z^{(\tau)}) = N(z^{(\tau)}, \sigma^2)$$

which is a kind of random walk MH algorithm.

To give some more details, the proposal distribution $Q(z|z^{(\tau)})$ of random walk MH is associated the trial random variable $z^* = z^{(\tau)} + \delta$, where $\delta \sim R(\delta)$ is independent of $z^{(\tau)}$. Such $R(\cdot)$ usually involves a step-size parameter such as σ^2 for Gaussian, α for $Uniform(-\alpha, \alpha)$ which influences how likely we can get distant trials from current state.

However, large step size will usually result in high rejection rate (small A value) since $R(\delta)$ is closed to 0 if faraway from its mode, making the transition too slow and harm the performance. Unfortunately if small step size is too small, the convergence rate is slow, the Markov Chain exhibits random walk behavior so that Markov Chains explores the state space (the domain of state variable $z^{(\tau)}$) too slowly. Indeed, finding a suitable step size in high dimension is a big challenge.

4 Efficiently explore the state space - HMC

Hamiltonian Monte Carlo (HMC, also known as hybrid Monte Carlo), defines its transition in a very different way to address the limitations of common MH methods metioned in section 3, while it is still using the framework of MH.

The general question is that can we design a proposal or a transition process on Markov Chain such that it explores the state space efficiently while having high acceptance rate. In other words, can we make large change to the state variable but let the states/samples still be representative for target distribution $p(z)$. Intuitively this means we need some more information extracted from $\tilde{p}(z)$ that will guide us on the movement of state variable on the Markov Chain. Typically we wish that we get more samples from the region near the mode of $p(z)$ or region that keeps $f(z)p(z)$ significant, as this contributes much to finite sum term eq(1).

Instead of going around the state space with random, uninformed jumps, we can follow the direction assigned to each point for a small distance. HMC is constructed based on this, using a vector field that incorporates gradient information of log-density of state variable through momentum variable r that provide efficient exploration aligned with typical region of target distribution.

Let's come to the detail. We consider the following setting inherited from MH: we target to sample from distribution $p(z) = \frac{1}{Z_p}\tilde{p}(z)$, where Z_p is an unknown constant and $\tilde{P}(z)$ can be readily calculated. Beside this we assume that gradient of negative log-density, which is $\frac{\partial - log\tilde{p}(z)}{\partial z}$, can be evaluated.

Denote the state variable as z, and we add a momentum variable r, of the same dimension of z, and consider the follow joint distribution of z and r:

$$p(z, r) = p(z)p(r|z) \tag{4}$$

where we get to choose $P(r|z)$. In practice, we choose $P(r|z) = N(0, \Sigma)$, and a simple choice is $\Sigma = I$, we will use it for this note.

Taking log of both side, and we get log of joint density called Hamiltonian function:

$$H(z, r) = -log\ p(z) - log\ p(r|z) = E(z) + K(z, r) \tag{5}$$

where $K(z, r)$ is called the kinetic energy and $E(z)$ is called potential energy, using an analogy to physical systems. And for $E(z)$, although we do not know $p(z)$, $E(z) = -log\ p(z) = -log(\tilde{p}(z))$.

Notice that we make $K(z, r)$ independent of z and as our choice of Gaussian distribution,

$$K(z, r) = K(r) = \frac{1}{2}||r||^2 \tag{6}$$

Clearly $E(z)$ decides marginal density of z and $K(r)$ decides marginal density of r.

We consider evolution of state variable $z = (z_1, z_2, ...z_d)$ under continuous time, which we denote by τ, and we get the equations to express the Hamiltonian dynamics:

$$\frac{dz_i}{d\tau} = r_i = \frac{\partial H}{\partial r_i} \tag{7}$$

$$\frac{dr_i}{d\tau} = -\frac{\partial E(z)}{\partial z_i} = -\frac{\partial H}{\partial z_i} \tag{8}$$

where $i = 1, 2, ..., d$.

During the evolution, the value of Hamiltonian H is invariant, as

$$\frac{dH}{d\tau} = \sum_i \{\frac{\partial H}{\partial z_i}\frac{dz_i}{d\tau} + \frac{\partial H}{\partial r_i}\frac{dr_i}{d\tau}\} = \sum_i \{\frac{\partial H}{\partial z_i}\frac{\partial H}{\partial r_i} - \frac{\partial H}{\partial r_i}\frac{\partial H}{\partial z_i}\} = 0 \tag{9}$$

The second property of Hamiltonian dynamical systems is that they preserve volume in phase space - space of (z, r). Consider the flow field

$$V = (\frac{dz}{d\tau}, \frac{dr}{d\tau}) \tag{10}$$

the divergence of this field vanishes

$$div\ V = \sum_i \{\frac{\partial}{\partial z_i}\frac{dz_i}{d\tau} + \frac{\partial}{\partial r_i}\frac{dr_i}{d\tau}\} = \sum_i \{\frac{\partial}{\partial z_i}\frac{\partial H}{\partial r_i} - \frac{\partial}{\partial r_i}\frac{\partial H}{\partial z_i}\} = 0 \tag{11}$$

Now we take a look at the joint distribution over phase space:

$$p(z, r) = p(z)p(r|z) = \frac{1}{Z_H}exp(-H(z, r)) = \frac{1}{Z_H}\tilde{p}(z, r) \tag{12}$$

where Z_H is a constant. From eq(9) we know that the system leaves $p(z, r)$ invariant. The advantage of such system is now presented in front of us, as we calculate the evolution over phase space, our proposal of (z,r) through such deterministic way is always accepted since $\tilde{p}(z^{(\tau+1)}, r^{(\tau+1)}) = \tilde{p}(z^{(\tau)}, r^{(\tau)})$ makes acceptance rate A equals to 1. In the other hand, momentum variable r can take large value, this means we are able to make large change to the state variable while maintaining high acceptance rate.

However, we can not calculate the exact trajectory under continuous time, for practical application, we need to get discrete-time approximations. One scheme to achieve this is called leapfrog discretization:

$$\begin{aligned}
\hat{r_i}(\tau + \epsilon/2) &= \hat{r_i}(\tau) - \frac{\epsilon}{2}\frac{\partial E}{\partial z_i}(\hat{z}(\tau)) \\
\hat{z_i}(\tau + \epsilon) &= \hat{z_r}(\tau) + \epsilon\hat{r_i}(\tau + \frac{\epsilon}{2}) \\
\hat{r_i}(\tau + \epsilon) &= \hat{r_i}(\tau + \frac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial E}{\partial z_i}(\hat{z}(\tau + \epsilon))
\end{aligned}$$

Therefore, through this scheme we can push forward τ to $\tau + 1$ through $1/\epsilon$ leapfrog steps of update.

Now that we solve the discretization, however, the problem is that this approximation harms invariant property of Hamiltonian H, but this can be adjusted by plug in it to the acceptance rate formula.

And we have

$$min(1,\ exp\{H(z, r) - exp(H(z^*, r^*))\}) \tag{13}$$

the numerical errors lower the acceptance rate, but still reasonably good enough.

The specific sampling process is like this, suppose the current time is τ, we first choose at random, with equal probability (1/2), whether to integrate forwards in time (using step size ϵ) or backwards in time (using step size -ϵ).

The only thing left is to prove such transition satisfies detailed balance. Considering a region R before the evolution that has volume δV, after the evolution the region becomes R', by conservation of volume property, it still has volume δV.

The probability of starting from R and transit to $R^{'}$ writes:

$$p(R)T(R,R^{'}) = \frac{1}{Z_H} exp(-H(R))\delta V \frac{1}{2} min(1,\ exp(H(R)-H(R^{'}))) \tag{14}$$

while the probability of reverse transition writes:

$$p(R^{'})T(R^{'},R) = \frac{1}{Z_H} exp(-H(R^{'}))\delta V \frac{1}{2} min(1,\ exp(H(R^{'})-H(R))) \tag{15}$$

These two probability are equal, thus detailed balance holds.

5 Discussion

The intuition behind using physical dynamics for HMC is that while we are in high potential energy state (where $p(z)$ is low), we intend to get away soon with large jump (large momentum), and vice versa. This mechanism can be simulated in Hamiltonian system so that change of potential energy (gradient information of $\tilde{p}(z)$) helps us explore the state space more efficiently. Since each direction has its own momentum r_i and can be calculated individually, this dynamics can fit high dimension of z well.

The time complexity for leapfrog algorithm scales linearly with dimension of state variable z, which is d. But we might need a lot steps to calculate the trajectory with small error (requires small ϵ). The overall time complexity for HMC writes $O(Nd/\epsilon)$ if we construct a Markov chain of length N. Empirical experiments show that HMC usually beat other algorithms using MH framework for untidy multi-modal distribution. The interesting thing for me is that it is a combination of molecular dynamics, statistical mechanism and sampling approach.

Apart from this, several other variations in MH framework will always focus on how to pick up a good proposal as well.