Abstract

Junpeng Gao

January 2019

MCMC algorithm is an important sampling algorithm.However, it also has confines and because of random walk property, it actually has potential to be more effective. Hamilton dynamics system is a classical physics system.It uses the generalized momenta and coordinates to describe the system. By lifting the dimensions of probabilistic model into phase space, we can actually gain better results in many aspects.Some theoretical analysis and the geometry intuition behind the HMC is also included.

Some numerical experiments are also implemented, by comparing with MH algorithm, we will intuitively feel the better efficacy of HMC. However, HMC is also needed to be choose some hyperparameters and select the model it is suitable. We step further from this step and combine it with sequential monte carlo to be Hamilton Sequential Monte Carlo(HSMC) so that it can sample noncontinuous and multimodality models,which just HMC can't give a good result.

MCMC based on Hamilton dynamics and Squential model sampling

Junpeng Gao

***Abstract*—See in the extended paper**

I. INTRODUCTION

The traditional Markov chain Monte Carlo method is mainly Gibbs sampling and Random Walk Metropolis method. They have a random walk trait that can be used in many situations. However, they have many constraints and always converge slowly. There is another method called Hamilton Monte Carlo method, which is based on the Hamilton dynamics in physics and it lifts a dimension into phase space to discuss the result and get good results in many situations.This report will mainly have a discussion about the basic operations of Hamilton Monte Carlo Method and the geometry intuition, basic conditions it satisfies,moreover,its combination with other method or problem.

II. THE BASIC OPERATIONS OF HAMILTON MONTE CARLO METHOD

A. Introduction to Hamilton dynamics

There are three kinds of systems to describe the classical dynamics, Newton system, Lagrange dynamics and Hamilton dynamics.

Newton system is based on the Cartesian coordinates or Polar coordinates system and use the Newton laws to describe the dynamics system.

Then we introduce the system of analytic dynamics, use generalized coordinates to describe the dynamics. What the difference is, Lagrange system use generalized coordinates q_α and generalized velocity $\dot{q_\alpha}$, when tackling and analyzing the problem in the real world concerning constraints,which provides great help. Hamilton system used the generalized coordinates and generalized momenta to describe the whole system,which is beneficial to the description of dimension, which provides great tool to generalize this tool into quantum mechanics and other fields, just like this report to discuss, the Hamilton Monte Carlo method.

With the Hamilton dynamics system, we have

$$\frac{\partial H}{\partial q_\alpha} = -\dot{p_\alpha}$$

$$\frac{\partial H}{\partial p_\alpha} = \dot{q_\alpha}$$

H has the dimension of energy, we can have it as $H(p,q) = K(p) + U(q)$,then we can see

$$\frac{\partial U(p)}{\partial q_\alpha} = -\dot{p_\alpha}$$

$$\frac{\partial K(p)}{\partial p_\alpha} = \dot{q_\alpha}$$

B. Numerical simulation of dynamic system

The best method used in the simulation of dynamic system is leap frog method.The operation is as below:

$$p_i(t+\epsilon/2) = p_i(t) - (\epsilon/2)\frac{\partial U}{\partial q_i}(q(t))$$

$$q_i(t+\epsilon) = q_i(t) + \epsilon\frac{p_i(t+\epsilon/2)}{m_i}$$

$$p_i(t+\epsilon) = p_i(t+\epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial q_i}(q(t+\epsilon))$$

Firstly,the leap frog method is shear transformation so that it conserves volume.

Compared with it, in fact, we can also have explicit Euler method and implicit Euler method.

Explicit Euler method:

$$p_i(t+\epsilon) = p_i(t) - \epsilon\frac{\partial U}{\partial q_i}(q(t))$$

$$q_i(t+\epsilon) = q_i(t) + \epsilon\frac{p_i(t)}{m_i}$$

As we step forward we change p and q simultaneously based on former step ,which makes diverge result.

Implicit Euler method:

$$p_i(t+\epsilon) = p_i(t) - \epsilon\frac{\partial U}{\partial q_i}(q(t))$$

$$q_i(t+\epsilon) = q_i(t) + \epsilon\frac{p_i(t+\epsilon)}{m_i}$$

Implicit Euler method should have better presentation than explicit Euler method, however, leapfrog method have the steadiest situation.

C. Operation of Hamilton Monte Carlo

The distribution to sample is related to a potential energy function via the concept of a canonical distribution from statistical mechanics.Given some energy function,E(x),for the state,x, of some physical system, the canonical distribution over states has probability density function

$$P(x)\frac{1}{Z}exp(-E(x)/T)$$

T is the temperature of the system,Z is the normalizing constant needed for this function.

In the Hamilton system, H has the same dimension as energy, we set

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

$$P(q,p) = \frac{1}{Z} exp(-U(q)/T) exp(-K(p)/T)$$

Besides,we firstly set $K(p) = p^T Mp/2$
HMC Algorithm

1: set t=0
2: generate an initial position state $x^{(0)} \sim \pi^{(0)}$
3: repeat until t = M
4: set t = t+1
5: sample a new initial momentum variable from the momentum canonical distribution $p_0 \sim P(p)$
6: set $x_0 = x^{(t-1)}$
7: run Leap Frog algorithm starting at $(x_0, p_0)$for L step and step size δ to obtain proposed states x^* and p^*
8: calculate the Metropolis acceptance probability: $\alpha = min(1, \exp(-U(x^*) + U(x_0) - K(p^*) + K(p_0)))$
9: draw a random number u from Unif(0,1)
10: if $u \leq \alpha$ accept the proposed state position x^* and set the next state in the Markov chain $x^{(t)} = x^*$
else set $x^{(t)} = x^{(t-1)}$
=0

III. HMC ANALYSIS

A. theoretical analysis

For Markov chain Monte Carlo method, we need to propose a Markov chain that has the property of Ergodicity, Reversibility.Here the Hamilton Monte Carlo also has these properties and it also has volume preservation in the phase space,namely(q,p)space .

We will discuss reversibility firstly. Hamilton dynamics is reversible, for a state (q_0, p_0) after L steps to (q_L, p_L), inverse the p as -p,at this time K(p)=K(-p), and again apply the mapping, we return to $(q_0, -p_0)$,and inverse the -p as p again. we are back to $(q_0.p_0)$ now.

Volume preservation: From classical Hamilton system, it is easy to know it satisfies. From the perspective of prbabilistic system, we can make an analogy to the classical system.We lift our target distribution onto a joint probability distribution on phase space,it is possible to do shear transformation, which keeps the volume of the (p,q), to keep transforming oppositely to each other. so here give a proof:

$$\sum_{i=1}^{d} [\frac{\partial}{\partial q_i}\frac{dq_i}{dt} + \frac{\partial}{\partial p_i}\frac{dp_i}{dt}] = \sum_{i=1}^{d} [\frac{\partial}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial}{\partial p_i}\frac{\partial H}{\partial q_i}] =$$

$$\sum_{i=1}^{d} [\frac{\partial^2 H}{\partial q_i \partial p_i} - \frac{\partial^2 H}{\partial p_i \partial q_i}] = 0$$

Ergodicity of HMC: Because the phase space contains the whole target trajectory, every time sampling through the phase trajectory and leap a random lift.To verify this property, it seems ergodic geometry and differential geometry needed. In some cases it seems not that good to be realized, we need to tune some hyperparameters sometimes.

B. geometry intuition in high dimensions space

First,we will discuss the geometry of high dimension space. Consider a point and its neighborhood with $r = \delta$ and $r = 2\delta$. With the dimension growing, the value of volume V2(neighborhood of radius 2δ) minus value of volume V1(neighborhood of radius δ) take an exponential growing.

Generally,the volume is largest out in the tails of the target distribution away from the mode,and this disparity grows exponentially with the dimension of parameter space.Consequently, the massive volume over which we integrate can compensate to give a significant contribution to the target expectation despite the smaller density.

The neighborhood immediately around the mode features large densities,but in more than a few dimensions the small volume of that neighborhood prevents it from having much contribution to any expectation.

The only significant contributions come from the neighborhood between these two extremes known as the typical set. With the dimension grows, the tension between the density and the volume grows so that the typical set takes a significant proportion.

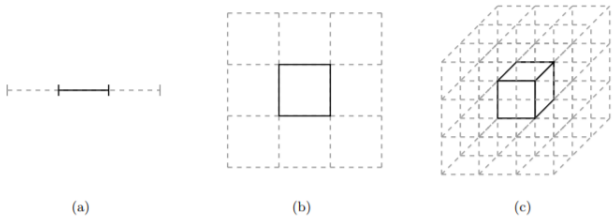

Fig. 1. dimension grows

C. Advantages and Numerical results

Here we will show examples to make a comparison between Hamilton Monte Carlo and Markov chain Monte Carlo method to show that our progress is robust and effective.
We firstly discuss the Rosenbrock's banana function defined as follow:

$$f(\theta) \propto exp((-5(y - x^2)^2 - x^2)/8)$$

Using the random work MH algorithm, we run two trials with different for the $N(\theta, \sigma I)$ proposal.
When $\sigma = 0.2$, the acceptance rate is 66.1% but the chain fails to cover the distribution after 1000iterations. When $\sigma = 1$, the coverage improves but the acceptance rate decreases to 39.6%.
Using the HMC approach, with $\epsilon = 0.05$ and L = 20 steps, the target density is well covered and the acceptance rate is 99.8%.

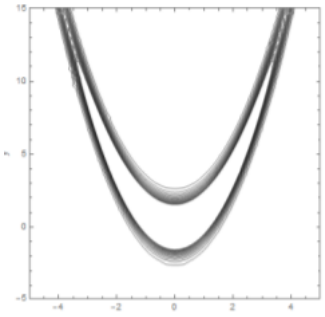

Fig. 2. the Rosenbrock's banana function

MH algorithm simulation:

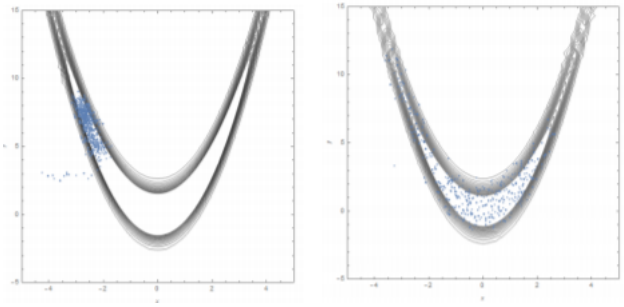

Fig. 3. HM simulation

HMC simulation:

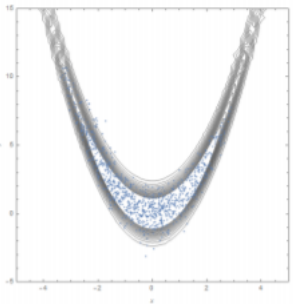

Fig. 4. HMC simulation

From this experiment,we can see that HMC is actually have a better convergent rate and ergodic property.This is for the ordinary function. In addition, in the situation that the dramatic variations in curvature happens, hamilton monte carlo will have apparently better presentation.
For a hierarchical model:

where Ds are data. For the infamous Eight Schools model, we follow the simulation,consider dataset:

$$\mu \sim N(0, 5)$$

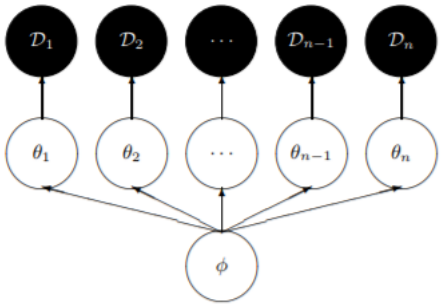

Fig. 5. hierarchical model

$$\tau \sim Half - Cauchy(0, 5)$$

$$\theta_n \sim N(\mu, \tau)$$

$$y_n \sim N(\theta_n, \sigma_n)$$

Where $n \in \{1, ..., 8\}$ and the $\{y_n, \sigma_n\}$ are given as data.We sample from such model and draw the plot between $theta_n$ and $log(\tau)$.The aiming figure should be funnel-shaped. First we do a centered eight schools implementation:

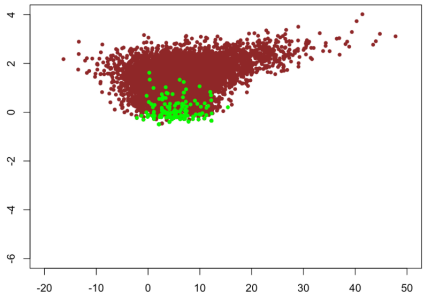

Fig. 6. hierarchical model

Non-Centered Eight schools simulation:

$$\mu \sim N(0, 5)$$

$$\tau \sim Half - Cauchy(0, 5)$$

$$\widetilde{\theta}_n \sim N(\mu, \tau)$$

$$_n = \mu + \tau\dot{\widetilde{\theta}}_n$$

From the results, we can see clearly that based on the non-centered model the HMC algorithm can simulate much better dramatic variations in curvature. This result can be greatly used in bayes networks.There are also other methods like combining with gibbs sampling so that it can get samples better. Besides, parameters about HMC and the model selection are also needed to be cautious to choose, in the next section, we will do more properly extension.

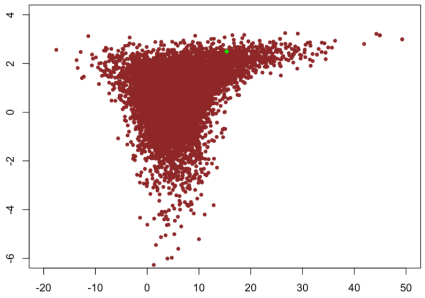

Fig. 7. hierarchical model

IV. Improvements for Hamilton Monte Carlo

A. HSMC introduction

Though Hamilton Monte Carlo gives a good presentation in many situations of what MCMC can not, however, HMC also confines in some simulations such as noncontinuous functions.
To alleviate such problem, we could combine it with Sequential Monte Carlo.

We introduce Sequential Monte Carlo firstly.SMC methods use multiple particles moving in parallel, the goal is to obtain a sample $\{\theta_n\}_{n=1}^N$ from a sequence of distributions with densities $f_1(\theta_n), ..., f_T(\theta_n)$.The simulator we propose works best when target densities in the sequence are smoother at the beginning and progressively converging toward the final sharper target density.
The algorithm is as follows:

1.initialization: Draw N particles $\{\theta_n^{(0)}\}_{n=1}^N$ from $f_0(\theta_n)$

2.Repeat for t = 1,...,T

(a) Correction: assign weight $w_n^{(t)} = f_t(\theta_n)/f_{t-1}(\theta_n)$ to each of the particles $\{\theta_n^{(t-1)}\}_{n=1}^N$

(b)Selection: draw N new particles $\{\hat{\theta}_n^{(t-1)}\}_{n=1}^N$ with replacement from the current sample of particles using weights $w_n^{(t)}$.Give the new particles a weight of 1

(c)Mutation: For each particle, perform a MH algorithm step to obtain a new sample of particles$\{\theta_n^{(t)}\}_{n=1}^N$

We will use HMC to take the place of the last step.The re-sampling is also done using leave-one-out approximation $\{\hat{\theta}_n^{(t-1)}\}_{n=1}^N$ of the observed distribution of the particles instead of the theoretical distribution $\{\theta_n^{(t-1)}\}_{n=1}^N$ as in specific cases the particles do not have the time to converge to their stationary distribution in one step.

1.initialization: Draw N particles $\{\theta_n^{(0)}\}_{n=1}^N$ from $f_0(\theta_n)$

2.Repeat for t = 1,...,T

(a) Correction: assign weight $w_n^{(t)} = f_t(\theta_n)/\hat{f}_{t-1}(\theta_n)$ to each of the particles $\{\theta_n^{(t-1)}\}_{n=1}^N$

(b)Selection: draw N new particles $\{\hat{\theta}_n^{(t-1)}\}_{n=1}^N$ with replacement from the current sample of particles using weights $w_n^{(t)}$.Give the new particles a weight of 1

(c)Mutation: For each particle, perform a HMC step to obtain a new sample of particles$\{\theta_n^{(t)}\}_{n=1}^N$
We call this algorithm Hamilton Sequential Monte Carlo (HSMC).It shows good representation in the noncontinuous function and multimodality function.

B. Numerical examples

Here we sample a dropwave function.We generate a sample of 4096 data points with coordinates(x,y) using a density proportional to the following function defined on [-2.5;2.5]×[-2.5;2.5]:

$$g(x,y) = exp(\frac{cos(5\sqrt{x^2+y^2}+1}{x^2+y^2+2})$$

We used 4 independent groups of 512 particles to simulate the kernel density for a total of 2048 particles. The data points are partitioned in 40 blocks of 100 points and 1 block of 96 points, for a total of 41 blocks Consequently, the sequence will include T = 41 target functions with the addition of the initial density.The lowest mutations acceptance rate we observed across several runs was 2023/2048. We can see from the result the simulation quite smooth.

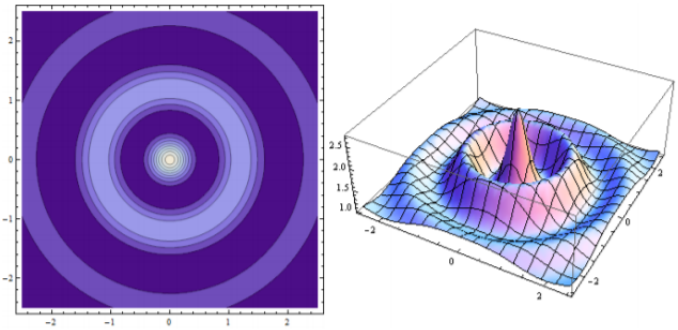

Fig. 8. Contour plot and 3D plot of the drop wave function

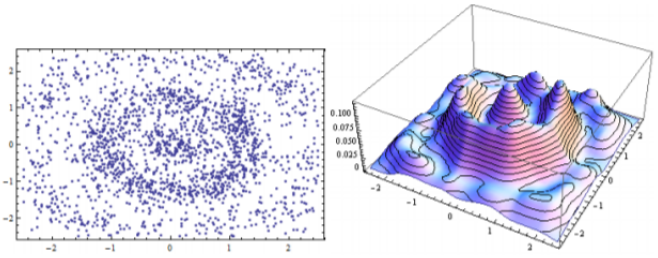

Fig. 9. HSMC simulation results

V. CONCLUSIONS

We improve the classical MCMC algorithm using Hamilton dynamics. It can be used into many fields and can sample many functions. The HMC totally has a better property concerning convergence and ergodicity. Its great use in the Hierarchical Model, which will be used for bayes networks.
We discuss the constraints about HMC, and give a concept of HSMC, which uses particles sampling parallel and solve the problem to some extent. Now we have great tool to sample many kinds of functions and get pretty good results.

REFERENCES

[1] Radford M. Neal, University of Toronto MCMC using Hamiltonian dynamics Published as Chapter 5 of the Handbook of Markov Chain Monte Carlo, 2011

[2] Michael Betancourt A Conceptual Introduction to Hamiltonian Monte Carlo arXiv:1701.02434 [stat.ME]

[3] Remi Daviet Inference with Hamiltonian Sequential Monte Carlo Simulators

[4] Michael Betancourt and Mark Girolami Department of Statistical Science, University College London, London, UK. Hamiltonian Monte Carlo for Hierarchical Models