Hamiltonian Monte Carlo - Write up

Paper: A Conceptual Introduction to Hamiltonian Monte Carlo [1]

Philip Kofoed-Djursner & Casper Wied
KU-ID: tkv976 & nqs117

8th of March 2023

Abstract

One of the most frequent implementations of Markov Chain Monte Carlo uses the Metropolis-Hastings Sampling algorithm. This method is defined by random walk behavior and is sensitive to correlated parameters, which plagues it in high dimensionality. Hamiltonian Monte Carlo proposes a solution to these problems by using first-order gradient information to determine steps. This allows it to converge rapidly in high dimensional parameter space. This method comes with its own drawback in the form of being incredibly sensitive to two user specified parameters, step size and number of steps.

1 Introduction

One of the main challenges in most scientific fields arises when one has to search a multidimensional parameter space and find the optimal answer. Markov Chain Monte Carlo (MCMC) is a computational method which efficiently samples the parameter space to find the typical set. There exist numerous different methods and algorithms which implement MCMC, but one of the most known and used is Metropolis-Hastings Sampling (MHS). This method is especially efficient in low dimensional problems, but suffers in high dimensional parameter space. This paper provides an alternative in the form of Hamiltonian Monte Carlo (HMC), which is based of first-order gradient information of the parameter space and an auxiliary momentum.

2 Review

2.1 The Typical Set

Regarding MCMC, it is important to have a robust understanding of the typical set. When exploring parameter space, especially with high dimensionality, all regions should not be treated equally. One needs to consider the balance between probability density ($\pi(q)$) and volume (dq) in parameter space. If one approximates the D-dimensional parameter space to a D-dimensional hyper sphere, it quickly becomes obvious that while the mode, which has the highest probability density, lies in the center of the sphere, it does not cover a large volume. And vice versa when looking at the edge of the sphere, the region contains a large volume, but a low probability density. The region that one truly wants to explore is then the balance between the two extremities, called the typical set, which is the area with highest probability. The width of the typical set narrows exponentially with increasing dimensions. Figure 1 shows a 1-dimensional visualization of the typical set as balance of probability density and volume.

2.2 Metropolis-Hastings

MHS is one of the simplest implementations of MCMC. It utilizes random walking to explore the typical set. The next point in the Markov Chain is initialized by the proposal function which lies a distance from the previous point which is chosen randomly from a predetermined Gaussian distribution. This proposal function is symmetric which reduces the acceptance criterion to a simple formula which is 1 if the new point lies in a region of higher probability density and the ratio between the new and old points' probability density if not.

$$a(q' \mid q) = \min\left(1, \frac{\pi(q')}{\pi(q)}\right) \tag{2.1}$$

The random walk nature of the method is both its strength and weakness. It makes it exceptionally easy to implement, but in higher dimensionality, as new points are Gaussian distributed about the old point, it explores the typical set increasingly inefficiently. Moreover, the points become correlated as can be seen in Figure 2. If the variance of the proposal is large, the MCMC will move away from the typical set, due to the large volume outside. If the variance is small, the MCMC explores the typical set very slowly.

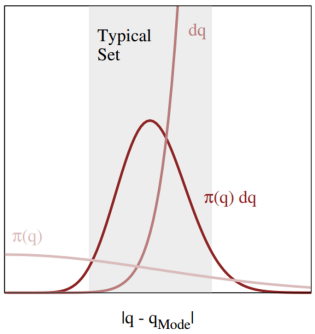

Figure 1: The typical set is the gaussian spanning the product of $\pi(q)$ and dq.

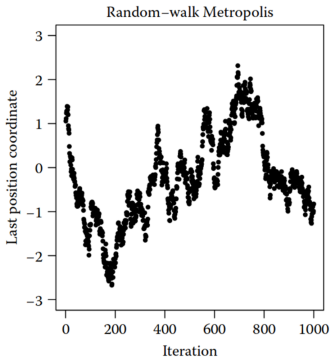

Figure 2: Correlation between iterations in MHS.[2]

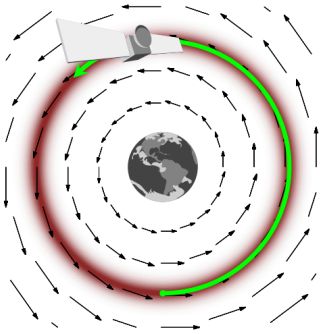

Figure 3: HMC moving along the typical set as a satellite orbiting earth.

2.3 Hamiltonian Monte Carlo

HMC is a MCMC method which utilizes first-order gradient information of parameter space to more efficiently survey the typical set. Naively following the gradient would always lead to the mode, which does not contain much information. Combining the gradient information with an auxiliary momentum allows the proposal function to glide through parameter space moving large distances and still remaining within the typical set, at least in theory as shown in Figure 3. The dimensionality is doubled to allow an accompanied momentum for each parameter thereby defining the joint density within phase space. The definition of the joint density, which is a function of both position and momentum, is shown Eq. 2.2.

$$\pi(q,p) = \pi(p \mid q)\pi(q) = e^{-H(q,p)} \tag{2.2}$$

The time-evolution of a physical system can be precisely described by Hamiltonian mechanics. The ideas developed in classical mechanics can inspire how to create the proposal of Hamiltonian Monte Carlo. Therefore, kinetic and potential energy is defined from the joint density (Eq. 2.3)

$$\begin{aligned} H(q,p) &= -\log\pi(p|q) - \log\pi(q) \\ &\equiv K(p,q) + V(q) \end{aligned} \tag{2.3}$$

Having defined the kinetic and potential energies, Hamilton's equations describe the evolution of position and momentum (Eq. 2.4).

$$\begin{aligned} \frac{\mathrm{d}q}{\mathrm{d}t} &= +\frac{\partial H}{\partial p} = \frac{\partial K}{\partial p} \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= -\frac{\partial H}{\partial q} = -\frac{\partial K}{\partial q} - \frac{\partial V}{\partial q} \end{aligned} \tag{2.4}$$

Having now the description of the time-evolution of the position and momentum it can be integrated over a time interval to arrive at a new state on the given Hamiltonian level curve. If tuned correctly this new point should be within the typical set.

In the implementation of Hamiltonian Monte Carlo the tuning of variables is the biggest challenge. Firstly, the parameter space needs to be sampled repeatedly to build up the covariance matrix (Σ) which is used to inform the momentum choice, which creates an unbiased MCMC.[3] Moreover, the integration over time is done by Leapfrog Updates which takes two variables; the step size ϵ, and integration time T. This results in T/ϵ Leapfrog Updates. ϵ needs to be balanced, as two small steps will slow down the algorithm and two large can make it leave the Hamiltonian level curve. If T is too small the parameter space will not be surveyed efficiently and if it is too large the point can end up returning to its original position and accomplish nothing. Even if well-calibrated, an acceptance criterion similar to that in MHS (Eq. 2.1), but based on the value of the Hamiltonian, is used to correct the small errors made by Leapfrogging. Each new integration starts with a random momentum to ensure the geometric ergodicity of the method.

3 Conclusion

The paper highlights the importance of understanding how the typical set emerges as a balance between probability density and parameter volume, and how the set narrows exponentially with increasing dimensionality. In conclusion, Hamiltonian Monte Carlo is a computational implementation of Markov Chain Monte Carlo, which is most effective in high dimensional problem solving, where simple implementations, like Metropolis-Hastings sampling, have limited effectiveness. The paper provides a clear guide to understanding Hamiltonian Monte Carlo, without the need to have an in-depth mathematical understanding of differential geometry, and instead draws on classical physics intuition. In addition, the paper also highlights the challenge of choosing optimal parameters which, if not chosen correctly, heavily limits the use of the method it.

References

[1] Michael Betancourt. *A Conceptual Introduction to Hamiltonian Monte Carlo*. 2017. DOI: 10.48550/ARXIV.1701.02434. URL: https://arxiv.org/abs/1701.02434.

[2] Lucas Mahler. "Randomized Algorithms - Using Randomness to Solve Increasingly Complex Problems - A Review". In: (Dec. 2018).

[3] Adrian Barbu and Song-Chun Zhu. *Monte Carlo Methods*. Springer Nature Singapore, 2020.