# Conceptual motivation of MCMC using Hamiltonian Dynamics

Anubhav Mittal

EE392A

Supervised by

Prof. Satyadev Nandakumar, Department of Computer Science and Engineering
Prof. K. Vasudevan, Department of Electrical Engineering

November 16, 2019

# Introduction

- The aim of this review is to first show the problem of computing expectations encountered in high dimensions and show how the usual MCMC performs.
- Then we describe Hamiltonian dynamics, and show how to use it to construct a Markov chain Monte Carlo method.
- We then discuss the different approximations involved, as well as the problems that arise and how to address them.

# Expectation of a given function

- The ultimate undertaking in statistical computing is evaluating expectations with respect to some distinguished target probability distribution. For example, we might be interested in extracting information from a posterior distribution over model configuration space in Bayesian inference.
- We consider a target distribution, $\pi$, on a D-dimensional sample space, $Q$, and the corresponding expectations of functions, $E_\pi[f]$. Assuming the distribution is smooth over the space $Q$, we have:

$$E_\pi[f] = \int_Q \pi(q)f(q)dq$$

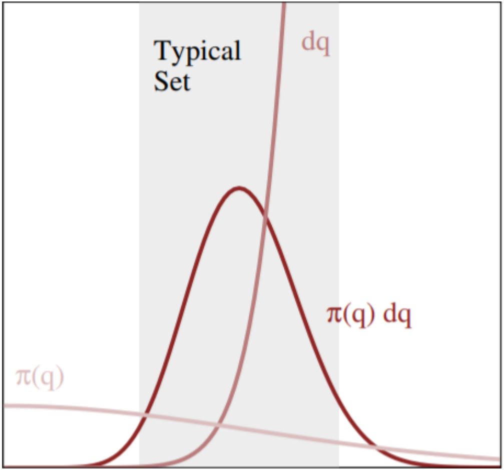
- For any nontrivial target distribution we will not be able to evaluate these integrals analytically, and we must instead resort to numerical methods which only approximate them.

# Typical set

- Intuitively, we should consider the neighborhood around the mode where the density is maximized.
- However, expectation values are given by accumulating the integrand over a volume of parameter space and, while the density is largest around the mode, there is not much volume there.
- 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. On the other hand, the complimentary neighborhood far away from the mode features a much larger volume, but the vanishing densities lead to similarly negligible contributions expectations

# Typical set

- The only significant contributions come from the neighborhood between these two extremes known as the **typical set**, as shown in the figure.



# Estimation using MCMC

- Markov chain Monte Carlo uses a Markov chain to stochastically explore the typical set, generating a random grid across the region of high probability from which we can construct accurate expectation estimates. Given sufficient computational resources a properly designed Markov chain will eventually explore the typical set of any distribution.
- The more practical, and much more challenging question, however, is whether a given Markov chain will explore a typical set in the finite time available in a real analysis.

# Estimation using MCMC

- For constructing a Markov chain, we can think of a Markov transition as a conditional probability density, $T(q'|q)$, defining to which point, $q'$, we are most likely to jump from the initial point, $q$.
- $$\pi(q) = \int_Q dq' \pi(q') T(q|q'))$$

  So long as this condition holds, at every initial point the Markov transition will concentrate towards the typical set.
- Using the samples generated by the Markov chain $q_0, ..., q_N$, we use a monte-carlo estimate of the function:

  $$\hat{f}_N = \frac{1}{N} \sum_{n=0}^{N} f(q_n)$$

- If the chain is run long enough, and under certain assumption on the target density, we can claim that the estimate will converge to the true expectation:

  $$\lim \quad \hat{f}_N = E_\pi[f]$$

# Estimation using MCMC

Ideally, the chain will have three phases:

- In the first phase the Markov chain converges towards the typical set from its initial position in parameter space while the Markov chain Monte Carlo estimators suffer from strong biases.
- The second phase begins once the Markov chain finds the typical set and persists through the first sojourn across the typical set. This initial exploration is extremely effective and the accuracy of Markov chain Monte Carlo estimators rapidly improves as the bias from the initial samples is eliminated.
- The third phase consists of all subsequent exploration where the Markov chain refines its exploration of the typical set and the precision of the Markov chain Monte Carlo estimators improves, albeit at a slower rate

# The Metropolis-Hastings Algorithm

- The Metropolis-Hastings algorithm is comprised of two steps: a proposal and a correction. The proposal is any stochastic perturbation of the initial state while the correction rejects any proposals that stray too far away from the typical set of the target distribution.
- More formally, let $Q(q'|q)$ be the probability density defining each proposal. The probability of accepting a given proposal is then given by

$$a(q'|q) = \min(1, \frac{Q(q|q')\pi(q')}{Q(q'|q)\pi(q)})$$

- The most common proposal distribution used is Gaussian, where the algorithm is now called **Random-Walk Metropolis** and due to symmetry of the gaussian distribution, the acceptance probability takes the following form:

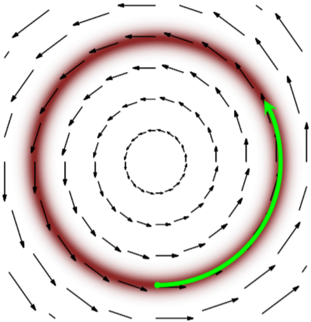
$$a(q'|q) = \min(1, \frac{\pi(q')}{\pi(q)})$$

# Problems with Random Walk Metropolis

- The idealized behavior of MCMC requires that the Markov transition is compatible with the structure of the target distribution. When the target distribution exhibits pathological behavior, like in a target probability distribution where the typical set pinches into a region of high curvature, Markov transitions will have trouble exploring and Markov chain Monte Carlo will fail.
- Even if we have a well-behaved target distribution, we may still encounter problems.As the dimension of the target distribution increases, the volume exterior to the typical set overwhelms the volume interior to the typical set, and almost every Random Walk Metropolis proposal will produce a point on the outside of the typical set, towards the tails. The density of these points, however, is so small, that the acceptance probability becomes negligible. In this case almost all of the proposals will be rejected and the resulting Markov chain will only rarely move.

# Need for better transition functions

- As seen in the problems encountered above, we need to define a different transition function. We need to exploit information about the geometry of the typical set. Specifically, we need transitions that can follow those contours of high probability mass, coherently gliding through the typical set.
- When the sample space is continuous, a natural way of encoding this direction information is with a vector field aligned with the typical set.



# Hamiltonian Dynamics

- Hamiltonian dynamics has a physical interpretation that can provide useful intuitions. In two dimensions, we can visualize the dynamics as that of a frictionless puck that slides over a surface of varying height.
- The state of this system consists of the position of the puck, given by a 2D vector $q$, and the momentum of the puck (its mass times its velocity), given by a 2D vector $p$.
- The potential energy, $U(q)$, of the puck is proportional to the height of the surface at its current position, and its kinetic energy, $K(p)$, is equal to $|p|^2/(2m)$, where $m$ is the mass of the puck.
- In non-physical MCMC applications of Hamiltonian dynamics, the position will correspond to the variables of interest. The potential energy will be minus the log of the probability density for these variables. Momentum variables, one for each position variable, will be introduced artificially.

# Hamilton’s equations

- Hamiltonian dynamics operates on a d-dimensional position vector, q, and a $d$-dimensional momentum vector, $p$, so that the full state space has $2d$ dimensions. The system is described by a function of $q$ and $p$ known as the Hamiltonian, $H(q, p)$.
- **Equations of motion** The partial derivatives of the Hamiltonian determine how q and p change over time, t, according to Hamiltons equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

  for $i = 1, ..., d$.

# Hamilton's equations

- **Potential and kinetic energy** For Hamiltonian Monte Carlo, we usually use Hamiltonian functions that can be written as follows:

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

  Here, $U(q)$ is called the potential energy, and will be defined to be minus the log probability density of the distribution for $q$ that we wish to sample, plus any constant that is convenient. $K(p)$ is called the kinetic energy, and is usually defined as

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

- With these forms, the Hamilton's equations reduce to the following:

  $$\frac{dq_i}{dt} = [M^{-1}p]_i$$

  $$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}$$

# Properties of Hamiltonian dynamics

Some of the general properties which are necessary for its use in MCMC are as follows:

- **Reversibility** Hamiltonian dynamics is reversible the mapping $T_s$ from the state at time $t, (q(t), p(t))$, to the state at time $t + s, (q(t + s), p(t + s))$, is one-to-one, and hence has an inverse, $T_s$.
- **Conservation of the Hamiltonian** A second property of the dynamics is that it keeps the Hamiltonian invariant (conserved). For Metropolis updates using a proposal found by Hamiltonian dynamics, which form part of the HMC method, the acceptance probability is one if H is kept invariant. However, in practice we can only make H approximately invariant, and hence we will not quite be able to achieve this.

# Properties of Hamiltonian dynamics

- **Volume preservation** A third fundamental property of Hamiltonian dynamics is that it preserves volume in $(q, p)$ space (a result known as Liouvilles Theorem). If we apply the mapping $T_s$ to the points in some region $R$ of $(q, p)$ space, with volume $V$ , the image of $R$ under $T_s$ will also have volume $V$ . The preservation of volume by Hamiltonian dynamics can be proved in several ways. One is to note that the divergence of the vector field defined in the Hamilton's equations is zero,

# Discretizing the Hamilton's equations

- For simplicity, we assume that M is diagonal, with diagonal elements

$$m_1, ..., m_d$$

, so that:

$$K(p) = \sum_{i=1}^{d} \frac{p_i^2}{m_i}$$

- **Eulers method** For Hamiltons equations, this method performs the following steps, for each component of position and momentum, indexed by $i = 1, ..., d$:

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

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

  In practise, however, the results are not that good, so we slightly modify the method to our liking.

# Discretizing the Hamilton's equations

- **A modification of Eulers method** Much better results can be obtained by slightly modifying Eulers method, as follows:

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

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

- The leapfrog method Even better results can be obtained with the leapfrog method, which works as follows:

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

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

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

# MCMC from Hamiltonian dynamics

- Using Hamiltonian dynamics to sample from a distribution requires translating the density function for this distribution to a potential energy function and introducing momentum variables to go with the original variables of interest (now seen as position variables).
- **Canonical distributions** The distribution we wish to sample can be 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 or probability density function:

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

# MCMC from Hamiltonian dynamics

- Viewing this the opposite way, if we are interested in some distribution with density function $P(x)$, we can obtain it as a canonical distribution with $T = 1$ by setting $E(x) = logP(x)logZ$, where $Z$ is any convenient positive constant.

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

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

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

- We see that $q$ and $p$ are independent, and each have canonical distributions, with energy functions $U(q)$ and $K(p)$. We will use $q$ to represent the variables of interest, and introduce $p$ just to allow Hamiltonian dynamics to operate.

# The algorithm for Hamiltonian MCMC

We use the values of $K(p)$ and $U(q)$ as defined before. There are two main steps in the algorithm:

- In the first step, new values for the momentum variables are randomly drawn from their Gaussian distribution, independently of the current values of the position variables. For the kinetic energy, the $d$ momentum variables are independent, with $p_i$ having mean zero and variance $m_i$ . Since $q$ isnt changed, and $p$ is drawn from its correct conditional distribution given $q$ (the same as its marginal distribution, due to independence), this step obviously leaves the canonical joint distribution invariant.

# The algorithm for Hamiltonian MCMC

- For the second step, we need some background for the details. There is one important exception to performance of the leapfrog method. Long time accuracy can be compromised when the exact energy level sets feature neighborhoods of high curvature that the finite time discretization is not able to resolve. These neighborhoods induce a divergence that almost immediately propels the numerical trajectory towards infinite energies.
- An intuitive way to correct this is to treat the Hamiltonian transition as the proposal for a Metropolis Hastings scheme on phase space.
- We encounter a problem here. Because Hamiltonian trajectories, and their numerical approximations, are deterministic and non-reversible, Metropolis-Hastings proposals are always rejected. In particular, we have positive proposal probabilities going forwards in time but vanishing proposal probabilities going backwards in time which renders the Metropolis-Hastings acceptance probability identically zero.

# The algorithm for Hamiltonian MCMC

- If we modify the Hamiltonian transition to be reversible, however, then the ratio of proposal densities becomes non-zero and we achieve a useful correction scheme. The simplest way of achieving a reversible proposal is to augment the the numerical integration with a negation step that flips the sign of momentum,

$$(p_L, q_L) \rightarrow (p_L, -q_L)$$

- Hence, the momentum variables at the end of the L-step trajectory are then negated, giving a proposed state $(q*, p*)$. This proposed state is accepted as the next state of the Markov chain with probability:

$$min[1, exp(-H(q*, p*) + H(q, p)] = min[1, exp(-U(q*) + U(q) - K(p*) +$$

- If the proposed state is not accepted (ie, it is rejected), the next state is the same as the current state (and is counted again when estimating the expectation of some function of state by its average over states of the Markov chain).

# Ergodicity of Hamiltonian Monte Carlo

- Typically, the HMC algorithm will also be ergodic  it will not be trapped in some subset of the state space, and hence will asymptotically converge to its (unique) invariant distribution.
- In an HMC iteration, any value can be sampled for the momentum variables, which can typically then affect the position variables in arbitrary ways.
- However, ergodicity can fail if the $L$ leapfrog steps in a trajectory produce an exact periodicity for some function of state.
- This potential problem of non-ergodicity can be solved by randomly choosing $\epsilon$ or $L$ (or both) from some fairly small interval.

# References I

M Betancourt.
A general metric for riemannian hamiltonian monte carlo.
In *First International Conference on the Geometric Science of Information (F. Nielsen and F. Barbaresco, eds.). Lecture Notes in Computer Science*, volume 8085, 2013.

Michael Betancourt.
A conceptual introduction to hamiltonian monte carlo.
2017.

Paul B Mackenze.
An improved hybrid monte carlo method.
*Physics Letters B*, 226(3-4):369–371, 1989.

Radford M Neal et al.
Mcmc using hamiltonian dynamics.
*Handbook of markov chain monte carlo*, 2(11):2, 2011.

# References II

Radford M Neal.
An improved acceptance procedure for the hybrid monte carlo algorithm.
*Journal of Computational Physics*, 111(1):194–203, 1994.

Radford M Neal.
*Bayesian learning for neural networks*, volume 118.
Springer Science & Business Media, 2012.