Hamiltonian Monte Carlo

Alexander Fengler

October 23, 2017

Table of contents

Standard Bivariate Gaussian

$$f(\mathbf{x}|\Sigma, \mu = 0) = \frac{1}{det(2\pi\Sigma)^{-\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x}}$$

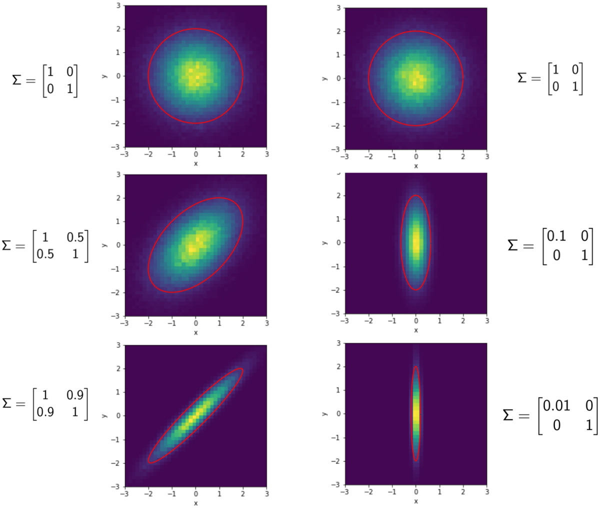
As a running example we will use the **Bivariate Gaussian** distribution.

It is simple, but enough to illustrate basic **shortcomings**, of the **Metropolis** and **Gibbs** samplers.

These shortcomings get exacerbated in **high dimensions**.

Standard Bivariate Gaussian

To show the limitations of **Metropolis** and **Gibbs** samplers, we consider the following covariance matrix structures, respectively.



Metropolis

Consider the basic **Metropolis** sampler with symmetric **proposal distribution**,

$$q \; N(0, \sigma \mathbf{I})$$

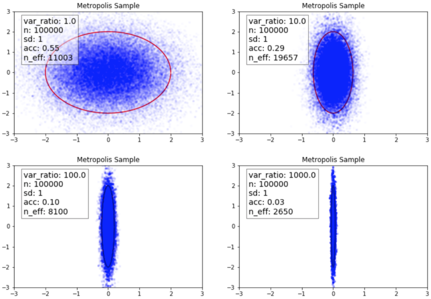
We have access to **one parameter**, σ, the **standard deviation of the proposal**.

Metropolis

Keeping $\sigma = 1$ constant, consider the following **target distributions**,

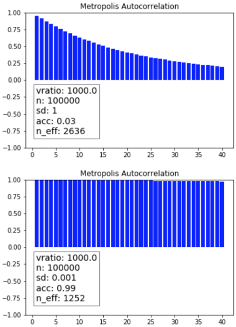
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 0.001 & 0 \\ 0 & 1 \end{bmatrix}$$



Metropolis

Changing σ helps in adjusting the **acceptance rate**, but also affects **autocorrelation**, and therefore the **effective sample size**.



It takes the random walk too long to cover significant distance in space.

Metropolis

For the **Metropolis Random Walk Algorithm**, the **proposal standard deviation** is limited by the **dimension with lowest variance**.

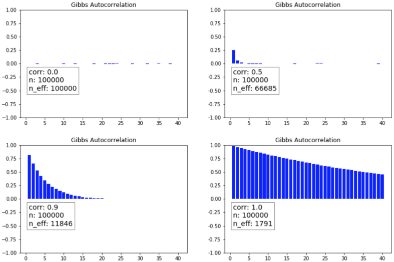
Greatly **uneven variances** across dimensions, negatively impact the performance of the sampler greatly.

We want to **control the acceptance rate**, but on the **cost of speed of exploration** of the target space (**mixing**).

Gibbs

Next we consider the **Gibbs sampler**. We observe that correlation in the **target distribution** introduces **autocorrelation** into the **Markov Chain**.

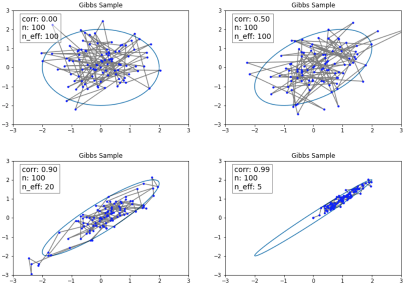
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$$

Gibbs

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$



$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

The **geometry** of a target distribution with **high correlation**, disfavors the constriction of the **Gibbs sampler** to consecutive vertical and horizontal steps.

Section Summary

Root cause for inefficiency of the **simple metropolis sampler** is the **random walk behavior**.

If the respective **standard deviations** of the target distribution greatly differ by dimension, exploration of the sampler space can be very inefficient. (Adaptations to overcome this issue exist, but not discussed here)

Root cause for inefficiency of the **gibbs sampler**, is the restriction to vertical and horizontal steps (in 2 dimensions), the efficiency of which depends on the geometry of the target distributions.

HMC: Goal

The motivation behind **Hamiltonian Monte Carlo** methods is to find a sampling scheme that is able to provide good **mixing properties**, from distributions with difficult geometric properties in high dimensions.

The **aim** is to allow for movements in arbitrary directions in space (overcoming limitations of **Gibbs sampler**), while avoiding the shortcomings of simple random walk behavior (overcoming limitations of **Metropolis sampler**).

HMC

Hamiltonian Monte Carlo falls under the class of **auxiliary variable methods**.

Mini-recap

$\mathbf{X} \sim f(\mathbf{x})$, where $f(\mathbf{x})$ can be **evaluated**, but not easily **sampled**. Then,

1. Augment $\mathbf{X}$, by a vector of **auxiliary variables**, $\mathbf{U}$
2. Construct a **Markov Chain** over $(\mathbf{X}, \mathbf{U})$, with **stationary distribution** $(\mathbf{X}, \mathbf{U}) \sim f(\mathbf{x}, \mathbf{u})$, that **marginalizes** to the **target**, $f(\mathbf{x})$
3. Discard $\mathbf{U}$ and do **inference** based on $\mathbf{X}$ only

HMC: Setup

Hamiltonian Dynamics

- d-dimensional **position vector** q
- d-dimensional **momentum vector** p
- **Hamiltonian** $H(q, p)$

Where,

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

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

Important Property: 1

Property: Reversibility

Define, T_s the mapping from **state** at **time** t, $(q(t), p(t)$ to the **state** at **time** $t + s$, $(q(t + s), p(t + s))$.

The mapping is **one-to-one** and therefore has an **inverse** T_{-s} (obtained by **negating derivatives** in **Hamiltonian equations**)

Important, because it is backbone of proof that **MCMC updates** by **Hamiltonian Dynamics** leave the desired distribution invariant.

Important Property: 2

Property: Conservation of Hamiltonian
Hamiltonian dynamics keep $H(q,p)$ invariant.

Proof:

$$\begin{aligned}\frac{d\mathbf{H}}{dt} &= \sum_{i=1}^{d}\left[\frac{dq_i}{dt}\frac{\partial \mathbf{H}}{\partial q_i} + \frac{dp_i}{dt}\frac{\partial \mathbf{H}}{\partial p_i}\right] \\ &= \sum_{i=1}^{d}\left[\frac{\partial \mathbf{H}}{\partial p_i}\frac{\partial \mathbf{H}}{\partial q_i} - \frac{\partial \mathbf{H}}{\partial q_i}\frac{\partial \mathbf{H}}{\partial p_i}\right] \\ &= 0\end{aligned}$$

Important, because this implies that for **metropolis updates** using a proposal found via **Hamiltonian Dynamics**, we get an **acceptance probability** of 1.

Important Property: 3

Property: Symplecticness

Let $z = (q, p)$, then we can write the **Hamiltonian equations** as:

$$\frac{dz}{dt} = \mathbf{J}\nabla\mathbf{H}$$

where,

$$J = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -I_{d\times d} & 0_{d\times d} \end{bmatrix}$$

Symplectiness means that the **Jacobian** of T_s, $\mathbf{B_s}$ satisfies,

$$\mathbf{B}_s^T\mathbf{J}^{-1}\mathbf{B_s} = \mathbf{J}^{-1} \rightarrow det(\mathbf{B}_s) = 1$$

Implies volume preservation of hamiltonian dynamics, which is important to avoid calculating **Jacobians** of the mapping T_s for acceptance probabilities in **Metropolis updates**.

Simulating Hamiltonian Dynamics

Leapfrog Method: (full step)

$$
\begin{aligned}
p_i(t + \epsilon/2) &= p_i(t) - (\epsilon/2)\frac{\partial \mathbf{H}}{\partial q_i}(q(t)) \\
q_i(t + \epsilon) &= q_i(t) + \epsilon \frac{\partial \mathbf{H}}{\partial p_i} p(t + \epsilon/2) \\
p_i(t + \epsilon) &= p_i(t + \epsilon/2) - (\epsilon/2)\frac{\partial \mathbf{H}}{\partial q_i}(q(t + \epsilon))
\end{aligned}
$$

Given suitable choice of $\mathbf{H}(q, p)$ the **leap-frog method**, **preserves volume**. (More on that later)
The method is **symmetric**, therefore **reversible** by simply negating p, the **momentum vector**.

Canonical Distributions

We can relate our target distribution $f(\mathbf{x}, \mathbf{u})$ to a **potential energy function**. Given **energy function** $E(\mathbf{x})$, for state $\mathbf{x}$ of some physical system, we can define a **canonical distribution** (**PDF**).

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

For our purposes we set,

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

setting $\mathbf{H}(q, p) = U(q) + K(p)$,

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

Canonical Distributions

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

Now we set $T = 1$, $U(q) = -log(f(\mathbf{x}))$ and choose a **kinetic energy function**, $K(p) = \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}/2$.

We get,

$$P(q,p) \propto f(\mathbf{x}) exp(-\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}/2)$$

crucially,

$$\int P(q,p) dp = \int f(\mathbf{x}) exp(-\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}/2) dx = f(\mathbf{x})$$

for appropriate normalization constants.

Hence, this construction is in line with the framework of **auxiliary variable methods**

The HMC Algorithm

$$P(q,p) \propto f(\mathbf{x}) exp(-\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}/2)$$

At time t, given $\mathbf{q}_t$,

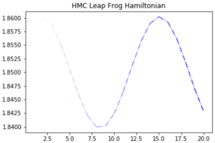
STEP 1:
Sample **momentum variables** from $N(0, \mathbf{M})$.
[By **independence**, $\mathbf{p}$ is drawn from its correct **conditional distribution**]

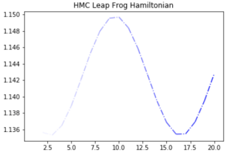
Now given $\mathbf{q}_t$ and $\mathbf{p}_t$,

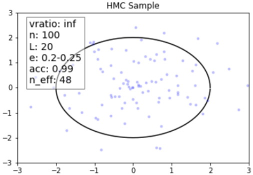
STEP 2:
Simulate L, ϵ-length **leap-frog steps** of the **Hamiltonian Dynamics**, to get proposed state $(\mathbf{q}^*, \mathbf{p}^*)$, and accept with probability,

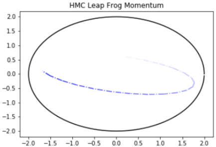
$$min\left[1, exp(-\mathbf{H}(\mathbf{q}^*, \mathbf{p}^*) + \mathbf{H}(\mathbf{q}_t, \mathbf{p}_t))\right] = min\left[1, exp(-U(\mathbf{q}^*) + U(\mathbf{q}_t)) - K(\mathbf{p}^*) + K(\mathbf{p}_t)\right]$$

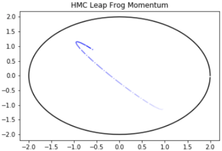
Example: 1



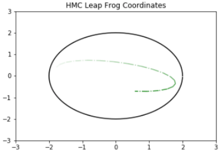


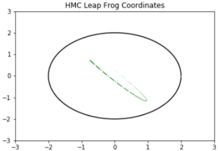




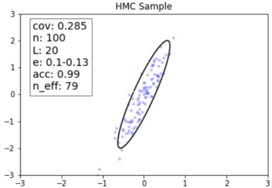


$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

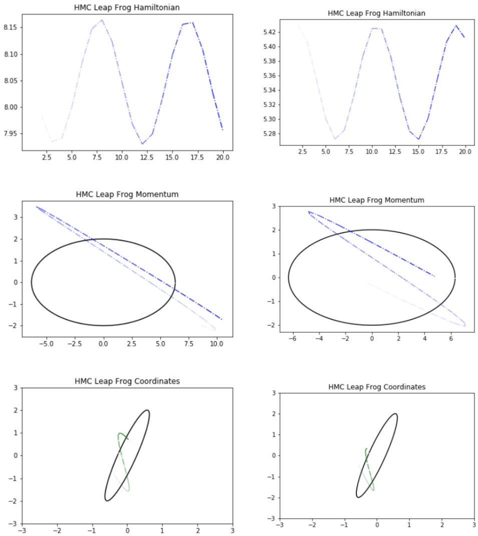




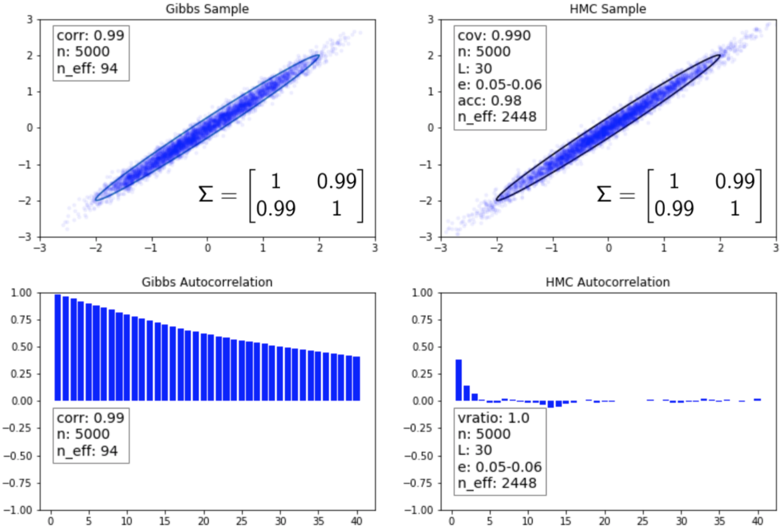
Example: 2



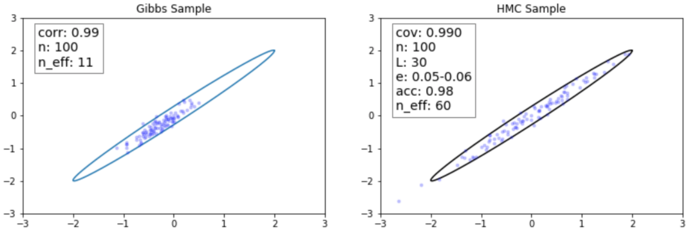
$$\Sigma = \begin{bmatrix} 0.1 & 0.285 \\ 0.285 & 1 \end{bmatrix}$$



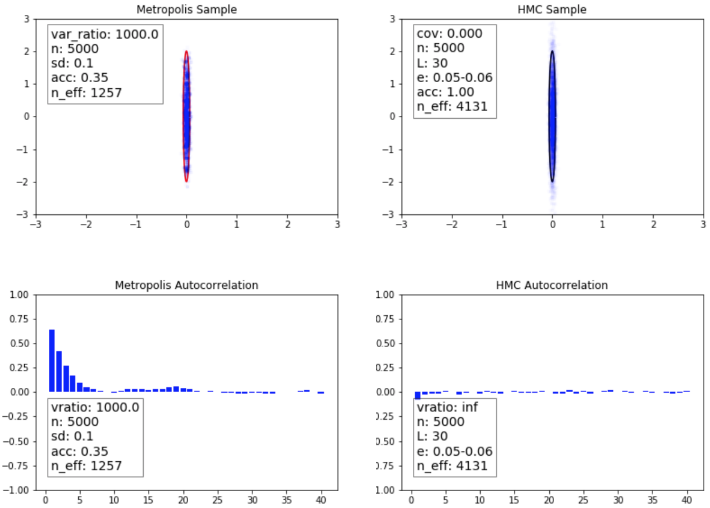
HMC VS. GIBBS


Gibbs Sample
corr: 0.99
n: 5000
n_eff: 94
$\Sigma = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$
HMC Sample
cov: 0.990
n: 5000
L: 30
e: 0.05-0.06
acc: 0.98
n_eff: 2448
$\Sigma = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$
Gibbs Autocorrelation
corr: 0.99
n: 5000
n_eff: 94
HMC Autocorrelation
vratio: 1.0
n: 5000
L: 30
e: 0.05-0.06
n_eff: 2448

HMC VS. GIBBS


Gibbs Sample
corr: 0.99
n: 100
n_eff: 11
HMC Sample
cov: 0.990
n: 100
L: 30
e: 0.05-0.06
acc: 0.98
n_eff: 60

HMC VS. METROPOLIS



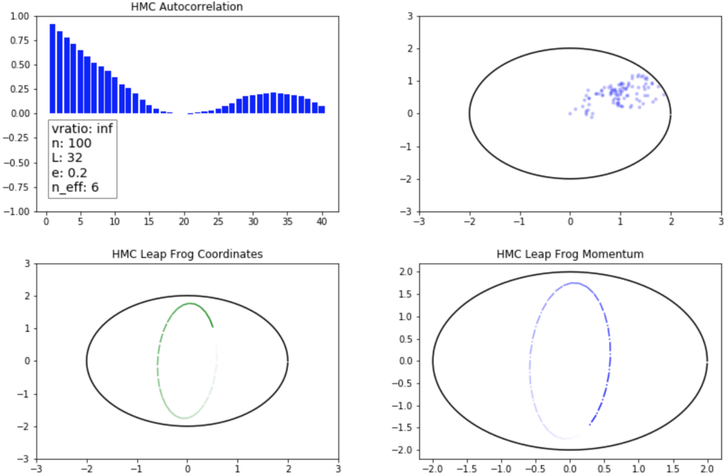
HMC: Concerns

The **performance** of **HMC** depends crucially on the choice of it's parameters, ϵ and L.

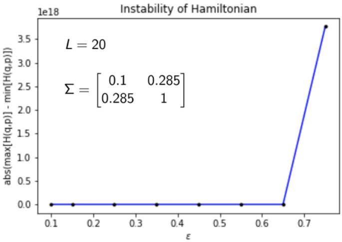
We want to achieve good **mixing**, (large movements in space for consecutive steps), while avoiding two pitfalls.

1. **Periodicity** in the Hamiltonian Dynamics.
2. **Instability** of the Hamiltonian

Tuning: Periodicity



Tuning: Instability of Hamiltonian



Tuning: Automatic Procedures

Tuning the **HMC** parameters is crucial because the sampler is very sensitive to the choice of L, and ϵ.

Automatic Procedures have been developed, of which the most widely used is the **No-U-Turn Sampler**.

The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, Matthew D. Hoffman, Andrew Gelman, *Journal of Machine Learning Research*, 15 (2014), Pages: 1351-1381

Idea, monitor **leapfrog steps** and interrupt progression if next step reduces distance to previous coordinate-position.

HMC: Limitations

The **HMC** sampler needs access to, and uses, the **gradients** of the **target distribution** at run-time.

- Not possible to use for **discrete distributions**. (Analytic tricks exists, but are not necessarily easy to handle)
- **Computational cost** of single iterations is **high** compared to Metropolis / Gibbs and other simpler samplers. [my own code is around 100 times slower than Gibbs and Metropolis]

Hence, for tractable problems for which standard samplers work reasonably well, they seem more advisable.

References and Code

REFERENCES

- The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, Matthew D. Hoffman, Andrew Gelman, *Journal of Machine Learning Research*, 15 (2014), Pages: 1351-1381
- Steve Brooks et. al. (2011). *Handbook of Markov Chain Monte Carlo*, CRC Press

CODE

askldksldjaksldj