On the geometric ergodicity of Hamiltonian Monte Carlo

Samuel Livingstone\textsuperscript{1*}, Michael Betancourt\textsuperscript{2}, Simon Byrne & Mark Girolami\textsuperscript{2}.

\textsuperscript{1}University of Bristol, \textsuperscript{2}University of Warwick.

January 6, 2016

*samuel.livingstone@bristol.ac.uk
Objectives

1) Establish some basic and general stability properties for Hamiltonian Monte Carlo
   - $\pi$-irreducibility
   - Geometric ergodicity

2) Infer useful guidelines for practitioners
   - Possible barriers to good performance
   - Integration time choice (motivation for heuristics)

Previous work of note.

A Hamiltonian function...

\[ H(x_t, p_t) = -\log \pi(x_t) + \frac{1}{2} p_t^T M^{-1} p_t \]

can be preserved using Hamilton’s equations...

If \( \frac{dx}{dt} = \frac{\partial H}{\partial p} \) and \( \frac{dp}{dt} = -\frac{\partial H}{\partial x} \) then \( \frac{dH}{dt} = 0. (*) \)

creating a measure-preserving flow

\[ H \circ \varphi_T(x_0, p_0) \approx H(x_0, p_0) \text{ for any } T \in \mathbb{R}. \]

where \( \varphi_T \) is a numerical solution to (*) - leapfrog integrator typical.
A Hamiltonian function...

\[ H(x_t, p_t) = -\log \pi(x_t) + \frac{1}{2} p_t^T M^{-1} p_t \]

can be preserved using Hamilton’s equations...

\[ \frac{dx}{dt} = \frac{\partial H}{\partial p} \quad \text{and} \quad \frac{dp}{dt} = -\frac{\partial H}{\partial x} \]

then \( \frac{dH}{dt} = 0. \) (*)

creating a measure-preserving flow

\[ H \circ \varphi_T(x_0, p_0) \approx H(x_0, p_0) \text{ for any } T \in \mathbb{R}. \]

where \( \varphi_T \) is a numerical solution to (*) - leapfrog integrator typical.

---

**Hamiltonian Monte Carlo**

**Require:** \( x_{i-1}, \epsilon \geq 0, L \in \mathbb{N} \)

Set \( x_0 \leftarrow x_{i-1}, \) draw \( p_0 \sim N(0, M), \) set \( T \leftarrow L\epsilon \)

Draw \( u \sim U[0, 1] \)

Set \( \delta \leftarrow H(x, p) - H \circ \varphi_T(x, p), \)

if \( \log(u) < \delta \) then

Set \( x_i \leftarrow \Pr_x \circ \varphi_T(x_0, p_0) \)

else

Set \( x_i \leftarrow x_{i-1} \)

end if
Possible $p_0$ values.

Figure 1: Phase space diagram for $H(x_t, p_t) = x_t^2/2 + p_t^2/2$.

Figure 2: Typical HMC proposal.
Stability of Markov chains

Measurable space $(X, \mathcal{B})$, $\pi(\cdot)$ a distribution, $P : X \times \mathcal{B} \to [0, 1]$ a transition kernel.

\textbf{π-irreducibility}

For every $A \in \mathcal{B}$ with $\pi(A) > 0$ and every $x \in X$, there is an $n = n(x, A)$ such that

$$P^n(x, A) > 0.$$

\textbf{Geometric ergodicity}

For some $\rho < 1$, $C < \infty$ and $\pi$-a.e. finite $V : X \to [1, \infty]$

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq CV(x)\rho^n$$
Intuition for GeoErg: for large $|x_0|$ how does the chain behave?

- Starting point $x_0$

How long until we reach the typical set?
And can we get back into the tails again?

Typical set

Idea formalised using Lyapunov functions: $\int V(y)P(x_0,dy) < \lambda V(x_0) + b\mathbb{1}_C(x_0)$. 
Viewing HMC *marginally* will be useful to us

The standard leapfrog integrator with $T = L\varepsilon$...

\[
\begin{align*}
\begin{aligned}
\dot{p}_{(i+1/2)} & = \dot{p}_i + \varepsilon \nabla \log \pi(x_i) / 2 \\
\dot{x}_{(i+1)} & = \dot{x}_i + \varepsilon \dot{p}_{(i+1/2)} \\
\dot{p}_{(i+1)} & = \dot{p}_{(i+1/2)} + \varepsilon \nabla \log \pi(x_{(i+1)}) / 2
\end{aligned}
\end{align*}
\]

Repeat $L$ times

... can be compressed into

\[
x_{L\varepsilon} = x_0 + \frac{L\varepsilon^2}{2} \nabla \log \pi(x_0) + \varepsilon^2 \sum_{i=1}^{L-1} (L - i) \nabla \log \pi(x_i) + L\varepsilon \dot{p}_0.
\]

(Possibly non-linear) function of $x_0, p_0$
Why is $\pi$-irreducibility difficult?

Take $\pi(x) \propto \exp\left(-\frac{x^2}{2}\right)$, meaning $\nabla \log \pi(x) = -x$.

Set $L = 2$, giving

$$x_{t+2\epsilon} = x_t - \epsilon^2 x_t - \epsilon^2 (x_t - \epsilon^2 x_t + \epsilon p_t) + 2\epsilon p_t,$$

$$= x_t - \epsilon^2 x_t - \epsilon^2 x_t + \epsilon^4 x_t - \epsilon^3 p_t + 2\epsilon p_t,$$

$$= (1 - 2\epsilon^2 + \epsilon^4) x_t + (2\epsilon - \epsilon^3) p_t.$$  

Set $\epsilon = \sqrt{2}$. Then $2\epsilon - \epsilon^3 = 0$, giving

$$x_{t+2\epsilon} = (1 - 4 + 4) x_t = x_t.$$
Why is $\pi$-irreducibility difficult?

Take $\pi(x) \propto \exp \left( -\frac{x^2}{2} \right)$, meaning $\nabla \log \pi(x) = -x$.

Set $L = 2$, giving

$$x_{t+2\varepsilon} = x_t - \varepsilon^2 x_t - \varepsilon^2 (x_t - \varepsilon^2 x_t + \varepsilon p_t) + 2\varepsilon p_t,$$

$$= x_t - \varepsilon^2 x_t - \varepsilon^2 x_t + \varepsilon^4 x_t - \varepsilon^3 p_t + 2\varepsilon p_t,$$

$$= (1 - 2\varepsilon^2 + \varepsilon^4) x_t + (2\varepsilon - \varepsilon^3) p_t.$$

Set $\varepsilon = \sqrt{2}$. Then $2\varepsilon - \varepsilon^3 = 0$, giving

$$x_{t+2\varepsilon} = (1 - 4 + 4) x_t = x_t.$$

The chain does not move $\Rightarrow$ not $\pi$-irreducible!
\( \pi \)-irreducibility (static integration time)

**Assumptions**

(A1) (Regularity) \( \nabla \log \pi \in C(X) \), \( \pi(x) \) bounded away from 0 and \( \infty \) on compact sets

(A2) (Gradient growth)

\[
\limsup_{|x| \to \infty} \frac{|\nabla \log \pi(x)|}{|x|} < C_{L,\epsilon} \text{ for some } C_{L,\epsilon} < \infty.
\]

**Proposition 1.** Under the above assumptions, HMC produces a \( \pi \)-irreducible Markov chain, and all compact sets are small.

Intuition for result (in 1D)...\[
x_{L\epsilon}(p_0) = x_0 + \frac{L\epsilon^2}{2} \nabla \log \pi(x_0) + \epsilon^2 \sum_{i=1}^{L-1} (L - i) \nabla \log \pi(x_{i\epsilon}) + L\epsilon p_0.
\]

Leading order term

- As \( p_0 \to \infty \), \( x_{L\epsilon}(p_0) \to \infty \), similarly for \( p_0 \to -\infty \)
- \( x_{L\epsilon}(p_0) \) continuous so Intermediate Value Theorem \( \Rightarrow \) surjective.
Geometric ergodicity (static integration time)

Further assumptions

(A3) (Direction of gradient)

\[
\limsup_{|x| \to \infty} \langle \nabla \log \pi(x), x \rangle < 0 
\]

(A4) (Inwards convergence)

\[
\lim_{|x| \to \infty} \int_{A(x) \setminus I(x)} Q(x, dy) = 0, 
\]

where \( A \triangle B = (A \cap B^c) \cup (R(x) \cap I(x)) \) is the symmetric difference of sets \( A \) and \( B \).


Theorem 3. Under (A1), HMC does not produce a geometrically ergodic Markov chain if either of the following hold

(i) (Vanishing gradient) \( |\nabla \log \pi(x)| \to 0 \) as \( |x| \to \infty \)

(ii) (Exploding gradient)

\[
\liminf_{|x| \to \infty} \frac{|\nabla \log \pi(x)|}{|x|} = \infty,
\]
Intuition for static results

Simplified ‘Exponential family’.

\[ \pi(x) \propto \exp \left( -\frac{1}{\beta} |x|^\beta \right) \]

for some \( \beta > 0 \).

\[ \nabla \log \pi(x) = -\text{sgn}(x) |x|^{\beta-1} \]

So proposal is

\[ x_{Le} = x_0 + \text{DRIFT} + L\epsilon p_0. \]

where

\[ |\text{DRIFT}| = O \left( \max \left( |p_0|^{\beta-1}, |x_0|^{\beta-1}, |x_\epsilon|^{\beta-1}, \ldots, |x_{(L-1)\epsilon}|^{\beta-1} \right) \right) \]
Intuition for static results

Simplified ‘Exponential family’

\[ \pi(x) \propto \exp \left( -\frac{1}{\beta} |x|^{\beta} \right), \]

for some \( \beta > 0 \).

\[ \nabla \log \pi(x) = -\text{sgn}(x) |x|^{\beta-1} \]

So proposal is

\[ x_{L\varepsilon} = x_0 + \text{DRIFT} + L\varepsilon p_0. \]

where

\[ |\text{DRIFT}| = O \left( \max \left( |p_0|^{\beta-1}, |x_0|^{\beta-1}, |x_\varepsilon|^{\beta-1}, \ldots, |x_{(L-1)\varepsilon}|^{\beta-1} \right) \right) \]

Take \( |p_0| \leq |x_0|^\delta \) for some \( 0 < \delta < 1 \), and \( |x_0| \) large.

- \( \beta < 1 \Rightarrow |\text{DRIFT}| \) typically small, \( |x_{L\varepsilon}| \approx |x_0| \). Not GeoErg.
- \( 1 \leq \beta \leq 2 \Rightarrow |\text{DRIFT}| \) is sub-linear, \( |x_{L\varepsilon}| < |x_0| \). GeoErg.
- \( \beta > 2 \Rightarrow |\text{DRIFT}| \) is super-linear, \( |x_{L\varepsilon}| \gg |x_0| \). Not GeoErg.
Geometric ergodicity (dynamic integration time)

Is HMC therefore the same as MALA? Not necessarily!

- MALA is HMC with $L = 1$
- $L = L(x_t, p_t)$ possible - Stan implementation does this

**Idealised Scenario**

- Take 1-dimensional targets $\pi(x) \propto \exp\left(-\frac{1}{\beta}|x|^\beta\right)$ for $\beta > 0$.
  $\implies$ flow periodic with period length $\eta_{x,p}$
- Assume exact integrator available
- Choose integration time as $T_{x,p} \sim U[0, \max(s, \eta_{x,p})]$, for some small $s > 0$
Geometric ergodicity (dynamic integration time)

Is HMC therefore the same as MALA? Not necessarily!

- MALA is HMC with $L = 1$
- $L = L(x_t, p_t)$ possible - **Stan implementation does this**

### Idealised Scenario

- Take 1-dimensional targets $\pi(x) \propto \exp\left(-\frac{1}{\beta}|x|^\beta\right)$ for $\beta > 0$.
  
  \[\implies \text{flow periodic} \text{ with period length } \eta_{x,p}\]

- Assume exact integrator available

- Choose integration time as $T_{x,p} \sim U[0, \max(s, \eta_{x,p})]$, for some small $s > 0$

**Theorem 4.** In the Idealised Scenario setting then HMC produces a geometrically ergodic Markov chain for any $0 < \beta < \infty$.

**Remark.** With the leapfrog integrator Theorem 4 will not hold for $\beta > 2$ but will for $\beta < 1$. 
Summary and future directions

- Established some conditions for stability of HMC
- Offered some insight into integration time choice

Open problems

- Incorporating computing time
- Direct algorithm comparison
- Results for other HMC variants