Geometry and ergodicity of Hamiltonian Monte Carlo

Simon Byrne
with Michael Betancourt and Sam Livingston

Department of Statistical Science
University College London

24th September 2015
Bayesian computational problem

We want to compute the posterior distribution of some quantity $x$ conditional on having observed $y$:

$$\Pi(dx) = \frac{p(y \mid x) \omega(dx)}{\int_X p(y \mid x') \omega(dx')}$$

- High-dimensional, complicated distributions.
- Cannot be integrated or sampled from directly.
- Normalisation constant is unknown.

Markov chain Monte Carlo (MCMC)

We construct a Markov chain $T(d\theta_{i+1} \mid \theta_i)$ whose invariant distribution is the posterior.
MCMC: Exploration vs. Acceptance

There is an inherent tradeoff:

**Small proposals**
⇒ highly correlated samples
⇒ poor estimators.

**Large proposals**
⇒ low acceptance rates
⇒ highly correlated samples
⇒ poor estimators.
Hamiltonian/Hybrid Monte Carlo

- Let $\mathcal{X}$ be a Riemannian manifold with metric $M$.
- **Auxiliary momentum variable** $p \in T^*_x$, with $p \mid x \sim \mathcal{N}(0, M(x))$ (Girolami and Calderhead 2011).
- The **Hamiltonian** is the negative log-density of the joint distribution
  \[ H(x, p) = -\log \pi(x) + \frac{1}{2} \log |M(x)| + \frac{1}{2} p^\top M(x)^{-1} p \]
- Defines a dynamical system (**Hamilton’s equations**):
  \[ \frac{dx}{dt}(t) = \frac{\partial H}{\partial p} = M^{-1}(x)p \quad \text{and} \quad \frac{dp}{dt}(t) = -\frac{\partial H}{\partial x} \]
- Unaffected by normalisation constants
- **Properties:**
  - $H$ constant over $t$.
  - Reversible by negation of $p$.
  - $(x(0), p(0)) \mapsto (x(t), p(t))$ has unit Jacobian determinant.
Leapfrog integrator

We can’t solve the system exactly, but we can approximate it using a leapfrog integrator, such that

- Reversible
- Unit Jacobian
- $H$ approximately preserved.
Sample $p^{(n)} \sim \mathcal{N}(0, M(x^{(n)}))$.

2. From $(x^{(n)}, p^{(n)})$, simulate $L$ leapfrog steps to obtain $(x^*, p^*)$.

3. Compute $\alpha = \exp\{-H(x^*, p^*) + H(x^{(n)}, p^{(n)})\}$.

4. Set $x^{(n+1)} = x^*$ with probability $\min(\alpha, 1)$, otherwise $x^{(n+1)} = x^{(n)}$.

- Can make large proposal moves, with high acceptance, yet low autocorrelation.
- General purpose: only requires derivatives.
- Can be largely automated (e.g. Stan library).
Cotangent bundle is a symplectic manifold.
- $H$ is negative log-density w.r.t. canonical symplectic measure.
- $(x(0), p(0)) \mapsto (x(t), p(t))$ is a symplectic map.

Choice of metric $M$:
- Constant/Euclidean is easiest to work with.
- Fisher–Rao typically intractable, and lacks invariance justification (known data, influence of prior).
- Observed information often requires modification to be positive definite (Betancourt 2013).

Geometric numerical integration (Hairer, Lubich, and Wanner 2006)
- Approximate integrator is actually exact solution to a different Hamiltonian $\tilde{H}$.

$$\tilde{H}(x, p) = H(x, p) + \epsilon^2 G(x, p) + \ldots$$

- Backward error analysis.
- Optimal tuning of $\epsilon$ (Betancourt, B., and Girolami 2014).
MCMC estimators

Samples are typically used to obtain estimates of moments of interest:

\[ \hat{f}_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \approx \mathbb{E}_\Pi[f] \]

- Samples are not independent: usual central limit theorem does not apply.
Geometric ergodicity

Total variation distance between two probability measures

\[ \| \mu - \nu \|_{TV} = \sup_{A \in \Omega} \| \mu(A) - \nu(A) \| \]

A Markov chain is **geometrically ergodic** if the \( n \)-step transition distance decays geometrically

\[ \| T^n(\cdot | x) - \Pi \|_{TV} \leq V(x) \rho^n, \quad \rho < 1 \]
If a Markov chain is geometrically ergodic, then subject to some moment conditions on $f$, there is a central limit theorem

$$\sqrt{n}\left(\hat{f}_n - \mathbb{E}_\Pi[f]\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

for some $\sigma^2 < \infty$.

- Justifies use of an MCMC approximation to the posterior.
- Known conditions for simple algorithms (Gibbs, RW, MALA).
- Difficult to establish for more complicated schemes, such as HMC.
Establishing geometric ergodicity

**Minorisation condition**

There exists a *small set* \( C \), integer \( n \) and \( \epsilon > 0 \) and probability measure \( \nu \) such that

\[
T^n(\cdot \mid x) \geq \epsilon \nu(\cdot)
\]

for all \( x \in C \).

**Drift condition**

There exist a *drift function* \( V : \mathcal{X} \to [1, \infty] \), and constants \( 0 < \lambda < 1, \ b < \infty \) such that

\[
\mathbb{E}_T[V \mid x] \leq \lambda V(x) + b1_C(x)
\]

where \( C \) is the small set.

If both conditions are satisfied then \( T \) is geometrically ergodic.
\[ \pi(x) \propto \exp\{-|x|^\beta\} \]

Then with standard HMC with Euclidean metric:

<table>
<thead>
<tr>
<th>Tails</th>
<th>(\beta)</th>
<th>Geom. Ergod.?</th>
</tr>
</thead>
<tbody>
<tr>
<td>heavy</td>
<td>(0, 1)</td>
<td>\xmark</td>
</tr>
<tr>
<td>intermediate</td>
<td>[1, 2]</td>
<td>\checkmark*</td>
</tr>
<tr>
<td>light</td>
<td>(2, (\infty))</td>
<td>\xmark</td>
</tr>
</tbody>
</table>

- Same as MALA algorithm (Roberts and Tweedie 1996): arises as a special case when \(L = 1\).

So why does it fail?
The problem is that it takes too long to come in from the tails:

- Lose in gradient information

Method behaves like a random walk, which performs poorly for heavy-tailed distributions (Jarner and Tweedie 2003).
Suppose that:

1. We can solve Hamilton’s equations exactly.
   - The paths are the contours of $H$, and therefore there exists $t_{rec}$ such that $(x(t_{rec}), p(t_{rec})) = (x(0), p(0))$.

2. We can sample integration time $I$ uniformly $[0, t_{rec})$. 

Diagram: A set of nested curves in the $x-p$ plane, indicating the contours of a function $H$. The curves are evenly spaced, suggesting a uniform sampling of the integration time $I$. The origin of the coordinates is marked, with axes labeled $x$ and $p$.
Virial theorem

**Theorem**

*In a Hamiltonian system with constant metric $M$,*

$$
\mathbb{E}_I \left[ p^\top M p \right] = \mathbb{E}_I \left[ x^\top \nabla_x \log \pi(x) \right]
$$

The *Virial* is the quantity

$$G = x^\top p$$

Then

$$
\frac{dG}{dt} = \frac{dx}{dt}^\top p + x^\top \frac{dp}{dt} = p^\top Mp - x^\top \nabla_x \log \pi(x)
$$

and hence

$$
\mathbb{E}_I \left[ p^\top M p \right] - \mathbb{E}_I \left[ x^\top \nabla_x \log \pi(x) \right] = \mathbb{E}_I \left[ \frac{dG}{dt} \right] = \frac{1}{t_{\text{rec}}} [G(t_{\text{rec}}) - G(0)] = 0.
$$
### Geometric ergodicity of HMC

**Theorem**

Suppose that for some $B > 0$,

$$\log \pi(x) \leq A + B x^\top \nabla_x \log \pi(x)$$

then the constant-metric Hamiltonian scheme

1. Sample $p(0) \sim \mathcal{N}(0, M)$,
2. Sample $t \sim I$, $(x(0), p(0)) \mapsto (x(t), p(t))$.

is geometrically ergodic.

$$H = -\log \pi(x) + \frac{1}{2} p^\top Mp$$

Rough idea: drift function $V = C - \log \pi(x)$.

1. “resets” the second term
2. “equilibrates” the two terms
Generalisations

Of course we can’t do this in typical practice, but

- Poincaré recurrence theorem says that we will *approximately* recur in finite time.

- No-U-turn sampler (NUTS, Hoffman and Gelman 2014) provide a framework for adaptively choosing the integration.
  - Can be adapted to utilise approximate Virial criterion.
Let $M$ be the Hessian of the log-density

$$M(x) = |\nabla^2 \log \pi(x)| \propto \beta^2 |x|^\beta - 2$$

This manifold has an isometric embedding into Euclidean space

$$x' = \text{sign}(x)|x|^\beta/2, \quad p' = \beta|q|^\beta/2 - 1 p$$

In other words, this is equivalent to Euclidean HMC with the target density

$$\pi'(x') \propto |x'|^{2/\beta - 1} \exp(-x'^2/\beta^2)$$

- The tails are dominated by the Gaussian term: geometrically ergodic?
Summary

- HMC is a deeply geometric algorithm.
- Geometry provides valuable insight, where other tools fail.
- Provided genuine qualitative improvements.

Lots of open questions:

- Incorporate effects of discretisation.
- Extend results to more general distributions.
- Quantitative results (e.g. geometric rate $\rho$).
- Diagnostics of Markov chain convergence.


