A General Metric for Riemannian Hamiltonian Monte Carlo

Michael Betancourt
University College London
August 30th, 2013
I’m going to talk about probability and geometry, but *not* information geometry!
Instead our interest is Bayesian inference

\[ \pi(\theta|D) \propto \pi(D|\theta) \pi(\theta) \]
Markov Chain Monte Carlo admits the practical analysis and manipulation of posteriors even in high dimensions.
Markov transitions can be though of as an “average” isomorphism that preserves a target distribution

\((\Omega, \mathcal{B}(\Omega), \pi)\)
Markov transitions can be though of as an “average” isomorphism that preserves a target distribution

$$(\Omega, \mathcal{B}(\Omega), \pi)$$

$$(\Gamma, \mathcal{B}(\Gamma), \gamma)$$

$t : \Omega \rightarrow \Omega, \forall t \in \Gamma$
Markov transitions can be though of as an “average” isomorphism that preserves a target distribution

\[(\Omega, \mathcal{B}(\Omega), \pi)\]

\[(\Gamma, \mathcal{B}(\Gamma), \gamma)\]

\[t : \Omega \to \Omega, \; \forall t \in \Gamma\]

\[\pi T = \pi\]
Random Walk Metropolis and the Gibbs sampler have been the workhorse Markov transitions
Random Walk Metropolis and the Gibbs sampler have been the workhorse Markov transitions

\[
T(\theta, \theta') = \mathcal{N}(\theta' | \theta, \sigma^2) \min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right)
\]
Random Walk Metropolis and the Gibbs sampler have been the workhorse Markov transitions

\[ T(\theta, \theta') = \mathcal{N}(\theta' | \theta, \sigma^2) \min \left( 1, \frac{\pi(\theta')}{\pi(\theta)} \right) \]

\[ t : \theta \rightarrow \theta + a \cdot \epsilon \]

\[ \epsilon \sim \mathcal{N}(0, \sigma^2) \]

\[ a|\epsilon \sim \text{Ber} \left( \min \left( 1, \frac{\pi(\theta + \epsilon)}{\pi(\theta)} \right) \right) \]
Random Walk Metropolis and the Gibbs sampler have been the workhorse Markov transitions

\[ T(\theta, \theta') = \prod_{i} \pi(\theta'_i | \theta_j \setminus i) \]
Random Walk Metropolis and the Gibbs sampler have been the workhorse Markov transitions

\[ T(\theta, \theta') = \prod_i \pi(\theta'_i | \theta_{\cdot \setminus i}) \]

\[ t_i : \theta_i \rightarrow \epsilon \]

\[ \epsilon \sim \pi(\epsilon | \theta_{\cdot \setminus i}) \]
MCMC performance is limited by complex posteriors, which are common in large dimensions.
Random walk Metropolis sampling explores only slowly
Random walk Metropolis sampling explores only slowly
Gibbs sampling doesn’t fare much better
Gibbs sampling doesn’t fare much better
RWM and Gibbs explore incoherently in large dimensions

\[ t : \theta \rightarrow \theta + a \cdot \epsilon \]

\[ \epsilon \sim \mathcal{N}(0, \sigma^2) \]

\[ a|\epsilon \sim \text{Ber}\left( \min\left( 1, \frac{\pi(\theta + \epsilon)}{\pi(\theta)} \right) \right) \]

\[ t_i : \theta_i \rightarrow \epsilon \]

\[ \epsilon \sim \pi(\epsilon|\theta_{j \setminus i}) \]
How do we generate coherent transitions?

$$(\Omega, \mathcal{B}(\Omega), \pi)$$

$$(\Gamma, \mathcal{B}(\Gamma), \gamma)$$

$t : \Omega \to \Omega$, $\forall t \in \Gamma$

$$\pi T = \pi$$
How do we generate coherent transitions?

$$(\mathcal{M}, \mathcal{B}(\mathcal{M}), \pi)$$

$$(\Gamma, \mathcal{B}(\Gamma), \gamma)$$

$t : \mathcal{M} \to \mathcal{M}, \ \forall t \in \Gamma$

$$\pi^T = \pi$$
Hamiltonian flow is a coherent, measure-preserving map

\[ T : \mathcal{M} \rightarrow T^{*}\mathcal{M} \rightarrow T^{*}\mathcal{M} \rightarrow \mathcal{M} \]
Hamiltonian flow is a coherent, measure-preserving map

\[ T : \mathcal{M} \rightarrow T^* \mathcal{M} \rightarrow T^* \mathcal{M} \rightarrow \mathcal{M} \]

Random Lift
Hamiltonian flow is a coherent, measure-preserving map

\[ T : \mathcal{M} \rightarrow T^*\mathcal{M} \rightarrow T^*\mathcal{M} \rightarrow \mathcal{M} \]

Random Lift

Hamiltonian Flow
Hamiltonian flow is a coherent, measure-preserving map

\[
T : \mathcal{M} \rightarrow T^*\mathcal{M} \rightarrow T^*\mathcal{M} \rightarrow \mathcal{M}
\]

Random Lift

Hamiltonian Flow

Marginalization
We just need to define a lift from the sample space to its cotangent bundle

\[ \pi(q) \rightarrow \pi(p|q) \pi(q) \]
We just need to define a lift from the sample space to its cotangent bundle

\[ \pi(q) \rightarrow \pi(p|q) \pi(q) \]

\[ H = - \log \pi(p|q) - \log \pi(q) \]
We just need to define a lift from the sample space to its cotangent bundle

\[ \pi(q) \rightarrow \pi(p|q) \pi(q) \]

\[ H = -\log \pi(p|q) - \log \pi(q) \]

\[ T \]
We just need to define a lift from the sample space to its cotangent bundle

\[ \pi(q) \rightarrow \pi(p|q) \pi(q) \]

\[ H = - \log \pi(p|q) - \log \pi(q) \]

\[ V \]
Quadratic kinetic energies with constant metrics emulate dynamics on a Euclidean manifold

\[ \pi(p|q) = \mathcal{N}(0, M) \]

\[ T = \frac{1}{2} p_i p_j (M^{-1})^{ij} \]
The coherent flow the Markov chain along the target distribution, avoiding random walk behavior
The coherent flow the Markov chain along the target distribution, avoiding random walk behavior
Unfortunately, EHMC is sensitive to large variations in curvature.
As well as variations in the target density

$$\Delta V = \Delta T = \frac{d}{2}$$
These weaknesses are particularly evident in hierarchical models

\[ \pi(x, v) = \prod_{i=1}^{n} \pi(x_i | v) \pi(v) \]
These weaknesses are particularly evident in hierarchical models
These weaknesses are particularly evident in hierarchical models.
These weaknesses are particularly evident in hierarchical models.
These weaknesses are particularly evident in hierarchical models.
Quadratic kinetic energies with dynamic metrics emulate dynamics on a Riemannian manifold

\[
\pi(p|q) = \mathcal{N}(0, \Sigma(q))
\]

\[
T = \frac{1}{2} p_i p_j \left( \Sigma^{-1}(q) \right)^{ij} + \frac{1}{2} \log |\Sigma(q)|
\]
Optimal numerical integration suggests using the Hessian, but the Hessian isn’t positive-definite

$$
\Sigma(q)_{ij} \neq \partial_i \partial_j V(q)
$$
Fisher-Rao is both impractical and ineffective

\[ \Sigma(q)_{ij} = \mathbb{E}_D \left[ \partial_i \partial_j V(q|\mathcal{D}) \right] \]
Fisher-Rao is both impractical and ineffective

$$\Sigma(q)_{ij} = E_D [\partial_i \partial_j V(q|D)]$$
Fisher-Rao is both impractical and ineffective

\[ \Sigma(q)_{ij} = \mathbb{E}_D \left[ \partial_i \partial_j V(q | D) \right] \]
We can regularize without appealing to expectations

$$\Sigma_{ij}(q) = \left[ \exp(\alpha H_{ik}) + \exp(-\alpha H_{ik}) \right]$$

$$\cdot H_{kl} \cdot$$

$$\left[ \exp(\alpha H_{lj}) - \exp(-\alpha H_{lj}) \right]^{-1}$$
The “SoftAbs” metric serves as a differentiable absolute value of the Hessian.
The SoftAbs metric locally standardizes the target distribution
The SoftAbs metric locally standardizes the target distribution.
And the log determinant admits full exploration of the funnel
And the log determinant admits full exploration of the funnel
The SoftAbs metric admits a general-purpose, practical implementation of RHMC.