Wormhole Hamiltonian Monte Carlo

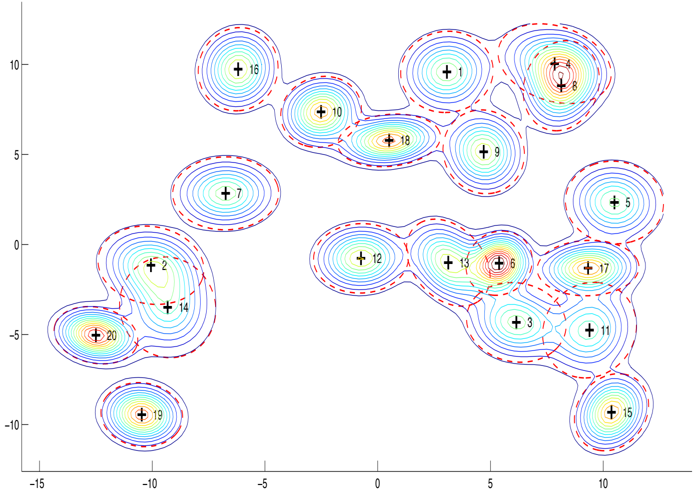
Babak Shahbaba[1]

Department of Statistics and Department of Computer Science, UCI

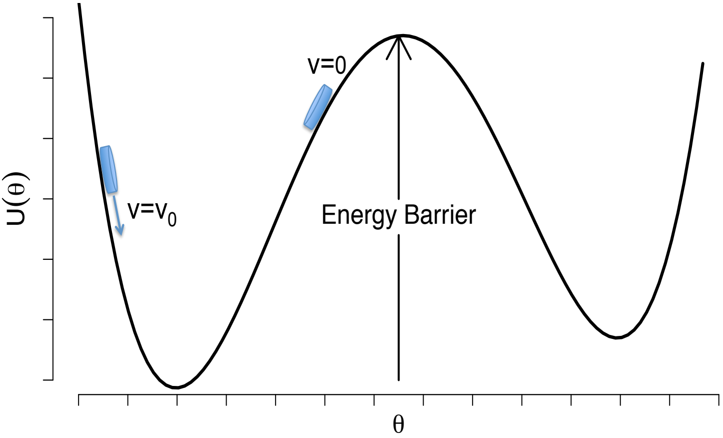
AAAI, 2014

[1] Joint work with Shiwei Lan and Jeffrey Streets

Motivation



Energy Barrier



Our Solution:
Exploiting and Modifying Geometry

Hamiltonian Monte Carlo (HMC)

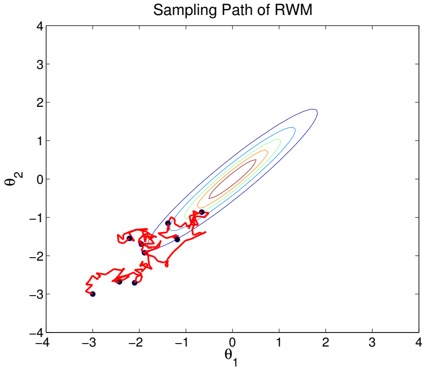
The Metropolis Algorithm

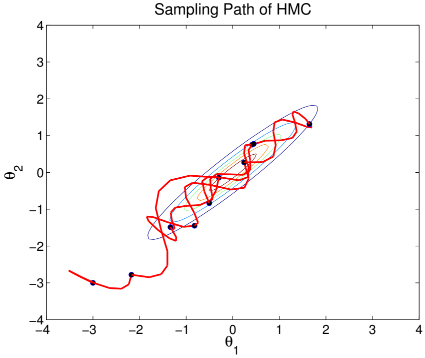
- Specify a symmetric transition probability $g(\theta, \theta^*)$ and repeat the following steps for many iterations:
  1. Given our current state $\theta^{(n)}$, we propose a new state θ^* according to g.
  2. Calculated the acceptance probability,

$$a(\theta^{(n)}, \theta^*) = \min(1, \frac{f(\theta^*)}{f(\theta^{(n)})})$$

  3. Accept the proposed state $\theta^{(n+1)} = \theta^*$ as the new state with probability $a(\theta^{(n)}, \theta^*)$ or remain at the current state $\theta^{(n+1)} = \theta^{(n)}$.

- Hamiltonian Monte Carlo (HMC) reduces the random walk behavior of Metropolis.





Hamiltonian Dynamics

- The dynamic system can be represented by the *Hamiltonian* function:

$$H(\theta, p) = U(\theta) + K(p)$$

- *Hamilton's equations* determine how θ and p change over time:

$$\begin{aligned} \dot{\theta} &= \nabla_p H(\theta, p) \\ \dot{p} &= -\nabla_\theta H(\theta, p) \end{aligned}$$

- They define a mapping, T_s, from the state at some time t to the state at time $t + s$.

Application in Statistics

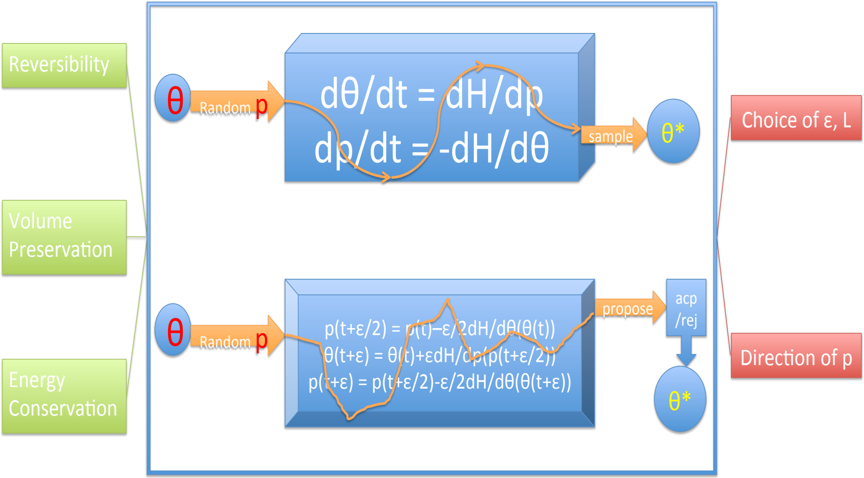
- In statistics, the potential energy $U(\theta)$ is the minus log density of the target distribution (e.g., posterior distribution).
- We also introduce fictitious momentum variables p.
- Typically, we set $p \sim N(0, M)$ so

$$K(p) = \sum_i p_i^2 / 2m_i$$

  where M is called the *mass matrix* and is usually set to I.
- The joint density of θ and p is

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

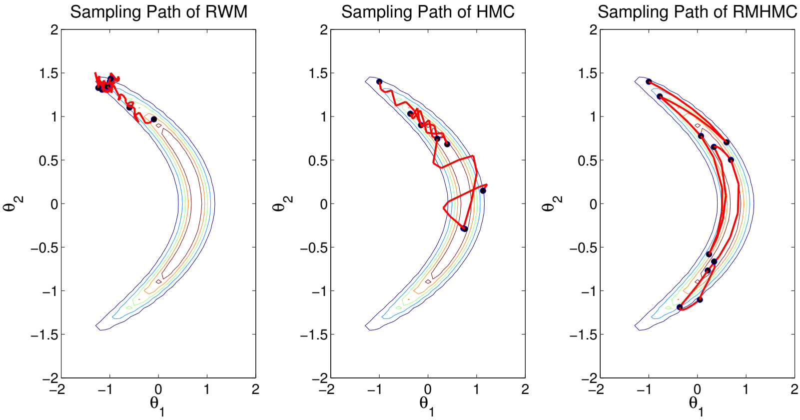
HMC Overview



Riemannian Manifold HMC

RMHMC

- Girolami and Calderhead (2011) have introduced a new method, called Riemannian Manifold HMC (RMHMC).
- They argue that it is more natural to put the Hamiltonian dynamic on Riemannian manifold of distributions rather than Euclidean space.
- They follow Amari (2000) and use the Fisher information matrix, $G(\theta) = -E\left[\nabla_\theta^2 \log f(\theta)\right]$, as a metric on the manifold.
- That is, they use position specific mass matrix, $M = G(\theta)$
- This way, we could explore the parameter space more efficiently by exploiting its geometric properties.


Sampling Path of RWM
Sampling Path of HMC
Sampling Path of RMHMC
θ1
θ2
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2
−1
0
1
2

Wormhole HMC

Known Modes

Geodesic

- For a manifold, $\mathcal{M}$, endowed with a generic metric $G(\theta)$, we define the arclength along the curve $\theta(t) : [0, T] \to \mathcal{M}$ as

$$\ell(\theta) := \int_0^T \sqrt{\dot{\theta}(t)^\mathsf{T} G(\theta(t)) \dot{\theta}(t)} dt$$

- Given any two points $\theta_1, \theta_2 \in \mathcal{M}$ there exists a curve

$$\theta(t) : [0, T] \to \mathcal{M}$$

satisfying the boundary conditions

$$\theta(0) = \theta_1, \theta(T) = \theta_2$$

whose arclength is minimal among such curves.

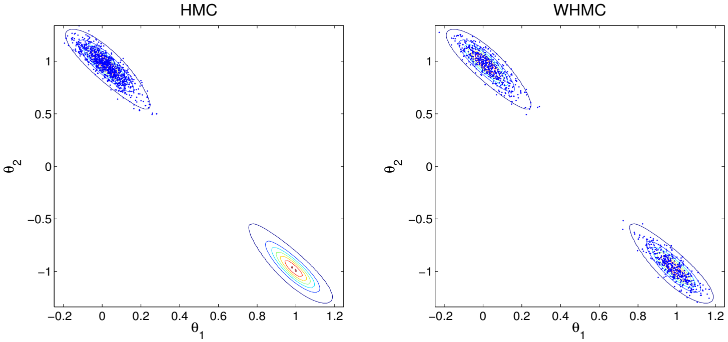
- The length of such a minimal curve defines a distance function on $\mathcal{M}$.

Wormhole Metric

- We start by assuming that the modes are known (possibly through some fast optimization methods).
- We define a new metric, G_W, for which the distance between modes is shortened.
- This way, we can facilitate moving between modes by creating "wormholes" between them.
- Next, we define the overall metric, G, for the whole parameter space of θ as a weighted sum of the base metric G_0 and the wormhole metric G_W,

$$G(\theta) = (1 - \mathfrak{m}(\theta))G_0(\theta) + \mathfrak{m}(\theta)G_W,$$

Illustration: 2d MoG with tied means



$$\theta_d \sim \mathcal{N}(\theta_d, \sigma_d^2), \quad d = 1, 2.$$

$$x_i \sim \frac{1}{2}\mathcal{N}(\theta_1, \sigma_x^2) + \frac{1}{2}\mathcal{N}(\theta_1 + \theta_2, \sigma_x^2)$$

Wormhole HMC

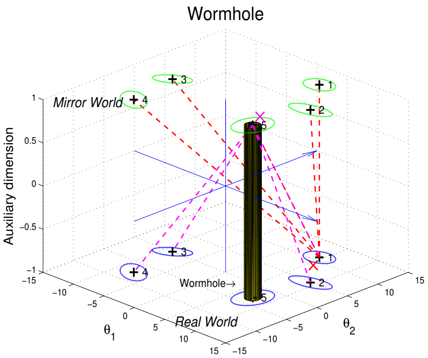
Effect of wormhole metric diminishes

$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$

$$\boxed{\textit{External force}} \Bigg\Downarrow + \mathbf{f}(\boldsymbol{\theta}, \mathbf{v})$$

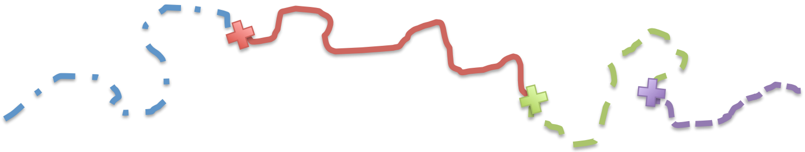
$$\dot{\boldsymbol{\theta}} = \mathbf{v} + \mathbf{f}(\boldsymbol{\theta}, \mathbf{v})$$

Wormholes interfere with each other



Wormhole HMC
Unknown Modes

Regeration

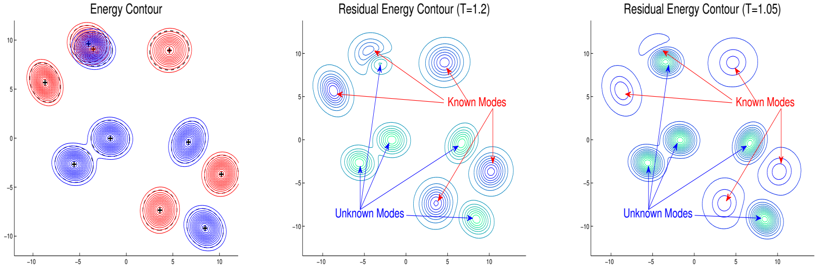


- Our solution: search for new modes and update wormhole network at *regeneration times*
- The transition kernel is regarded as a mixture of two kernels

$$\mathcal{T}(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) = S(\boldsymbol{\theta}_t)Q(\boldsymbol{\theta}_{t+1}) + (1 - S(\boldsymbol{\theta}_t))R(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)$$

- Regeneration: a state is deemed to come from an independence kernel *retrospectively*

Mode Searching

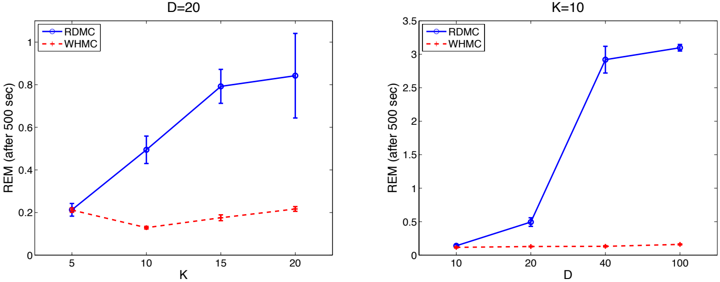


- At regeneration times, proactively search new modes by optimizing the *tempered residual potential energy*:

$$U_{\mathbf{r}}(\boldsymbol{\theta}, T) = -\log\left(f(\boldsymbol{\theta}) - \exp\left(\frac{1}{T}\log q(\boldsymbol{\theta})\right) + c\right)$$

Experiments

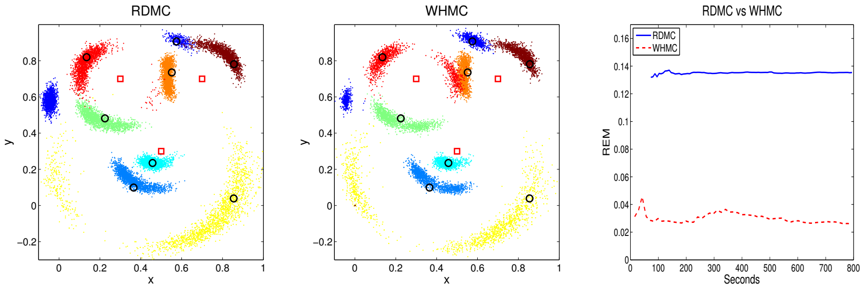
Mixture of Gaussians with known modes



$$\mathrm{REM}(t) = \|\overline{\boldsymbol{\theta}(t)} - \boldsymbol{\theta}^*\|_1 / \|\boldsymbol{\theta}^*\|_1$$

where $\overline{\boldsymbol{\theta}(t)}$ is the mean of MCMC samples obtained by time t and $\boldsymbol{\theta}^*$ is the true mean.

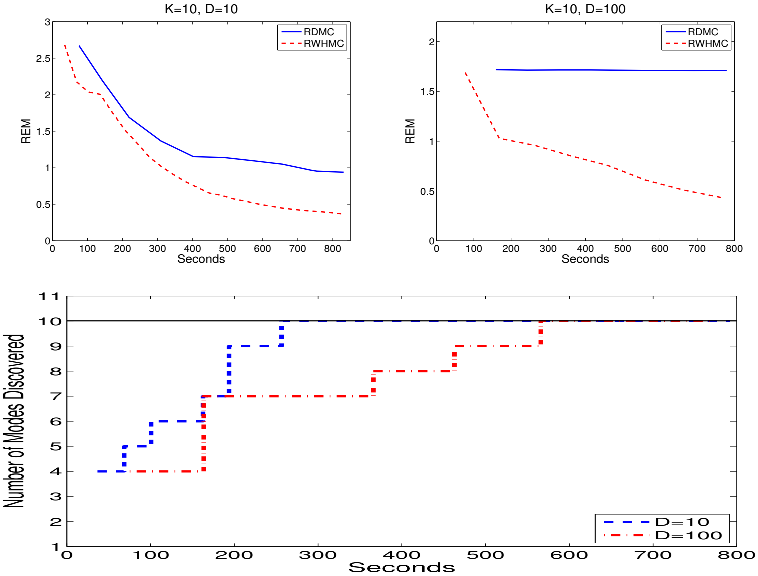
Sensor Network Localization



$$Z_{ij} := I(Y_{ij} > 0) | x \sim \mathrm{Binom}(1, \pi(x_i, x_j))$$

$$Y_{ij} | Z_{ij} = 1, x \sim \mathcal{N}(\|x_i - x_j\|, \sigma^2)$$

Mixture of Gaussians with unknown modes



Thank You!