Manifold Monte Carlo Methods

Mark Girolami

Department of Statistical Science
University College London

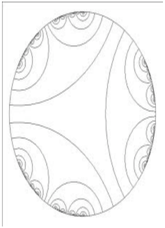
Joint work with Ben Calderhead

Research Section Ordinary Meeting

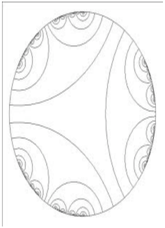
The Royal Statistical Society
October 13, 2010

Riemann manifold Langevin and Hamiltonian Monte Carlo Methods

Riemann manifold Langevin and Hamiltonian Monte Carlo Methods

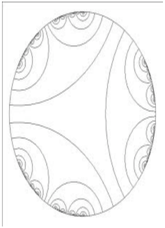




Riemann manifold Langevin and Hamiltonian Monte Carlo Methods

- Riemann manifold Langevin and Hamiltonian Monte Carlo Methods
  Girolami, M. & Calderhead, B., *J.R.Statist. Soc*. B (2011), **73**, *2*, 1 - 37.

Riemann manifold Langevin and Hamiltonian Monte Carlo Methods

- Riemann manifold Langevin and Hamiltonian Monte Carlo Methods
  Girolami, M. & Calderhead, B., *J.R.Statist. Soc.* B (2011), **73**, *2*, 1 - 37.
- Advanced Monte Carlo methodology founded on geometric principles

Talk Outline

- Motivation to improve MCMC capability for challenging problems

Talk Outline

- Motivation to improve MCMC capability for challenging problems
- Stochastic diffusion as adaptive proposal process

Talk Outline

- Motivation to improve MCMC capability for challenging problems
- Stochastic diffusion as adaptive proposal process
- Exploring geometric concepts in MCMC methodology

Talk Outline

- Motivation to improve MCMC capability for challenging problems
- Stochastic diffusion as adaptive proposal process
- Exploring geometric concepts in MCMC methodology
- Diffusions across Riemann manifold as proposal mechanism

Talk Outline

- Motivation to improve MCMC capability for challenging problems
- Stochastic diffusion as adaptive proposal process
- Exploring geometric concepts in MCMC methodology
- Diffusions across Riemann manifold as proposal mechanism
- Deterministic geodesic flows on manifold form basis of MCMC methods

Talk Outline

- Motivation to improve MCMC capability for challenging problems
- Stochastic diffusion as adaptive proposal process
- Exploring geometric concepts in MCMC methodology
- Diffusions across Riemann manifold as proposal mechanism
- Deterministic geodesic flows on manifold form basis of MCMC methods
- Illustrative Examples:-
  - Warped Bivariate Gaussian

Talk Outline

- Motivation to improve MCMC capability for challenging problems
- Stochastic diffusion as adaptive proposal process
- Exploring geometric concepts in MCMC methodology
- Diffusions across Riemann manifold as proposal mechanism
- Deterministic geodesic flows on manifold form basis of MCMC methods
- Illustrative Examples:-
  - Warped Bivariate Gaussian
  - Gaussian mixture model

Talk Outline

- Motivation to improve MCMC capability for challenging problems
- Stochastic diffusion as adaptive proposal process
- Exploring geometric concepts in MCMC methodology
- Diffusions across Riemann manifold as proposal mechanism
- Deterministic geodesic flows on manifold form basis of MCMC methods
- Illustrative Examples:-
  - Warped Bivariate Gaussian
  - Gaussian mixture model
  - Log-Gaussian Cox process

Talk Outline

- Motivation to improve MCMC capability for challenging problems
- Stochastic diffusion as adaptive proposal process
- Exploring geometric concepts in MCMC methodology
- Diffusions across Riemann manifold as proposal mechanism
- Deterministic geodesic flows on manifold form basis of MCMC methods
- Illustrative Examples:-
  - Warped Bivariate Gaussian
  - Gaussian mixture model
  - Log-Gaussian Cox process
- Conclusions

Motivation Simulation Based Inference

- Monte Carlo method employs samples from $p(\boldsymbol{\theta})$ to obtain estimate

$$\int \phi(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{N} \sum_n \phi(\boldsymbol{\theta}^n) + \mathcal{O}(N^{-\frac{1}{2}})$$

Motivation Simulation Based Inference

- Monte Carlo method employs samples from $p(\boldsymbol{\theta})$ to obtain estimate

$$\int \phi(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{1}{N}\sum_n \phi(\boldsymbol{\theta}^n) + \mathcal{O}(N^{-\frac{1}{2}})$$

- Draw $\boldsymbol{\theta}^n$ from ergodic Markov process with stationary distribution $p(\boldsymbol{\theta})$

Motivation Simulation Based Inference

- Monte Carlo method employs samples from $p(\boldsymbol{\theta})$ to obtain estimate

$$\int \phi(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{1}{N}\sum\nolimits_n \phi(\boldsymbol{\theta}^n) + \mathcal{O}(N^{-\frac{1}{2}})$$

- Draw $\boldsymbol{\theta}^n$ from ergodic Markov process with stationary distribution $p(\boldsymbol{\theta})$
- Construct process in two parts

Motivation Simulation Based Inference

- Monte Carlo method employs samples from $p(\boldsymbol{\theta})$ to obtain estimate

$$\int \phi(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{N} \sum_n \phi(\boldsymbol{\theta}^n) + \mathcal{O}(N^{-\frac{1}{2}})$$

- Draw $\boldsymbol{\theta}^n$ from ergodic Markov process with stationary distribution $p(\boldsymbol{\theta})$
- Construct process in two parts
  - Propose a move $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'$ with probability $p_p(\boldsymbol{\theta}' | \boldsymbol{\theta})$

Motivation Simulation Based Inference

- Monte Carlo method employs samples from $p(\boldsymbol{\theta})$ to obtain estimate

$$\int \phi(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{1}{N}\sum_n \phi(\boldsymbol{\theta}^n) + \mathcal{O}(N^{-\frac{1}{2}})$$

- Draw $\boldsymbol{\theta}^n$ from ergodic Markov process with stationary distribution $p(\boldsymbol{\theta})$
- Construct process in two parts
  - Propose a move $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'$ with probability $p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})$
  - accept or reject proposal with probability

$$p_a(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left[1, \frac{p(\boldsymbol{\theta}')p_p(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta})p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right]$$

Motivation Simulation Based Inference

- Monte Carlo method employs samples from $p(\boldsymbol{\theta})$ to obtain estimate

$$\int \phi(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{N} \sum_n \phi(\boldsymbol{\theta}^n) + \mathcal{O}(N^{-\frac{1}{2}})$$

- Draw $\boldsymbol{\theta}^n$ from ergodic Markov process with stationary distribution $p(\boldsymbol{\theta})$
- Construct process in two parts
  - Propose a move $\boldsymbol{\theta} \to \boldsymbol{\theta}'$ with probability $p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})$
  - accept or reject proposal with probability

$$p_a(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left[1, \frac{p(\boldsymbol{\theta}')p_p(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta})p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right]$$

- Efficiency dependent on $p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})$ defining proposal mechanism

Motivation Simulation Based Inference

- Monte Carlo method employs samples from $p(\boldsymbol{\theta})$ to obtain estimate

$$\int \phi(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{1}{N}\sum_n \phi(\boldsymbol{\theta}^n) + \mathcal{O}(N^{-\frac{1}{2}})$$

- Draw $\boldsymbol{\theta}^n$ from ergodic Markov process with stationary distribution $p(\boldsymbol{\theta})$
- Construct process in two parts
  - Propose a move $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'$ with probability $p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})$
  - accept or reject proposal with probability

$$p_a(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left[1, \frac{p(\boldsymbol{\theta}')p_p(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta})p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right]$$

- Efficiency dependent on $p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})$ defining proposal mechanism
- Success of MCMC reliant upon appropriate proposal design

Adaptive Proposal Distributions - Exploit Discretised Diffusion

Adaptive Proposal Distributions - Exploit Discretised Diffusion

- For $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$, $\mathcal{L}(\boldsymbol{\theta}) \equiv \log p(\boldsymbol{\theta})$, define Langevin diffusion

$$d\boldsymbol{\theta}(t) = \frac{1}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(t))dt + d\mathbf{b}(t)$$

Adaptive Proposal Distributions - Exploit Discretised Diffusion

- For $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$, $\mathcal{L}(\boldsymbol{\theta}) \equiv \log p(\boldsymbol{\theta})$, define Langevin diffusion

$$d\boldsymbol{\theta}(t) = \frac{1}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(t))dt + d\mathbf{b}(t)$$

- First order Euler-Maruyama discrete integration of diffusion

$$\boldsymbol{\theta}(\tau+\epsilon) = \boldsymbol{\theta}(\tau) + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(\tau)) + \epsilon\mathbf{z}(\tau)$$

Adaptive Proposal Distributions - Exploit Discretised Diffusion

- For $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$, $\mathcal{L}(\boldsymbol{\theta}) \equiv \log p(\boldsymbol{\theta})$, define Langevin diffusion

$$d\boldsymbol{\theta}(t) = \frac{1}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(t))dt + d\mathbf{b}(t)$$

- First order Euler-Maruyama discrete integration of diffusion

$$\boldsymbol{\theta}(\tau + \epsilon) = \boldsymbol{\theta}(\tau) + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(\tau)) + \epsilon\mathbf{z}(\tau)$$

- Proposal

$$p_p(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}'|\boldsymbol{\mu}(\boldsymbol{\theta},\epsilon),\epsilon^2\mathbf{I}) \quad \text{with} \quad \boldsymbol{\mu}(\boldsymbol{\theta},\epsilon) = \boldsymbol{\theta} + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$$

Adaptive Proposal Distributions - Exploit Discretised Diffusion

- For $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$, $\mathcal{L}(\boldsymbol{\theta}) \equiv \log p(\boldsymbol{\theta})$, define Langevin diffusion

$$d\boldsymbol{\theta}(t) = \frac{1}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(t))dt + d\mathbf{b}(t)$$

- First order Euler-Maruyama discrete integration of diffusion

$$\boldsymbol{\theta}(\tau + \epsilon) = \boldsymbol{\theta}(\tau) + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(\tau)) + \epsilon\mathbf{z}(\tau)$$

- Proposal

$$p_p(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}'|\boldsymbol{\mu}(\boldsymbol{\theta}, \epsilon), \epsilon^2\mathbf{I}) \quad \text{with} \quad \boldsymbol{\mu}(\boldsymbol{\theta}, \epsilon) = \boldsymbol{\theta} + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$$

- Acceptance probability to correct for bias

$$p_a(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left[1, \frac{p(\boldsymbol{\theta}')p_p(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta})p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right]$$

Adaptive Proposal Distributions - Exploit Discretised Diffusion

- For $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$, $\mathcal{L}(\boldsymbol{\theta}) \equiv \log p(\boldsymbol{\theta})$, define Langevin diffusion

$$d\boldsymbol{\theta}(t) = \frac{1}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(t))dt + d\mathbf{b}(t)$$

- First order Euler-Maruyama discrete integration of diffusion

$$\boldsymbol{\theta}(\tau + \epsilon) = \boldsymbol{\theta}(\tau) + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(\tau)) + \epsilon\mathbf{z}(\tau)$$

- Proposal

$$p_p(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}'|\boldsymbol{\mu}(\boldsymbol{\theta},\epsilon), \epsilon^2\mathbf{I}) \quad \text{with} \quad \boldsymbol{\mu}(\boldsymbol{\theta},\epsilon) = \boldsymbol{\theta} + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$$

- Acceptance probability to correct for bias

$$p_a(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left[1, \frac{p(\boldsymbol{\theta}')p_p(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta})p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right]$$

- Isotropic diffusion inefficient, employ pre-conditioning

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \epsilon^2\mathbf{M}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})/2 + \epsilon\sqrt{\mathbf{M}}\mathbf{z}$$

Adaptive Proposal Distributions - Exploit Discretised Diffusion

- For $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$, $\mathcal{L}(\boldsymbol{\theta}) \equiv \log p(\boldsymbol{\theta})$, define Langevin diffusion

$$d\boldsymbol{\theta}(t) = \frac{1}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(t))dt + d\mathbf{b}(t)$$

- First order Euler-Maruyama discrete integration of diffusion

$$\boldsymbol{\theta}(\tau + \epsilon) = \boldsymbol{\theta}(\tau) + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(\tau)) + \epsilon\mathbf{z}(\tau)$$

- Proposal

$$p_p(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}'|\boldsymbol{\mu}(\boldsymbol{\theta},\epsilon), \epsilon^2\mathbf{I}) \quad \text{with} \quad \boldsymbol{\mu}(\boldsymbol{\theta},\epsilon) = \boldsymbol{\theta} + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$$

- Acceptance probability to correct for bias

$$p_a(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left[1, \frac{p(\boldsymbol{\theta}')p_p(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta})p_p(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right]$$

- Isotropic diffusion inefficient, employ pre-conditioning

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \epsilon^2\mathbf{M}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})/2 + \epsilon\sqrt{\mathbf{M}}\mathbf{z}$$

- How to set **M** systematically? Tuning in transient & stationary phases

Geometric Concepts in MCMC

Geometric Concepts in MCMC

- Rao, 1945; Jeffreys, 1948, to first order

$$\int p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})\log\frac{p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}d\boldsymbol{\theta}\approx\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$$

Geometric Concepts in MCMC

- Rao, 1945; Jeffreys, 1948, to first order

$$\int p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})\log\frac{p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}d\boldsymbol{\theta}\approx\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$$

where

$$\mathbf{G}(\boldsymbol{\theta})=E_{\mathbf{y}|\boldsymbol{\theta}}\left\{\frac{\nabla_{\boldsymbol{\theta}}p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}\frac{\nabla_{\boldsymbol{\theta}}p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}^{\mathsf{T}}\right\}$$

Geometric Concepts in MCMC

- Rao, 1945; Jeffreys, 1948, to first order

$$\int p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta}) \log \frac{p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})} d\boldsymbol{\theta} \approx \delta\boldsymbol{\theta}^{\mathsf{T}} \mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$$

  where

$$\mathbf{G}(\boldsymbol{\theta}) = E_{\mathbf{y}|\boldsymbol{\theta}} \left\{ \frac{\nabla_{\boldsymbol{\theta}} p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})} \frac{\nabla_{\boldsymbol{\theta}} p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}^{\mathsf{T}} \right\}$$

- Fisher Information $\mathbf{G}(\boldsymbol{\theta})$ is p.d. metric defining a Riemann manifold

Geometric Concepts in MCMC

- Rao, 1945; Jeffreys, 1948, to first order

$$\int p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta}) \log \frac{p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})} d\boldsymbol{\theta} \approx \delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$$

  where

$$\mathbf{G}(\boldsymbol{\theta}) = E_{\mathbf{y}|\boldsymbol{\theta}} \left\{ \frac{\nabla_{\boldsymbol{\theta}} p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})} \frac{\nabla_{\boldsymbol{\theta}} p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}^{\mathsf{T}} \right\}$$

- Fisher Information $\mathbf{G}(\boldsymbol{\theta})$ is p.d. metric defining a Riemann manifold
- Non-Euclidean geometry for probabilities - distances, metrics, invariants, curvature, geodesics

Geometric Concepts in MCMC

- Rao, 1945; Jeffreys, 1948, to first order

$$\int p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})\log\frac{p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}d\boldsymbol{\theta}\approx\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$$

where

$$\mathbf{G}(\boldsymbol{\theta})=E_{\mathbf{y}|\boldsymbol{\theta}}\left\{\frac{\nabla_{\boldsymbol{\theta}}p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}\frac{\nabla_{\boldsymbol{\theta}}p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}^{\mathsf{T}}\right\}$$

- Fisher Information $\mathbf{G}(\boldsymbol{\theta})$ is p.d. metric defining a Riemann manifold
- Non-Euclidean geometry for probabilities - distances, metrics, invariants, curvature, geodesics
- Asymptotic statistical analysis. Amari, 1981, 85, 90; Murray & Rice, 1993; Critchley *et al*, 1993; Kass, 1989; Efron, 1975; Dawid, 1975; Lauritsen, 1989

Geometric Concepts in MCMC

- Rao, 1945; Jeffreys, 1948, to first order

$$\int p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})\log\frac{p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}d\boldsymbol{\theta}\approx\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$$

  where

$$\mathbf{G}(\boldsymbol{\theta})=E_{\mathbf{y}|\boldsymbol{\theta}}\left\{\frac{\nabla_{\boldsymbol{\theta}}p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}\frac{\nabla_{\boldsymbol{\theta}}p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}^{\mathsf{T}}\right\}$$

- Fisher Information $\mathbf{G}(\boldsymbol{\theta})$ is p.d. metric defining a Riemann manifold
- Non-Euclidean geometry for probabilities - distances, metrics, invariants, curvature, geodesics
- Asymptotic statistical analysis. Amari, 1981, 85, 90; Murray & Rice, 1993; Critchley *et al*, 1993; Kass, 1989; Efron, 1975; Dawid, 1975; Lauritsen, 1989
- Statistical shape analysis Kent *et al*, 1996; Dryden & Mardia, 1998

Geometric Concepts in MCMC

- Rao, 1945; Jeffreys, 1948, to first order

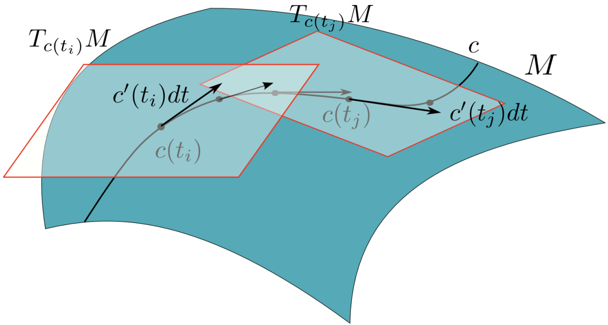
$$\int p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})\log\frac{p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}d\boldsymbol{\theta} \approx \delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$$

  where

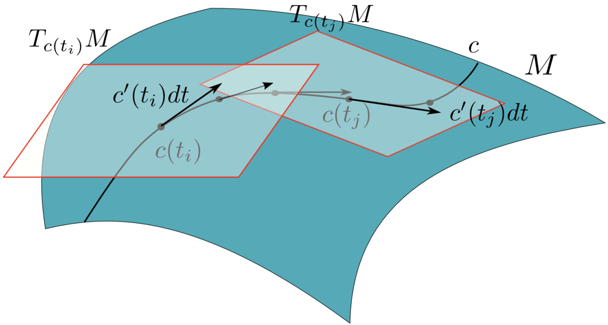
$$\mathbf{G}(\boldsymbol{\theta}) = E_{\mathbf{y}|\boldsymbol{\theta}}\left\{\frac{\nabla_{\boldsymbol{\theta}}p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}\frac{\nabla_{\boldsymbol{\theta}}p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta})}^{\mathsf{T}}\right\}$$

- Fisher Information $\mathbf{G}(\boldsymbol{\theta})$ is p.d. metric defining a Riemann manifold
- Non-Euclidean geometry for probabilities - distances, metrics, invariants, curvature, geodesics
- Asymptotic statistical analysis. Amari, 1981, 85, 90; Murray & Rice, 1993; Critchley *et al*, 1993; Kass, 1989; Efron, 1975; Dawid, 1975; Lauritsen, 1989
- Statistical shape analysis Kent *et al*, 1996; Dryden & Mardia, 1998
- Can geometric structure be employed in MCMC methodology?

Geometric Concepts in MCMC

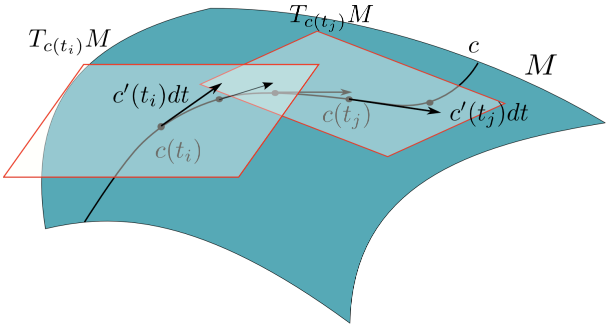


Geometric Concepts in MCMC



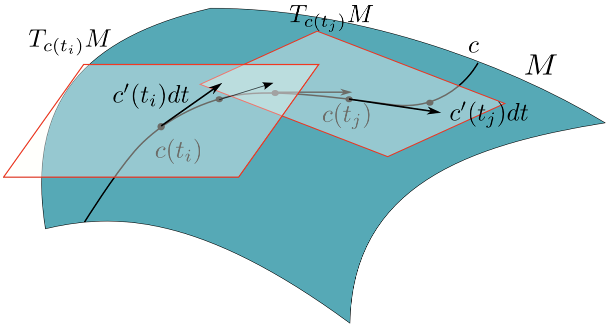
- Tangent space - local metric defined by $\delta\boldsymbol{\theta}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta} = \sum_{k,l} g_{kl}\delta\theta_k\delta\theta_l$

Geometric Concepts in MCMC



- Tangent space - local metric defined by $\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta} = \sum_{k,l} g_{kl}\delta\theta_k\delta\theta_l$
- Christoffel symbols - characterise connection on curved manifold

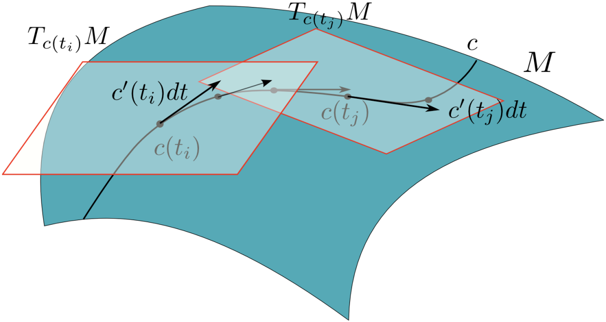
Geometric Concepts in MCMC



- Tangent space - local metric defined by $\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta} = \sum_{k,l} g_{kl}\delta\theta_k\delta\theta_l$
- Christoffel symbols - characterise connection on curved manifold

$$\Gamma^i_{kl} = \frac{1}{2}\sum_m g^{im}\left(\frac{\partial g_{mk}}{\partial\theta^l} + \frac{\partial g_{ml}}{\partial\theta^k} - \frac{\partial g_{kl}}{\partial\theta^m}\right)$$

Geometric Concepts in MCMC

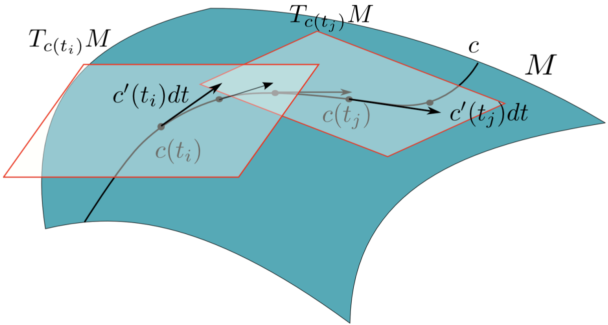


- Tangent space - local metric defined by $\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta} = \sum_{k,l} g_{kl}\delta\theta_k\delta\theta_l$
- Christoffel symbols - characterise connection on curved manifold

$$\Gamma^i_{kl} = \frac{1}{2}\sum_m g^{im}\left(\frac{\partial g_{mk}}{\partial\theta^l} + \frac{\partial g_{ml}}{\partial\theta^k} - \frac{\partial g_{kl}}{\partial\theta^m}\right)$$

- Geodesics - shortest path between two points on manifold

Geometric Concepts in MCMC



- Tangent space - local metric defined by $\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta} = \sum_{k,l} g_{kl}\delta\theta_k\delta\theta_l$
- Christoffel symbols - characterise connection on curved manifold

$$\Gamma^i_{kl} = \frac{1}{2}\sum_m g^{im}\left(\frac{\partial g_{mk}}{\partial\theta^l} + \frac{\partial g_{ml}}{\partial\theta^k} - \frac{\partial g_{kl}}{\partial\theta^m}\right)$$

- Geodesics - shortest path between two points on manifold

$$\frac{d^2\theta^i}{dt^2} + \sum_{k,l}\Gamma^i_{kl}\frac{d\theta^k}{dt}\frac{d\theta^l}{dt} = 0$$

Illustration of Geometric Concepts

- Consider Normal density $p(x|\mu, \sigma) = \mathcal{N}_x(\mu, \sigma)$

Illustration of Geometric Concepts

- Consider Normal density $p(x|\mu, \sigma) = \mathcal{N}_x(\mu, \sigma)$
- Local inner product on tangent space defined by metric tensor, i.e. $\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\mu, \sigma)^{\mathsf{T}}$

Illustration of Geometric Concepts

- Consider Normal density $p(x|\mu, \sigma) = \mathcal{N}_x(\mu, \sigma)$
- Local inner product on tangent space defined by metric tensor, i.e. $\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\mu, \sigma)^{\mathsf{T}}$
- Metric is Fisher Information

$$\mathbf{G}(\mu, \sigma) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{bmatrix}$$

Illustration of Geometric Concepts

- Consider Normal density $p(x|\mu,\sigma) = \mathcal{N}_x(\mu,\sigma)$
- Local inner product on tangent space defined by metric tensor, i.e. $\delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\mu,\sigma)^{\mathsf{T}}$
- Metric is Fisher Information

$$\mathbf{G}(\mu,\sigma) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{bmatrix}$$

- Inner-product $\sigma^{-2}(\delta\mu^2 + 2\delta\sigma^2)$ so densities $\mathcal{N}(0,1)$ & $\mathcal{N}(1,1)$ further apart than the densities $\mathcal{N}(0,2)$ & $\mathcal{N}(1,2)$ - distance non-Euclidean

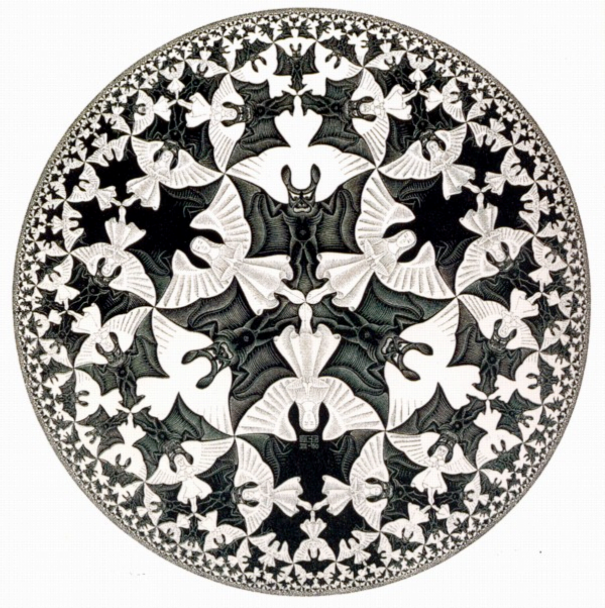
Illustration of Geometric Concepts

- Consider Normal density $p(x|\mu, \sigma) = \mathcal{N}_x(\mu, \sigma)$
- Local inner product on tangent space defined by metric tensor, i.e. $\delta\boldsymbol{\theta}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\mu, \sigma)^\mathsf{T}$
- Metric is Fisher Information

$$\mathbf{G}(\mu, \sigma) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{bmatrix}$$

- Inner-product $\sigma^{-2}(\delta\mu^2 + 2\delta\sigma^2)$ so densities $\mathcal{N}(0, 1)$ & $\mathcal{N}(1, 1)$ further apart than the densities $\mathcal{N}(0, 2)$ & $\mathcal{N}(1, 2)$ - distance non-Euclidean
- Metric tensor for univariate Normal defines a Hyperbolic Space

M.C. Escher, Heaven and Hell, 1960



Langevin Diffusion on Riemannian manifold

- Discretised Langevin diffusion on manifold defines proposal mechanism

$$\theta'_d = \theta_d + \frac{\epsilon^2}{2}\left(\boldsymbol{G}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta)\right)_d - \epsilon^2 \sum_{i,j}^{D} \boldsymbol{G}(\theta)^{-1}_{ij}\Gamma^d_{ij} + \epsilon\left(\sqrt{\mathbf{G}^{-1}(\theta)}\mathbf{z}\right)_d$$

Langevin Diffusion on Riemannian manifold

- Discretised Langevin diffusion on manifold defines proposal mechanism

$$\theta'_d = \theta_d + \frac{\epsilon^2}{2}\left(\boldsymbol{G}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta)\right)_d - \epsilon^2 \sum_{i,j}^{D} \boldsymbol{G}(\theta)^{-1}_{ij}\Gamma^d_{ij} + \epsilon\left(\sqrt{\mathbf{G}^{-1}(\theta)}\mathbf{z}\right)_d$$

- Manifold with constant curvature then proposal mechanism reduces to

$$\theta' = \theta + \frac{\epsilon^2}{2}\boldsymbol{G}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta) + \epsilon\sqrt{\mathbf{G}^{-1}(\theta)}\mathbf{z}$$

Langevin Diffusion on Riemannian manifold

- Discretised Langevin diffusion on manifold defines proposal mechanism

$$\theta'_d = \theta_d + \frac{\epsilon^2}{2}\left(\boldsymbol{G}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta)\right)_d - \epsilon^2 \sum_{i,j}^{D} \boldsymbol{G}(\theta)^{-1}_{ij}\Gamma^d_{ij} + \epsilon\left(\sqrt{\mathbf{G}^{-1}(\theta)}\mathbf{z}\right)_d$$

- Manifold with constant curvature then proposal mechanism reduces to

$$\theta' = \theta + \frac{\epsilon^2}{2}\boldsymbol{G}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta) + \epsilon\sqrt{\mathbf{G}^{-1}(\theta)}\mathbf{z}$$

- MALA proposal with preconditioning

$$\theta' = \theta + \frac{\epsilon^2}{2}\mathbf{M}\nabla_\theta \mathcal{L}(\theta) + \epsilon\sqrt{\mathbf{M}}\mathbf{z}$$

Langevin Diffusion on Riemannian manifold

- Discretised Langevin diffusion on manifold defines proposal mechanism

$$\theta'_d = \theta_d + \frac{\epsilon^2}{2}\left(\boldsymbol{G}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta)\right)_d - \epsilon^2 \sum_{i,j}^{D} \boldsymbol{G}(\theta)^{-1}_{ij}\Gamma^d_{ij} + \epsilon\left(\sqrt{\mathbf{G}^{-1}(\theta)}\mathbf{z}\right)_d$$

- Manifold with constant curvature then proposal mechanism reduces to

$$\theta' = \theta + \frac{\epsilon^2}{2}\boldsymbol{G}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta) + \epsilon\sqrt{\mathbf{G}^{-1}(\theta)}\mathbf{z}$$

- MALA proposal with preconditioning

$$\theta' = \theta + \frac{\epsilon^2}{2}\mathbf{M}\nabla_\theta \mathcal{L}(\theta) + \epsilon\sqrt{\mathbf{M}}\mathbf{z}$$

- Proposal and acceptance probability

$$p_p(\theta'|\theta) = \mathcal{N}(\theta'|\mu(\theta,\epsilon), \epsilon^2\mathbf{G}(\theta))$$

$$p_a(\theta'|\theta) = \min\left[1, \frac{p(\theta')p_p(\theta|\theta')}{p(\theta)p_p(\theta'|\theta)}\right]$$

Langevin Diffusion on Riemannian manifold

- Discretised Langevin diffusion on manifold defines proposal mechanism

$$\theta'_d = \theta_d + \frac{\epsilon^2}{2}\left(\mathbf{G}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta)\right)_d - \epsilon^2 \sum_{i,j}^{D} \mathbf{G}(\theta)^{-1}_{ij}\Gamma^d_{ij} + \epsilon\left(\sqrt{\mathbf{G}^{-1}(\theta)}\mathbf{z}\right)_d$$

- Manifold with constant curvature then proposal mechanism reduces to

$$\theta' = \theta + \frac{\epsilon^2}{2}\mathbf{G}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta) + \epsilon\sqrt{\mathbf{G}^{-1}(\theta)}\mathbf{z}$$

- MALA proposal with preconditioning

$$\theta' = \theta + \frac{\epsilon^2}{2}\mathbf{M}\nabla_\theta \mathcal{L}(\theta) + \epsilon\sqrt{\mathbf{M}}\mathbf{z}$$

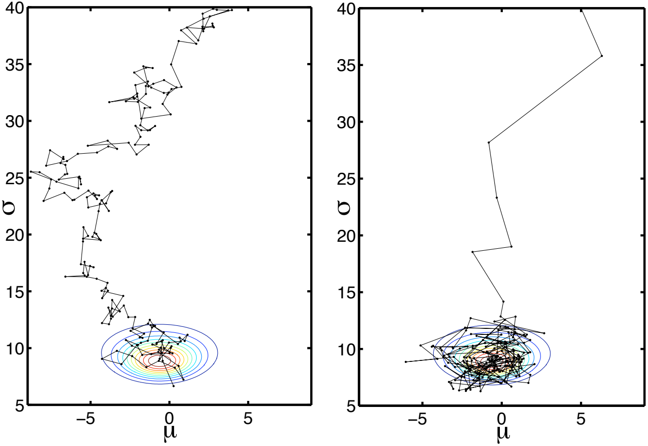
- Proposal and acceptance probability

$$p_p(\theta'|\theta) = \mathcal{N}(\theta'|\mu(\theta,\epsilon), \epsilon^2\mathbf{G}(\theta))$$

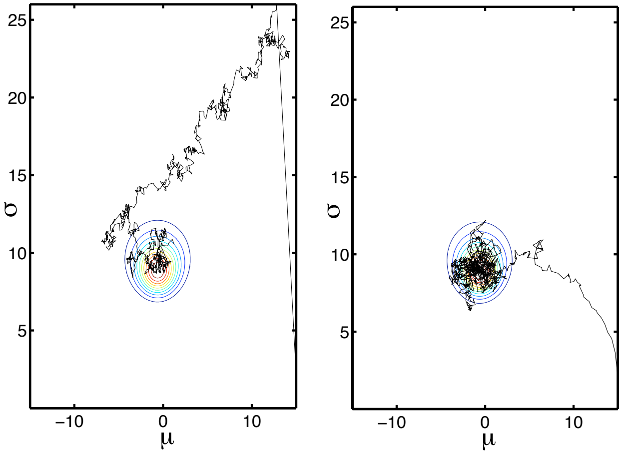
$$p_a(\theta'|\theta) = \min\left[1, \frac{p(\theta')p_p(\theta|\theta')}{p(\theta)p_p(\theta'|\theta)}\right]$$

- Proposal mechanism diffuses approximately along the manifold

Langevin Diffusion on Riemannian manifold



Langevin Diffusion on Riemannian manifold



Geodesic flow as proposal mechanism

- Desirable that proposals follow direct path on manifold - geodesics

Geodesic flow as proposal mechanism

- Desirable that proposals follow direct path on manifold - geodesics
- Define geodesic flow on manifold by solving

$$\frac{d^2\theta^i}{dt^2} + \sum_{k,l} \Gamma^i_{kl} \frac{d\theta^k}{dt}\frac{d\theta^l}{dt} = 0$$

Geodesic flow as proposal mechanism

- Desirable that proposals follow direct path on manifold - geodesics
- Define geodesic flow on manifold by solving

$$\frac{d^2\theta^i}{dt^2} + \sum_{k,l} \Gamma^i_{kl} \frac{d\theta^k}{dt} \frac{d\theta^l}{dt} = 0$$

- How can this be exploited in the design of a transition operator?

Geodesic flow as proposal mechanism

- Desirable that proposals follow direct path on manifold - geodesics
- Define geodesic flow on manifold by solving

$$\frac{d^2\theta^i}{dt^2} + \sum_{k,l} \Gamma^i_{kl} \frac{d\theta^k}{dt}\frac{d\theta^l}{dt} = 0$$

- How can this be exploited in the design of a transition operator?
- Need slight detour

Geodesic flow as proposal mechanism

- Desirable that proposals follow direct path on manifold - geodesics
- Define geodesic flow on manifold by solving

$$\frac{d^2\theta^i}{dt^2} + \sum_{k,l} \Gamma^i_{kl} \frac{d\theta^k}{dt} \frac{d\theta^l}{dt} = 0$$

- How can this be exploited in the design of a transition operator?
- Need slight detour - first define log-density under model as $\mathcal{L}(\boldsymbol{\theta})$

Geodesic flow as proposal mechanism

- Desirable that proposals follow direct path on manifold - geodesics
- Define geodesic flow on manifold by solving

$$\frac{d^2\theta^i}{dt^2} + \sum_{k,l} \Gamma^i_{kl} \frac{d\theta^k}{dt} \frac{d\theta^l}{dt} = 0$$

- How can this be exploited in the design of a transition operator?
- Need slight detour - first define log-density under model as $\mathcal{L}(\boldsymbol{\theta})$
- Introduce auxiliary variable $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}))$

Geodesic flow as proposal mechanism

- Desirable that proposals follow direct path on manifold - geodesics
- Define geodesic flow on manifold by solving

$$\frac{d^2\theta^i}{dt^2} + \sum_{k,l} \Gamma^i_{kl} \frac{d\theta^k}{dt}\frac{d\theta^l}{dt} = 0$$

- How can this be exploited in the design of a transition operator?
- Need slight detour - first define log-density under model as $\mathcal{L}(\boldsymbol{\theta})$
- Introduce auxiliary variable $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}))$
- Negative joint log density is

$$H(\boldsymbol{\theta}, \mathbf{p}) = -\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\log 2\pi^D |\mathbf{G}(\boldsymbol{\theta})| + \frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$$

Riemannian Hamiltonian Monte Carlo

- Negative joint log-density $\equiv$ Hamiltonian defined on Riemann manifold

$$H(\theta, \mathbf{p}) = \underbrace{-\mathcal{L}(\theta) + \frac{1}{2}\log 2\pi^{D}|\mathbf{G}(\theta)|}_{\text{Potential Energy}} + \underbrace{\frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\theta)^{-1}\mathbf{p}}_{\text{Kinetic Energy}}$$

Riemannian Hamiltonian Monte Carlo

- Negative joint log-density $\equiv$ Hamiltonian defined on Riemann manifold

$$H(\boldsymbol{\theta}, \mathbf{p}) = \underbrace{-\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2} \log 2\pi^D |\mathbf{G}(\boldsymbol{\theta})|}_{\text{Potential Energy}} + \underbrace{\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}}_{\text{Kinetic Energy}}$$

- Marginal density follows as required

$$p(\boldsymbol{\theta}) \propto \frac{\exp\{\mathcal{L}(\boldsymbol{\theta})\}}{\sqrt{2\pi^D |\mathbf{G}(\boldsymbol{\theta})|}} \int \exp\left\{-\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\right\} d\mathbf{p} = \exp\{\mathcal{L}(\boldsymbol{\theta})\}$$

Riemannian Hamiltonian Monte Carlo

- Negative joint log-density $\equiv$ Hamiltonian defined on Riemann manifold

$$H(\boldsymbol{\theta}, \mathbf{p}) = \underbrace{-\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\log 2\pi^D |\mathbf{G}(\boldsymbol{\theta})|}_{\text{Potential Energy}} + \underbrace{\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}}_{\text{Kinetic Energy}}$$

- Marginal density follows as required

$$p(\boldsymbol{\theta}) \propto \frac{\exp\{\mathcal{L}(\boldsymbol{\theta})\}}{\sqrt{2\pi^D|\mathbf{G}(\boldsymbol{\theta})|}} \int \exp\left\{-\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\right\} d\mathbf{p} = \exp\{\mathcal{L}(\boldsymbol{\theta})\}$$

- Obtain samples from marginal $p(\boldsymbol{\theta})$ using Gibbs sampler for $p(\boldsymbol{\theta}, \mathbf{p})$

$$\begin{aligned} \mathbf{p}^{n+1}|\boldsymbol{\theta}^n &\sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}^n)) \\ \boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1} &\sim p(\boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1}) \end{aligned}$$

Riemannian Hamiltonian Monte Carlo

- Negative joint log-density $\equiv$ Hamiltonian defined on Riemann manifold

$$H(\boldsymbol{\theta}, \mathbf{p}) = \underbrace{-\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\log 2\pi^D|\mathbf{G}(\boldsymbol{\theta})|}_{\text{Potential Energy}} + \underbrace{\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}}_{\text{Kinetic Energy}}$$

- Marginal density follows as required

$$p(\boldsymbol{\theta}) \propto \frac{\exp\{\mathcal{L}(\boldsymbol{\theta})\}}{\sqrt{2\pi^D|\mathbf{G}(\boldsymbol{\theta})|}} \int \exp\left\{-\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\right\} d\mathbf{p} = \exp\{\mathcal{L}(\boldsymbol{\theta})\}$$

- Obtain samples from marginal $p(\boldsymbol{\theta})$ using Gibbs sampler for $p(\boldsymbol{\theta}, \mathbf{p})$

$$\begin{aligned} \mathbf{p}^{n+1}|\boldsymbol{\theta}^n &\sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}^n)) \\ \boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1} &\sim p(\boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1}) \end{aligned}$$

- Employ Hamiltonian dynamics to propose samples for $p(\boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1})$, Duane *et al*, 1987; Neal, 2010.

Riemannian Hamiltonian Monte Carlo

- Negative joint log-density ≡ Hamiltonian defined on Riemann manifold

$$H(\boldsymbol{\theta}, \mathbf{p}) = \underbrace{-\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\log 2\pi^D |\mathbf{G}(\boldsymbol{\theta})|}_{\text{Potential Energy}} + \underbrace{\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}}_{\text{Kinetic Energy}}$$

- Marginal density follows as required

$$p(\boldsymbol{\theta}) \propto \frac{\exp\{\mathcal{L}(\boldsymbol{\theta})\}}{\sqrt{2\pi^D|\mathbf{G}(\boldsymbol{\theta})|}} \int \exp\left\{-\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\right\} d\mathbf{p} = \exp\{\mathcal{L}(\boldsymbol{\theta})\}$$

- Obtain samples from marginal $p(\boldsymbol{\theta})$ using Gibbs sampler for $p(\boldsymbol{\theta}, \mathbf{p})$

$$\begin{aligned} \mathbf{p}^{n+1}|\boldsymbol{\theta}^n &\sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}^n)) \\ \boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1} &\sim p(\boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1}) \end{aligned}$$

- Employ Hamiltonian dynamics to propose samples for $p(\boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1})$, Duane *et al*, 1987; Neal, 2010.

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial}{\partial \mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}) \qquad \frac{d\mathbf{p}}{dt} = -\frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p})$$

Riemannian Manifold Hamiltonian Monte Carlo

- Consider the Hamiltonian $\tilde{H}(\theta, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\tilde{\mathbf{G}}(\theta)^{-1}\mathbf{p}$

Riemannian Manifold Hamiltonian Monte Carlo

- Consider the Hamiltonian $\tilde{H}(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\tilde{\mathbf{G}}(\boldsymbol{\theta})^{-1}\mathbf{p}$
- Hamiltonians with only a quadratic kinetic energy term exactly describe geodesic flow on the coordinate space $\boldsymbol{\theta}$ with metric $\tilde{\mathbf{G}}$

Riemannian Manifold Hamiltonian Monte Carlo

- Consider the Hamiltonian $\tilde{H}(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\tilde{\mathbf{G}}(\boldsymbol{\theta})^{-1}\mathbf{p}$
- Hamiltonians with only a quadratic kinetic energy term exactly describe geodesic flow on the coordinate space $\boldsymbol{\theta}$ with metric $\tilde{\mathbf{G}}$
- However our Hamiltonian is $H(\boldsymbol{\theta}, \mathbf{p}) = V(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$

Riemannian Manifold Hamiltonian Monte Carlo

- Consider the Hamiltonian $\tilde{H}(\theta, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\tilde{\mathbf{G}}(\theta)^{-1}\mathbf{p}$
- Hamiltonians with only a quadratic kinetic energy term exactly describe geodesic flow on the coordinate space θ with metric $\tilde{\mathbf{G}}$
- However our Hamiltonian is $H(\theta, \mathbf{p}) = V(\theta) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\theta)^{-1}\mathbf{p}$
- If we define $\tilde{\mathbf{G}}(\theta) = \mathbf{G}(\theta) \times (h - V(\theta))$, where h is a constant $H(\theta, \mathbf{p})$

Riemannian Manifold Hamiltonian Monte Carlo

- Consider the Hamiltonian $\tilde{H}(\theta, \mathbf{p}) = \frac{1}{2}\mathbf{p}^\mathsf{T}\tilde{\mathbf{G}}(\theta)^{-1}\mathbf{p}$
- Hamiltonians with only a quadratic kinetic energy term exactly describe geodesic flow on the coordinate space θ with metric $\tilde{\mathbf{G}}$
- However our Hamiltonian is $H(\theta, \mathbf{p}) = V(\theta) + \frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\theta)^{-1}\mathbf{p}$
- If we define $\tilde{\mathbf{G}}(\theta) = \mathbf{G}(\theta) \times (h - V(\theta))$, where h is a constant $H(\theta, \mathbf{p})$
- Then the Maupertuis principle tells us that the Hamiltonian flow for $H(\theta, \mathbf{p})$ and $\tilde{H}(\theta, \mathbf{p})$ are equivalent along energy level h

Riemannian Manifold Hamiltonian Monte Carlo

- Consider the Hamiltonian $\tilde{H}(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\tilde{\mathbf{G}}(\boldsymbol{\theta})^{-1}\mathbf{p}$
- Hamiltonians with only a quadratic kinetic energy term exactly describe geodesic flow on the coordinate space $\boldsymbol{\theta}$ with metric $\tilde{\mathbf{G}}$
- However our Hamiltonian is $H(\boldsymbol{\theta}, \mathbf{p}) = V(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$
- If we define $\tilde{\mathbf{G}}(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta}) \times (h - V(\boldsymbol{\theta}))$, where h is a constant $H(\boldsymbol{\theta}, \mathbf{p})$
- Then the Maupertuis principle tells us that the Hamiltonian flow for $H(\boldsymbol{\theta}, \mathbf{p})$ and $\tilde{H}(\boldsymbol{\theta}, \mathbf{p})$ are equivalent along energy level h
- The solution of

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial}{\partial \mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}) \qquad \frac{d\boldsymbol{p}}{dt} = -\frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p})$$

Riemannian Manifold Hamiltonian Monte Carlo

- Consider the Hamiltonian $\tilde{H}(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\tilde{\mathbf{G}}(\boldsymbol{\theta})^{-1}\mathbf{p}$
- Hamiltonians with only a quadratic kinetic energy term exactly describe geodesic flow on the coordinate space $\boldsymbol{\theta}$ with metric $\tilde{\mathbf{G}}$
- However our Hamiltonian is $H(\boldsymbol{\theta}, \mathbf{p}) = V(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$
- If we define $\tilde{\mathbf{G}}(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta}) \times (h - V(\boldsymbol{\theta}))$, where h is a constant $H(\boldsymbol{\theta}, \mathbf{p})$
- Then the Maupertuis principle tells us that the Hamiltonian flow for $H(\boldsymbol{\theta}, \mathbf{p})$ and $\tilde{H}(\boldsymbol{\theta}, \mathbf{p})$ are equivalent along energy level h
- The solution of

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial}{\partial \mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}) \quad \frac{d\boldsymbol{p}}{dt} = -\frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p})$$

is therefore equivalent to the solution of

$$\frac{d^2\theta^i}{dt^2} + \sum_{k,l} \tilde{\Gamma}^i_{kl} \frac{d\theta^k}{dt}\frac{d\theta^l}{dt} = 0$$

Riemannian Manifold Hamiltonian Monte Carlo

- Consider the Hamiltonian $\tilde{H}(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\tilde{\mathbf{G}}(\boldsymbol{\theta})^{-1}\mathbf{p}$
- Hamiltonians with only a quadratic kinetic energy term exactly describe geodesic flow on the coordinate space $\boldsymbol{\theta}$ with metric $\tilde{\mathbf{G}}$
- However our Hamiltonian is $H(\boldsymbol{\theta}, \mathbf{p}) = V(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$
- If we define $\tilde{\mathbf{G}}(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta}) \times (h - V(\boldsymbol{\theta}))$, where h is a constant $H(\boldsymbol{\theta}, \mathbf{p})$
- Then the Maupertuis principle tells us that the Hamiltonian flow for $H(\boldsymbol{\theta}, \mathbf{p})$ and $\tilde{H}(\boldsymbol{\theta}, \mathbf{p})$ are equivalent along energy level h
- The solution of

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial}{\partial \mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}) \qquad \frac{d\mathbf{p}}{dt} = -\frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p})$$

  is therefore equivalent to the solution of

$$\frac{d^2\theta^i}{dt^2} + \sum_{k,l} \tilde{\Gamma}^i_{kl} \frac{d\theta^k}{dt} \frac{d\theta^l}{dt} = 0$$

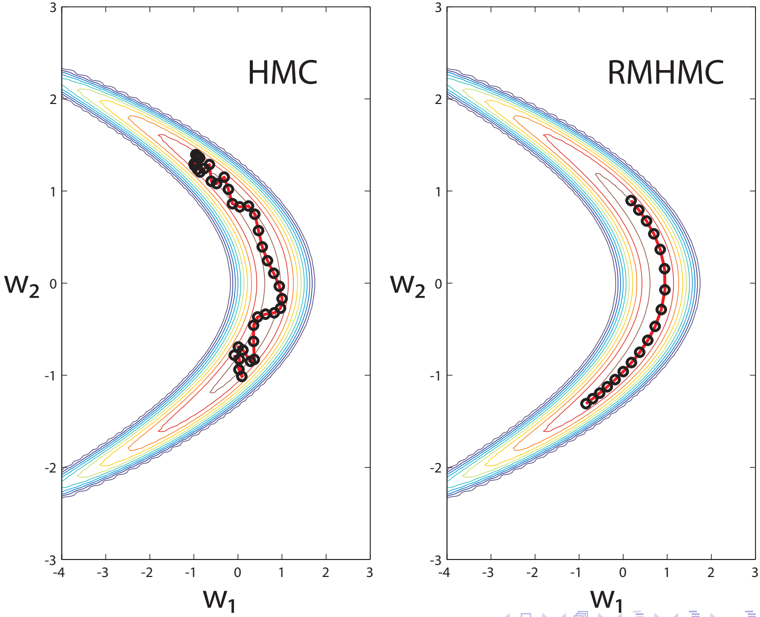
- RMHMC proposals are along the manifold geodesics

Warped Bivariate Gaussian

- $p(w_1, w_2|\mathbf{y}, \sigma_x, \sigma_y) \propto \prod_{n=1}^{N} \mathcal{N}(y_n|w_1 + w_2^2, \sigma_y^2)\mathcal{N}(w_1, w_2|\mathbf{0}, \sigma_x^2\mathbf{I})$

Warped Bivariate Gaussian

- $p(w_1, w_2|\mathbf{y}, \sigma_x, \sigma_y) \propto \prod_{n=1}^{N} \mathcal{N}(y_n|w_1 + w_2^2, \sigma_y^2)\mathcal{N}(w_1, w_2|\mathbf{0}, \sigma_x^2\mathbf{I})$



Gaussian Mixture Model

- Univariate finite mixture model

$$p(x_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k^2)$$

Gaussian Mixture Model

- Univariate finite mixture model

$$p(x_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k^2)$$

- FI based metric tensor non-analytic - employ empirical FI

Gaussian Mixture Model

- Univariate finite mixture model

$$p(x_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k^2)$$

- FI based metric tensor non-analytic - employ empirical FI

$$\boldsymbol{G}(\boldsymbol{\theta}) = \frac{1}{N}\boldsymbol{S}^T\boldsymbol{S} - \frac{1}{N^2}\bar{\boldsymbol{s}}\bar{\boldsymbol{s}}^T \quad \xrightarrow[N\to\infty]{} \quad \text{cov}\left(\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\right) = \boldsymbol{I}(\boldsymbol{\theta})$$

$$\frac{\partial \boldsymbol{G}(\boldsymbol{\theta})}{\partial \theta_d} = \frac{1}{N}\left(\frac{\partial \boldsymbol{S}^T}{\partial \theta_d}\boldsymbol{S} + \boldsymbol{S}^T\frac{\partial \boldsymbol{S}}{\partial \theta_d}\right) - \frac{1}{N^2}\left(\frac{\partial \bar{\boldsymbol{s}}}{\partial \theta_d}\bar{\boldsymbol{s}}^T + \bar{\boldsymbol{s}}\frac{\partial \bar{\boldsymbol{s}}^T}{\partial \theta_d}\right)$$

with score matrix $\boldsymbol{S}$ with elements $S_{i,d} = \frac{\partial \log p(x_i|\boldsymbol{\theta})}{\partial \theta_d}$ and $\bar{\boldsymbol{s}} = \sum_{i=1}^{N} \boldsymbol{S}_{i,\cdot}^T$

Gaussian Mixture Model

- Univariate finite mixture model

$$p(x|\mu, \sigma^2) = 0.7 \times \mathcal{N}(x|0, \sigma^2) + 0.3 \times \mathcal{N}(x|\mu, \sigma^2)$$

Gaussian Mixture Model

- Univariate finite mixture model

$$p(x|\mu,\sigma^2) = 0.7 \times \mathcal{N}(x|0,\sigma^2) + 0.3 \times \mathcal{N}(x|\mu,\sigma^2)$$

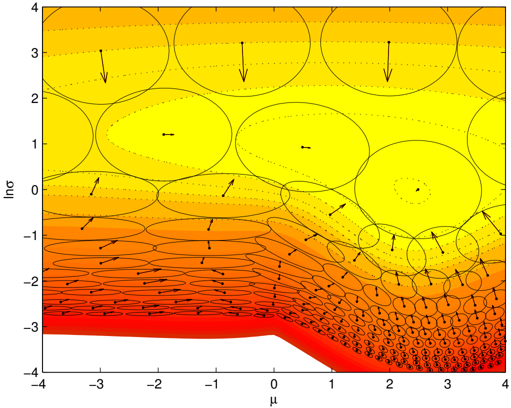


Figure: Arrows correspond to the gradients and ellipses to the inverse metric tensor. Dashed lines are isocontours of the joint log density

Gaussian Mixture Model

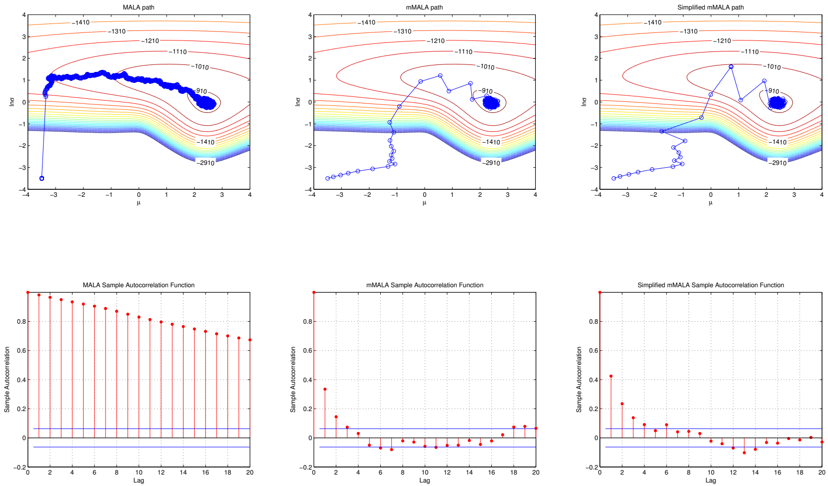


Figure: Comparison of MALA (left), mMALA (middle) and simplified mMALA (right) convergence paths and autocorrelation plots. Autocorrelation plots are from the stationary chains, i.e. once the chains have converged to the stationary distribution.

Gaussian Mixture Model

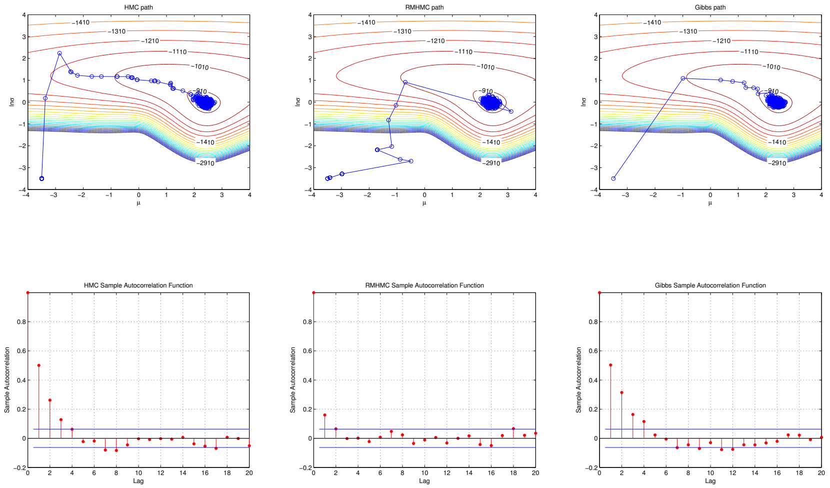


Figure: Comparison of HMC (left), RMHMC (middle) and GIBBS (right) convergence paths and autocorrelation plots. Autocorrelation plots are from the stationary chains, i.e. once the chains have converged to the stationary distribution.

Log-Gaussian Cox Point Process with Latent Field

- The joint density for Poisson counts and latent Gaussian field

$$p(\mathbf{y}, \mathbf{x} | \mu, \sigma, \beta) \propto \prod_{i,j}^{64} \exp\{y_{i,j} x_{i,j} - m \exp(x_{i,j})\} \exp(-(\mathbf{x} - \mu \mathbf{1})^{\mathsf{T}} \Sigma_{\theta}^{-1} (\mathbf{x} - \mu \mathbf{1})/2)$$

Log-Gaussian Cox Point Process with Latent Field

- The joint density for Poisson counts and latent Gaussian field

$$p(\mathbf{y}, \mathbf{x}|\mu, \sigma, \beta) \propto \prod_{i,j}^{64} \exp\{y_{i,j}x_{i,j} - m\exp(x_{i,j})\} \exp(-(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}\Sigma_{\theta}^{-1}(\mathbf{x}-\mu\mathbf{1})/2)$$

- Metric tensors

$$\begin{aligned} \mathbf{G}(\boldsymbol{\theta})_{i,j} &= \frac{1}{2} trace\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial\theta_i}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial\theta_j}\right) \\ \mathbf{G}(\mathbf{x}) &= \Lambda + \Sigma_{\boldsymbol{\theta}}^{-1} \end{aligned}$$

where Λ is diagonal with elements $m\exp(\mu + (\Sigma_{\boldsymbol{\theta}})_{i,i})$

Log-Gaussian Cox Point Process with Latent Field

- The joint density for Poisson counts and latent Gaussian field

$$p(\mathbf{y},\mathbf{x}|\mu,\sigma,\beta) \propto \prod_{i,j}^{64} \exp\{y_{i,j}x_{i,j} - m\exp(x_{i,j})\}\exp(-(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}\Sigma_{\theta}^{-1}(\mathbf{x}-\mu\mathbf{1})/2)$$

- Metric tensors

$$\begin{aligned} \mathbf{G}(\boldsymbol{\theta})_{i,j} &= \frac{1}{2} trace\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial\theta_i}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial\theta_j}\right) \\ \mathbf{G}(\mathbf{x}) &= \Lambda + \Sigma_{\boldsymbol{\theta}}^{-1} \end{aligned}$$

where Λ is diagonal with elements $m\exp(\mu + (\Sigma_{\boldsymbol{\theta}})_{i,i})$

- Latent field metric tensor defining flat manifold is 4096×4096, $\mathcal{O}(N^3)$ obtained from parameter conditional

Log-Gaussian Cox Point Process with Latent Field

- The joint density for Poisson counts and latent Gaussian field

$$p(\mathbf{y}, \mathbf{x}|\mu, \sigma, \beta) \propto \prod_{i,j}^{64} \exp\{y_{i,j}x_{i,j} - m\exp(x_{i,j})\} \exp(-(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}\Sigma_{\theta}^{-1}(\mathbf{x}-\mu\mathbf{1})/2)$$

- Metric tensors

$$\begin{aligned} \mathbf{G}(\boldsymbol{\theta})_{i,j} &= \frac{1}{2} trace\left(\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \mathbf{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \mathbf{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_j}\right) \\ \mathbf{G}(\mathbf{x}) &= \Lambda + \Sigma_{\boldsymbol{\theta}}^{-1} \end{aligned}$$

where Λ is diagonal with elements $m\exp(\mu + (\Sigma_{\boldsymbol{\theta}})_{i,i})$

- Latent field metric tensor defining flat manifold is 4096×4096, $\mathcal{O}(N^3)$ obtained from parameter conditional
- MALA requires transformation of latent field to sample - with separate tuning in transient and stationary phases of Markov chain
- Manifold methods requires no pilot tuning or additional transformations

RMHMC for Log-Gaussian Cox Point Processes

Table: Sampling the latent variables of a Log-Gaussian Cox Process - Comparison of sampling methods

Method	Time	ESS (Min, Med, Max)	s/Min ESS	Rel. Speed
MALA (Transient)	31,577	(3, 8, 50)	10,605	$\times 1$
MALA (Stationary)	31,118	(4, 16, 80)	7836	$\times 1.35$
mMALA	634	(26, 84, 174)	24.1	$\times 440$
RMHMC	2936	(1951, 4545, 5000)	1.5	$\times 7070$

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models
  - Promising capability of methodology

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models
  - Promising capability of methodology
- Ongoing development

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models
  - Promising capability of methodology
- Ongoing development
  - Potential bottleneck at metric tensor and Christoffel symbols

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models
  - Promising capability of methodology
- Ongoing development
  - Potential bottleneck at metric tensor and Christoffel symbols
  - Theoretical analysis of convergence

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models
  - Promising capability of methodology
- Ongoing development
  - Potential bottleneck at metric tensor and Christoffel symbols
  - Theoretical analysis of convergence
  - Investigate alternative manifold structures

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models
  - Promising capability of methodology
- Ongoing development
  - Potential bottleneck at metric tensor and Christoffel symbols
  - Theoretical analysis of convergence
  - Investigate alternative manifold structures
  - Design and effect of metric

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models
  - Promising capability of methodology
- Ongoing development
  - Potential bottleneck at metric tensor and Christoffel symbols
  - Theoretical analysis of convergence
  - Investigate alternative manifold structures
  - Design and effect of metric
  - Optimality of Hamiltonian flows as local geodesics

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models
  - Promising capability of methodology
- Ongoing development
  - Potential bottleneck at metric tensor and Christoffel symbols
  - Theoretical analysis of convergence
  - Investigate alternative manifold structures
  - Design and effect of metric
  - Optimality of Hamiltonian flows as local geodesics
  - Alternative transition kernels

Conclusion and Discussion

- Geometry of statistical models harnessed in Monte Carlo methods
  - Diffusions that respect structure and curvature of space - Manifold MALA
  - Geodesic flows on model manifold - RMHMC - generalisation of HMC
  - Assessed on correlated & high-dimensional latent variable models
  - Promising capability of methodology
- Ongoing development
  - Potential bottleneck at metric tensor and Christoffel symbols
  - Theoretical analysis of convergence
  - Investigate alternative manifold structures
  - Design and effect of metric
  - Optimality of Hamiltonian flows as local geodesics
  - Alternative transition kernels
- No silver bullet or cure all - new powerful methodology for MC toolkit

Funding Acknowledgment

- Girolami funded by EPSRC Advanced Research Fellowship and BBSRC project grant
- Calderhead supported by Microsoft Research PhD Scholarship