Introduction to the Read Paper

Young Statisticians Section

Mark Girolami

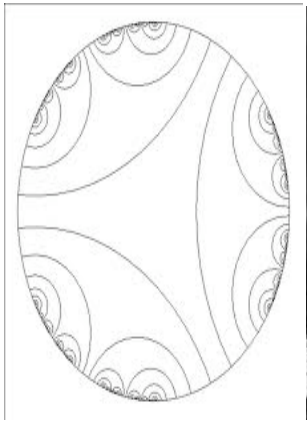
Department of Statistical Science
University College London

The Royal Statistical Society
Errol Street
London

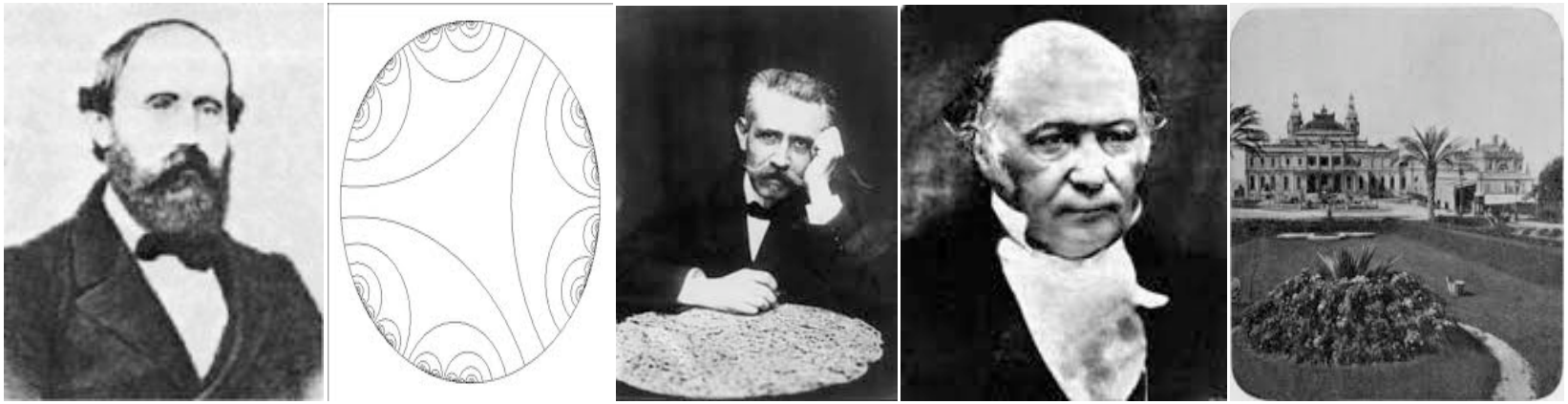
October 13, 2010

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

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

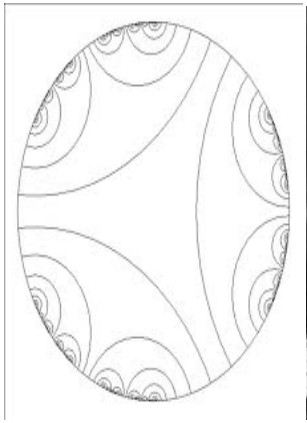




Riemann manifold Langevin and Hamiltonian Monte Carlo Methods, Girolami, M. & Calderhead, B., *J.R.Statist. Soc.* B (2011), **73**, *Part 2*



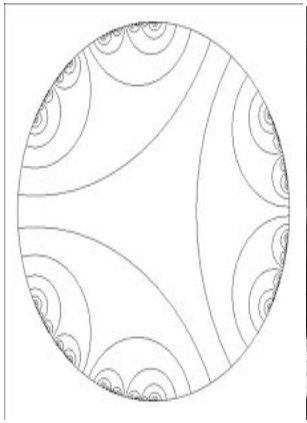
- Advancing MC methods via underlying geometry of fundamental objects

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

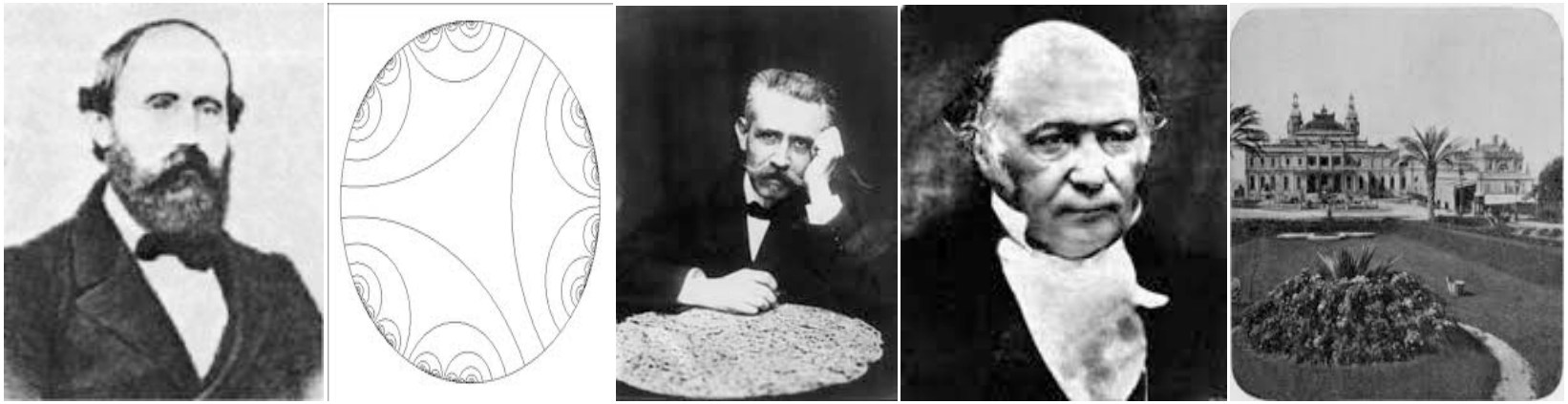

- Advancing MC methods via underlying geometry of fundamental objects
- Develop proposal mechanisms based on
  - Stochastic diffusions on Riemann manifold

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


- Advancing MC methods via underlying geometry of fundamental objects
- Develop proposal mechanisms based on
  - Stochastic diffusions on Riemann manifold
  - Deterministic mechanics on Riemann manifold

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



- Advancing MC methods via underlying geometry of fundamental objects
- Develop proposal mechanisms based on
  - Stochastic diffusions on Riemann manifold
  - Deterministic mechanics on Riemann manifold
- Focus on Hamiltonian Monte Carlo for next 27 minutes

Hamiltonian Monte Carlo for Computational Statistical Inference

- Target density $p(\theta)$, introduce auxiliary variable $\mathbf{p} \sim p(\mathbf{p}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$.

Hamiltonian Monte Carlo for Computational Statistical Inference

- Target density $p(\theta)$, introduce auxiliary variable $\mathbf{p} \sim p(\mathbf{p}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$.
- Negative log-density $\mathcal{L}(\theta) \equiv \log p(\theta)$, then

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

Hamiltonian Monte Carlo for Computational Statistical Inference

- Target density $p(\theta)$, introduce auxiliary variable $\mathbf{p} \sim p(\mathbf{p}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$.
- Negative log-density $\mathcal{L}(\theta) \equiv \log p(\theta)$, then

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

- Interpreted as separable Hamiltonian in position & momentum variables

$$\frac{d\theta}{d\tau} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \qquad \frac{d\mathbf{p}}{d\tau} = -\frac{\partial H}{\partial \theta} = \nabla_{\theta}\mathcal{L}(\theta)$$

Hamiltonian Monte Carlo for Computational Statistical Inference

- Target density $p(\theta)$, introduce auxiliary variable $\mathbf{p} \sim p(\mathbf{p}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$.
- Negative log-density $\mathcal{L}(\theta) \equiv \log p(\theta)$, then

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

- Interpreted as separable Hamiltonian in position & momentum variables

$$\frac{d\theta}{d\tau} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \qquad \frac{d\mathbf{p}}{d\tau} = -\frac{\partial H}{\partial \theta} = \nabla_\theta \mathcal{L}(\theta)$$

- Energy (approximate), volume preserving and reversible integrator follows

$$\begin{aligned} \mathbf{p}(\tau + \epsilon/2) &= \mathbf{p}(\tau) + \epsilon\nabla_\theta \mathcal{L}(\theta(\tau))/2 \\ \theta(\tau + \epsilon) &= \theta(\tau) + \epsilon\mathbf{M}^{-1}\mathbf{p}(\tau + \epsilon/2) \\ \mathbf{p}(\tau + \epsilon) &= \mathbf{p}(\tau + \epsilon/2) + \epsilon\nabla_\theta \mathcal{L}(\theta(\tau + \epsilon))/2 \end{aligned}$$

Hamiltonian Monte Carlo for Computational Statistical Inference

- Target density $p(\theta)$, introduce auxiliary variable $\mathbf{p} \sim p(\mathbf{p}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$.
- Negative log-density $\mathcal{L}(\theta) \equiv \log p(\theta)$, then

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

- Interpreted as separable Hamiltonian in position & momentum variables

$$\frac{d\theta}{d\tau} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \qquad \frac{d\mathbf{p}}{d\tau} = -\frac{\partial H}{\partial \theta} = \nabla_\theta \mathcal{L}(\theta)$$

- Energy (approximate), volume preserving and reversible integrator follows

$$\begin{aligned} \mathbf{p}(\tau + \epsilon/2) &= \mathbf{p}(\tau) + \epsilon\nabla_\theta \mathcal{L}(\theta(\tau))/2 \\ \theta(\tau + \epsilon) &= \theta(\tau) + \epsilon\mathbf{M}^{-1}\mathbf{p}(\tau + \epsilon/2) \\ \mathbf{p}(\tau + \epsilon) &= \mathbf{p}(\tau + \epsilon/2) + \epsilon\nabla_\theta \mathcal{L}(\theta(\tau + \epsilon))/2 \end{aligned}$$

- Detailed balance satisfied by $\min\{1, \exp\{-H(\theta^*, \mathbf{p}^*) + H(\theta, \mathbf{p})\}\}$

Hamiltonian Monte Carlo for Computational Statistical Inference

- Target density $p(\theta)$, introduce auxiliary variable $\mathbf{p} \sim p(\mathbf{p}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$.
- Negative log-density $\mathcal{L}(\theta) \equiv \log p(\theta)$, then

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

- Interpreted as separable Hamiltonian in position & momentum variables

$$\frac{d\theta}{d\tau} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \qquad \frac{d\mathbf{p}}{d\tau} = -\frac{\partial H}{\partial \theta} = \nabla_\theta \mathcal{L}(\theta)$$

- Energy (approximate), volume preserving and reversible integrator follows

$$\begin{aligned} \mathbf{p}(\tau + \epsilon/2) &= \mathbf{p}(\tau) + \epsilon\nabla_\theta \mathcal{L}(\theta(\tau))/2 \\ \theta(\tau + \epsilon) &= \theta(\tau) + \epsilon\mathbf{M}^{-1}\mathbf{p}(\tau + \epsilon/2) \\ \mathbf{p}(\tau + \epsilon) &= \mathbf{p}(\tau + \epsilon/2) + \epsilon\nabla_\theta \mathcal{L}(\theta(\tau + \epsilon))/2 \end{aligned}$$

- Detailed balance satisfied by $\min\{1, \exp\{-H(\theta^*, \mathbf{p}^*) + H(\theta, \mathbf{p})\}\}$
- The complete method to sample from the desired marginal $p(\theta)$ follows

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

Hamiltonian Monte Carlo for Computational Statistical Inference

- Target density $p(\theta)$, introduce auxiliary variable $\mathbf{p} \sim p(\mathbf{p}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$.
- Negative log-density $\mathcal{L}(\theta) \equiv \log p(\theta)$, then

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

- Interpreted as separable Hamiltonian in position & momentum variables

$$\frac{d\theta}{d\tau} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \qquad \frac{d\mathbf{p}}{d\tau} = -\frac{\partial H}{\partial \theta} = \nabla_\theta \mathcal{L}(\theta)$$

- Energy (approximate), volume preserving and reversible integrator follows

$$\begin{aligned} \mathbf{p}(\tau + \epsilon/2) &= \mathbf{p}(\tau) + \epsilon\nabla_\theta \mathcal{L}(\theta(\tau))/2 \\ \theta(\tau + \epsilon) &= \theta(\tau) + \epsilon\mathbf{M}^{-1}\mathbf{p}(\tau + \epsilon/2) \\ \mathbf{p}(\tau + \epsilon) &= \mathbf{p}(\tau + \epsilon/2) + \epsilon\nabla_\theta \mathcal{L}(\theta(\tau + \epsilon))/2 \end{aligned}$$

- Detailed balance satisfied by $\min\{1, \exp\{-H(\theta^*, \mathbf{p}^*) + H(\theta, \mathbf{p})\}\}$
- The complete method to sample from the desired marginal $p(\theta)$ follows

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

- Integrator provides proposals for $p(\theta|\mathbf{p})$ conditional

Illustrative Example - Bivariate Gaussian

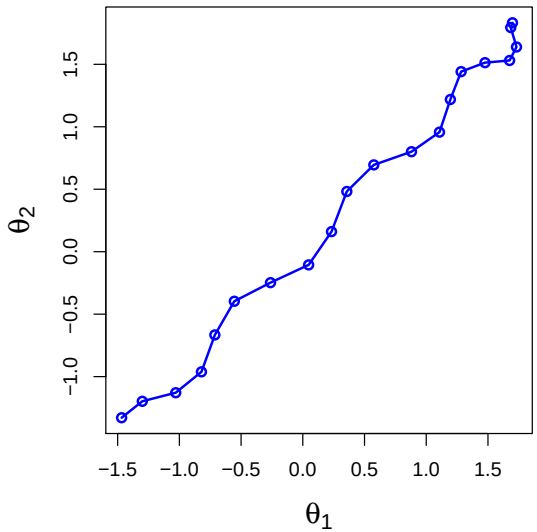
- Target density $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where

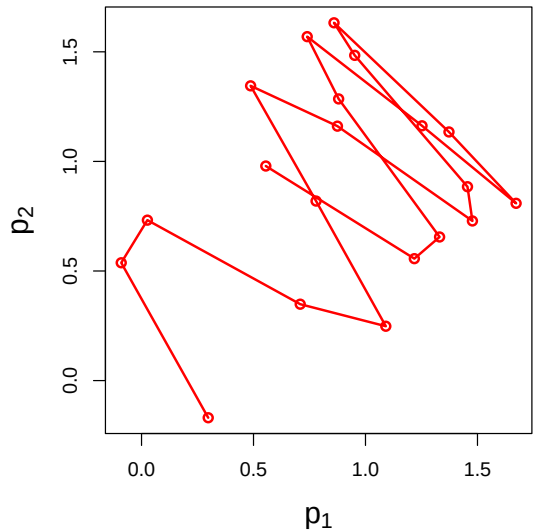
$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

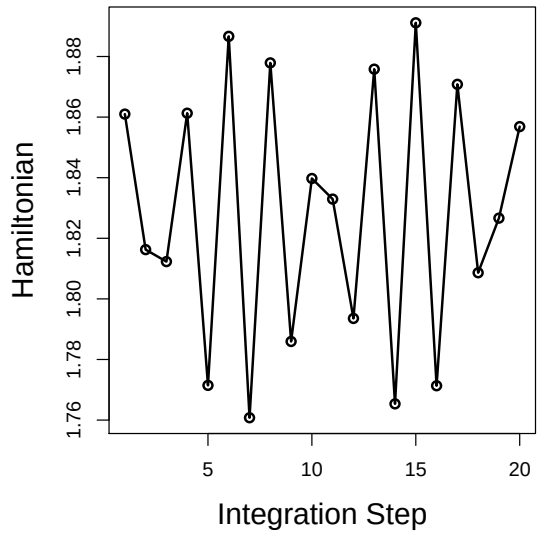
- For ρ large e.g. 0.98 sampling from this distribution is challenging
- Overall Hamiltonian

$$\frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{\Sigma}^{-1}\mathbf{x} + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{p}$$

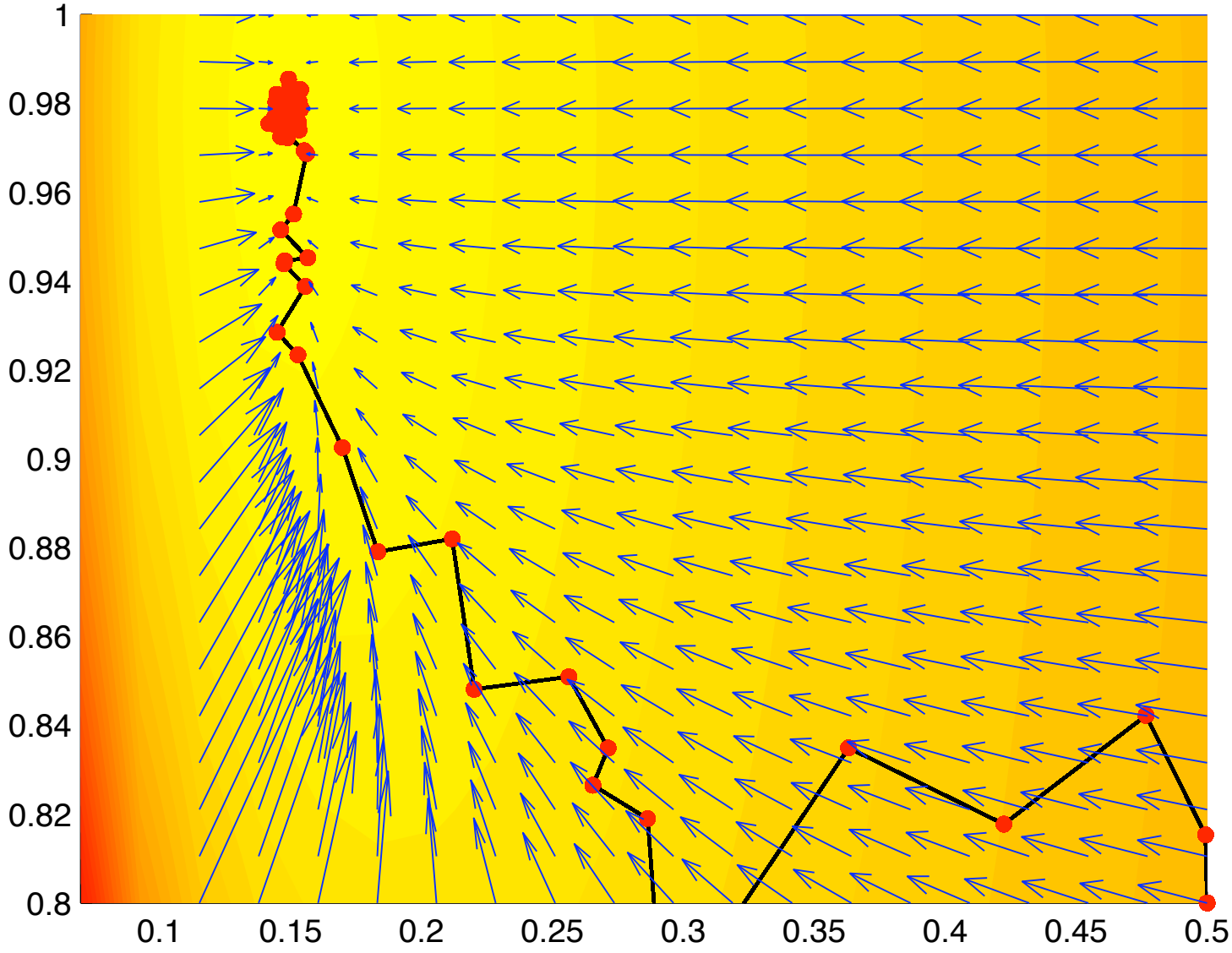
HMC Integration $\epsilon = 0.18$, $L = 20$, implicit identity matrix for metric



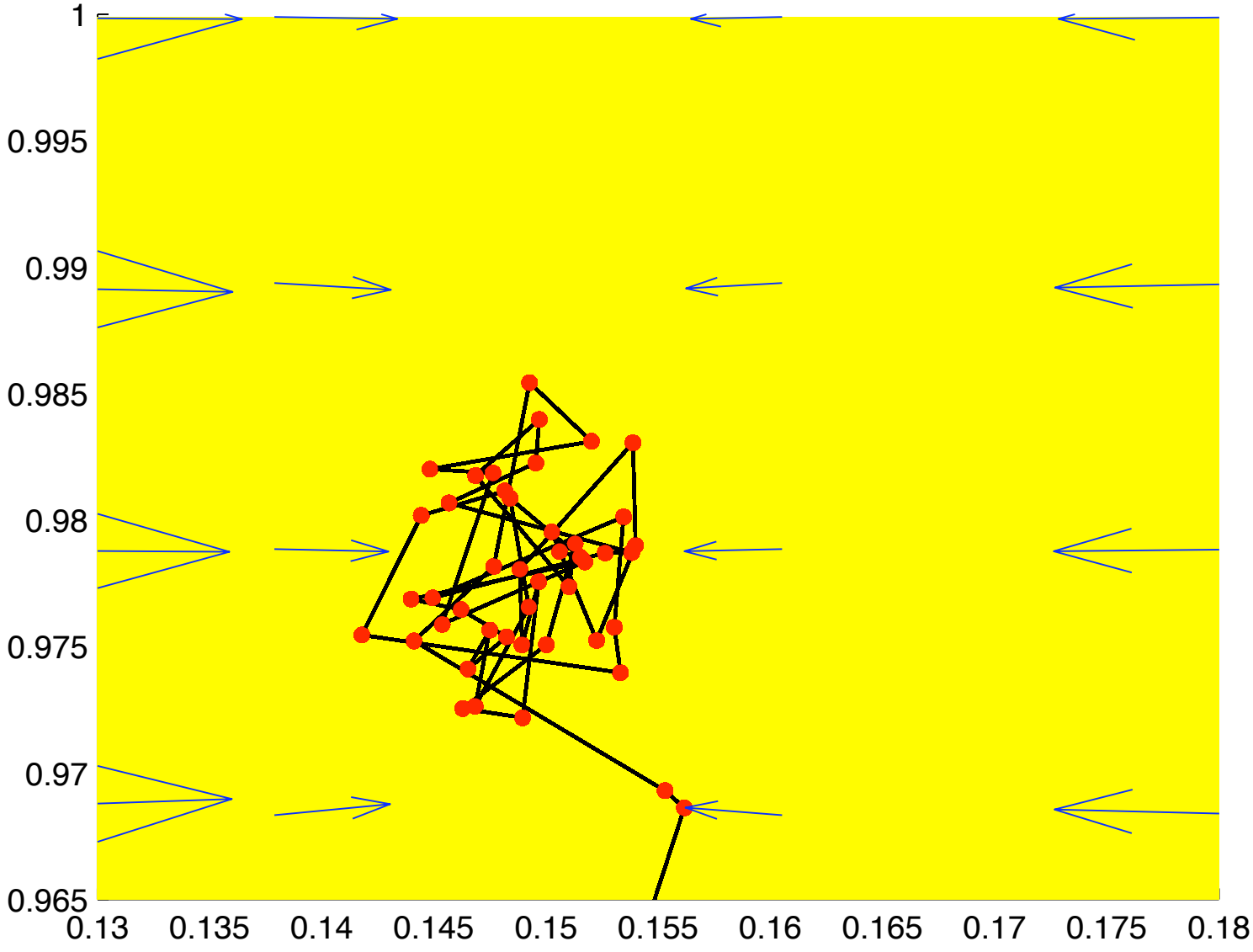




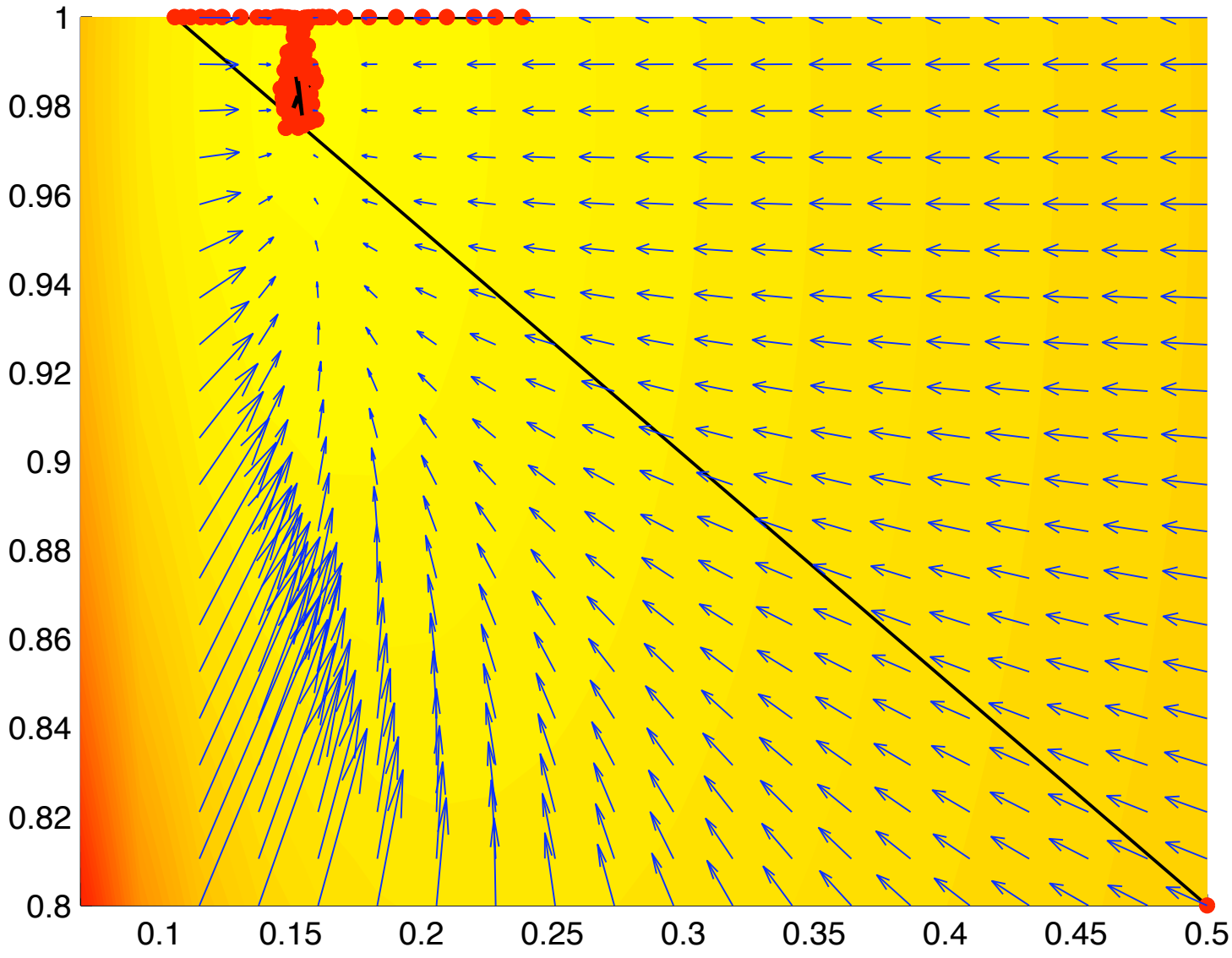
Metropolis Algorithm, Parameters of Stoch Vol Model, Acc Rate 25%



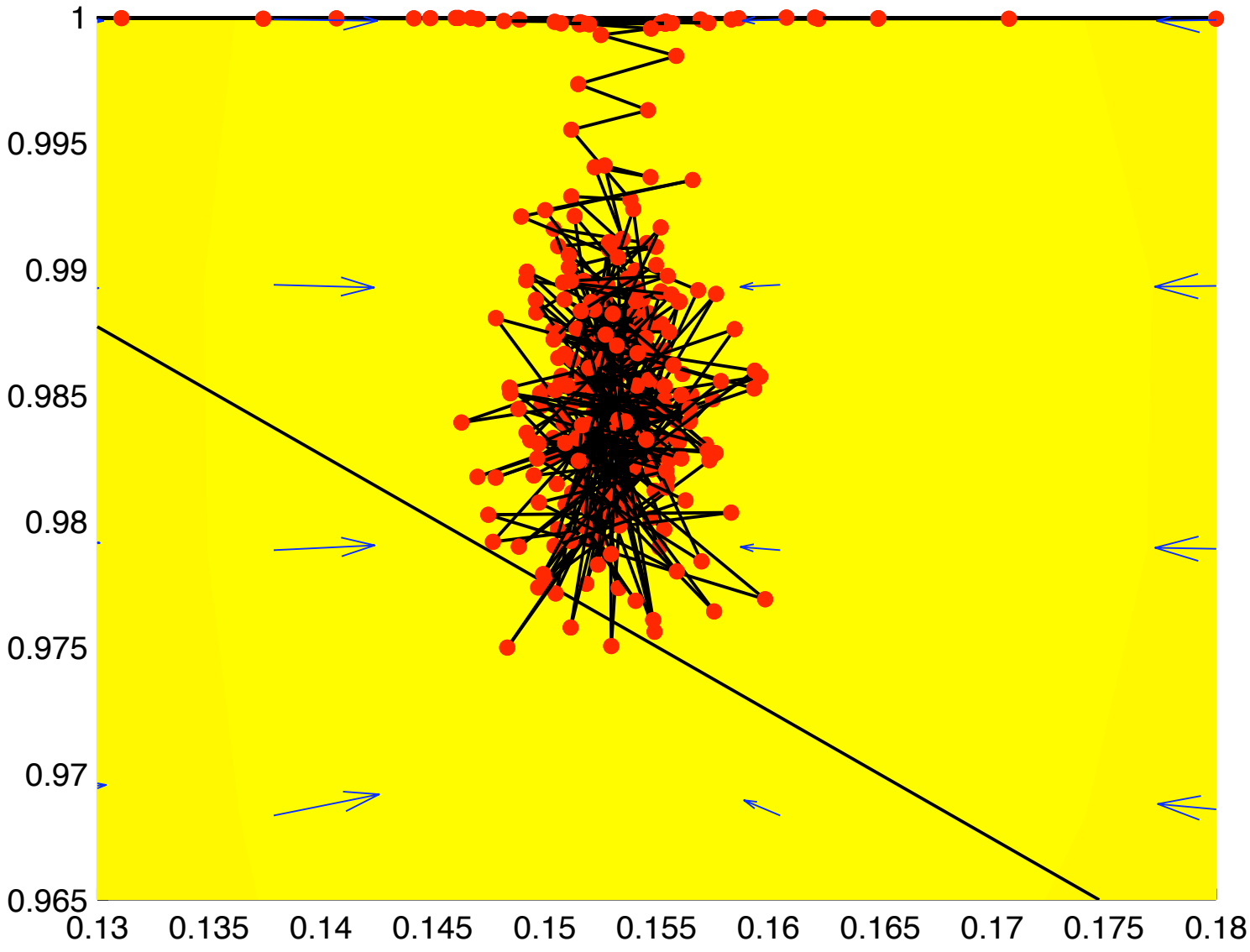
Metropolis Algorithm, Parameters of Stoch Vol Model, Acc Rate 25%



HMC Algorithm, Parameters of Stoch Vol Model, Acc Rate 95%



HMC Algorithm, Parameters of Stoch Vol Model, Acc Rate 95%



Hamiltonian Monte Carlo for Posterior Inference

- Deterministic proposal for θ ensures greater efficiency over Metropolis random walk

Hamiltonian Monte Carlo for Posterior Inference

- Deterministic proposal for θ ensures greater efficiency over Metropolis random walk
- Small fly in the ointment - tuning of values of matrix **M** essential for efficient performance of HMC

Hamiltonian Monte Carlo for Posterior Inference

- Deterministic proposal for θ ensures greater efficiency over Metropolis random walk
- Small fly in the ointment - tuning of values of matrix **M** essential for efficient performance of HMC
- Diagonal elements of **M** reflect scale and off-diagonal elements capture correlation structure of target - (no off-diagonal terms in physical interpretation)

Hamiltonian Monte Carlo for Posterior Inference

- Deterministic proposal for θ ensures greater efficiency over Metropolis random walk
- Small fly in the ointment - tuning of values of matrix **M** essential for efficient performance of HMC
- Diagonal elements of **M** reflect scale and off-diagonal elements capture correlation structure of target - (no off-diagonal terms in physical interpretation)
- Require knowledge of target density to set **M** - this requires extensive tuning via pilot runs of sampler

Hamiltonian Monte Carlo for Posterior Inference

- Deterministic proposal for θ ensures greater efficiency over Metropolis random walk
- Small fly in the ointment - tuning of values of matrix **M** essential for efficient performance of HMC
- Diagonal elements of **M** reflect scale and off-diagonal elements capture correlation structure of target - (no off-diagonal terms in physical interpretation)
- Require knowledge of target density to set **M** - this requires extensive tuning via pilot runs of sampler
- Common reason given for lack of HMC take-up in non-trivial applications - e.g. see recent discussion on Gelman blog

Hamiltonian Monte Carlo for Posterior Inference

- Deterministic proposal for θ ensures greater efficiency over Metropolis random walk
- Small fly in the ointment - tuning of values of matrix **M** essential for efficient performance of HMC
- Diagonal elements of **M** reflect scale and off-diagonal elements capture correlation structure of target - (no off-diagonal terms in physical interpretation)
- Require knowledge of target density to set **M** - this requires extensive tuning via pilot runs of sampler
- Common reason given for lack of HMC take-up in non-trivial applications - e.g. see recent discussion on Gelman blog
- Can this weakness be resolved?

Geometric Concepts in MCMC

- What is the distance between probabilities?

Geometric Concepts in MCMC

- What is the distance between probabilities?
- Rao, 1945, distance between $p(\mathbf{y}; \theta + \delta\theta)$ and $p(\mathbf{y}; \theta)$ follows as

Geometric Concepts in MCMC

- What is the distance between probabilities?
- Rao, 1945, distance between $p(\mathbf{y}; \boldsymbol{\theta} + \delta\boldsymbol{\theta})$ and $p(\mathbf{y}; \boldsymbol{\theta})$ follows as

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

Geometric Concepts in MCMC

- What is the distance between probabilities?
- Rao, 1945, distance between $p(\mathbf{y}; \boldsymbol{\theta} + \delta\boldsymbol{\theta})$ and $p(\mathbf{y}; \boldsymbol{\theta})$ follows as

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

- Rao, 1945, noted that $\mathbf{G}(\boldsymbol{\theta})$ is positive definite - defines a Riemann manifold - obtain complete non-Euclidean geometry for probabilities - distances, metrics, invariants, curvature, geodesics

Geometric Concepts in MCMC

- What is the distance between probabilities?
- Rao, 1945, distance between $p(\mathbf{y}; \theta + \delta\theta)$ and $p(\mathbf{y}; \theta)$ follows as

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

- Rao, 1945, noted that $\mathbf{G}(\theta)$ is positive definite - defines a Riemann manifold - obtain complete non-Euclidean geometry for probabilities - distances, metrics, invariants, curvature, geodesics
- Can this geometric structure also be employed in addressing problems with MCMC in complex inference problems?

Geometric Concepts in MCMC

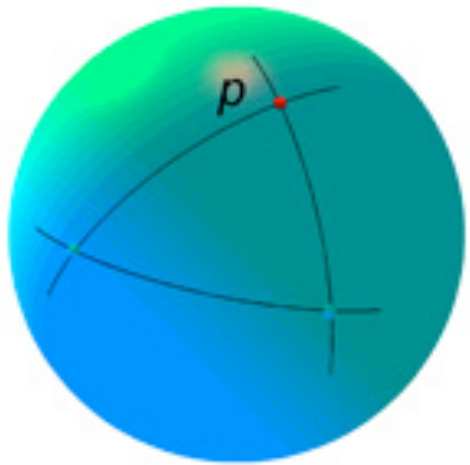
- What is the distance between probabilities?
- Rao, 1945, distance between $p(\mathbf{y}; \theta + \delta\theta)$ and $p(\mathbf{y}; \theta)$ follows as

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

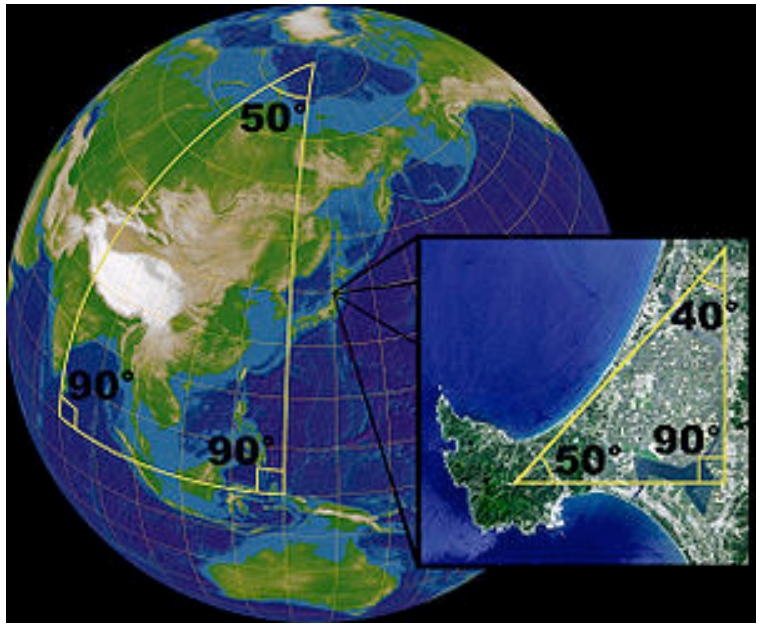
- Rao, 1945, noted that $\mathbf{G}(\theta)$ is positive definite - defines a Riemann manifold - obtain complete non-Euclidean geometry for probabilities - distances, metrics, invariants, curvature, geodesics
- Can this geometric structure also be employed in addressing problems with MCMC in complex inference problems?

Riemann Manifold

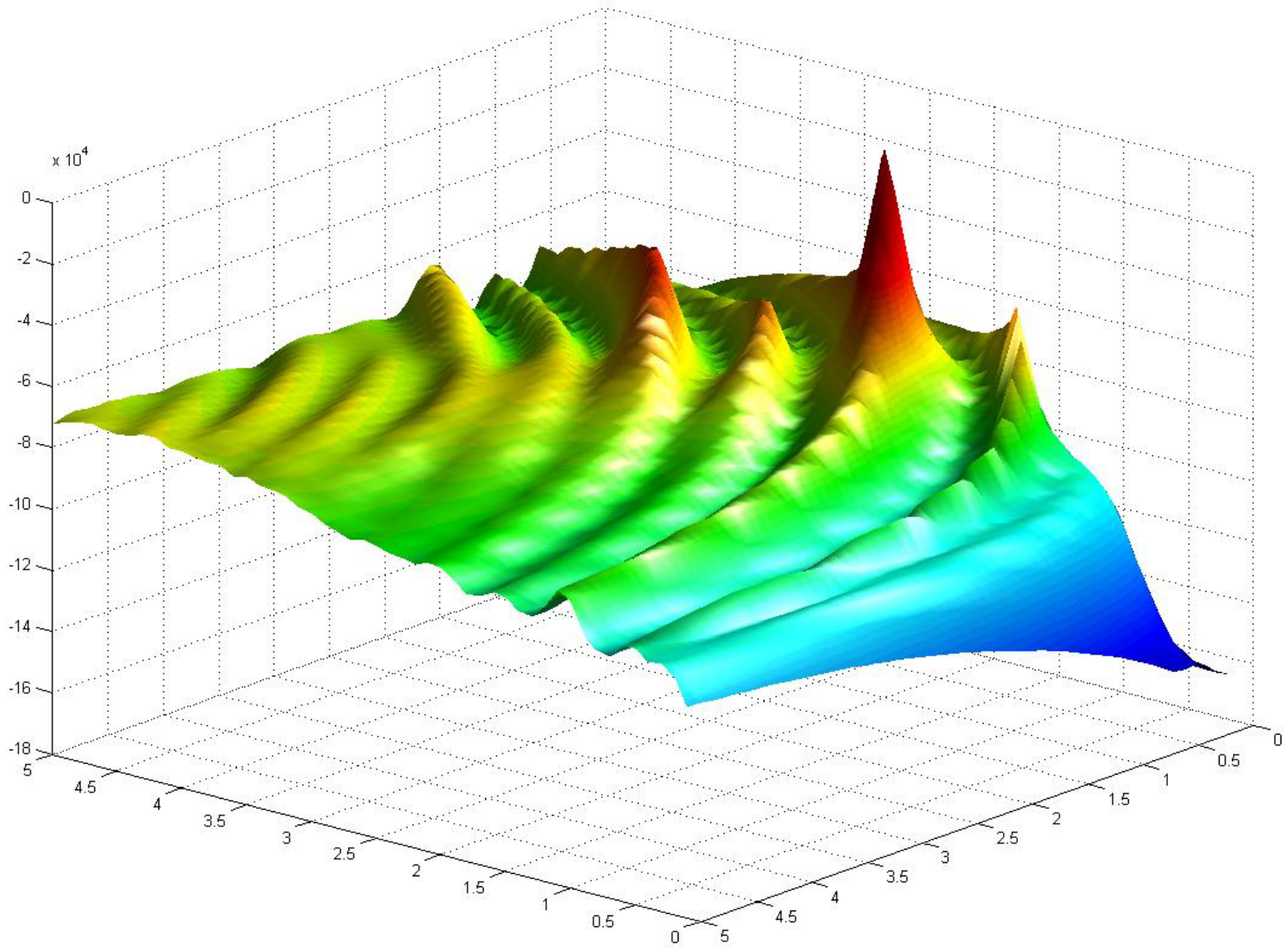
Riemann Manifold



Riemann Manifold



Manifold Structure in Density



Riemannian Hamiltonian Monte Carlo

- Manifold defined by metric $\mathbf{G}(\theta)$ hence Hamiltonian kinetic energy defined in terms of contraviant form i.e. $\dot{\theta}^{\mathsf{T}}\mathbf{G}(\theta)\dot{\theta} = \mathbf{p}^{\mathsf{T}}\mathbf{G}^{-1}(\theta)\mathbf{p}$

Riemannian Hamiltonian Monte Carlo

- Manifold defined by metric $\mathbf{G}(\theta)$ hence Hamiltonian kinetic energy defined in terms of contraviant form i.e. $\dot{\theta}^{\mathsf{T}}\mathbf{G}(\theta)\dot{\theta} = \mathbf{p}^{\mathsf{T}}\mathbf{G}^{-1}(\theta)\mathbf{p}$

Riemannian Hamiltonian Monte Carlo

- Manifold defined by metric $\mathbf{G}(\theta)$ hence Hamiltonian kinetic energy defined in terms of contraviant form i.e. $\dot{\theta}^{\mathsf{T}}\mathbf{G}(\theta)\dot{\theta} = \mathbf{p}^{\mathsf{T}}\mathbf{G}^{-1}(\theta)\mathbf{p}$
- Hamiltonian defined on Riemann manifold is non-separable

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

Riemannian Hamiltonian Monte Carlo

- Manifold defined by metric $\mathbf{G}(\theta)$ hence Hamiltonian kinetic energy defined in terms of contraviant form i.e. $\dot{\theta}^{\mathsf{T}}\mathbf{G}(\theta)\dot{\theta} = \mathbf{p}^{\mathsf{T}}\mathbf{G}^{-1}(\theta)\mathbf{p}$
- Hamiltonian defined on Riemann manifold is non-separable

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

- Hamiltonian in HMC artificially imposes a position independent metric tensor, $\mathbf{M}$, defining a flat manifold, upon the statistical model

Riemannian Hamiltonian Monte Carlo

- Manifold defined by metric $\mathbf{G}(\theta)$ hence Hamiltonian kinetic energy defined in terms of contraviant form i.e. $\dot{\theta}^{\mathsf{T}}\mathbf{G}(\theta)\dot{\theta} = \mathbf{p}^{\mathsf{T}}\mathbf{G}^{-1}(\theta)\mathbf{p}$
- Hamiltonian defined on Riemann manifold is non-separable

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

- Hamiltonian in HMC artificially imposes a position independent metric tensor, $\mathbf{M}$, defining a flat manifold, upon the statistical model
- Marginal density follows as required

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

Riemannian Hamiltonian Monte Carlo

- Manifold defined by metric $\mathbf{G}(\theta)$ hence Hamiltonian kinetic energy defined in terms of contraviant form i.e. $\dot{\theta}^{\mathsf{T}}\mathbf{G}(\theta)\dot{\theta} = \mathbf{p}^{\mathsf{T}}\mathbf{G}^{-1}(\theta)\mathbf{p}$
- Hamiltonian defined on Riemann manifold is non-separable

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

- Hamiltonian in HMC artificially imposes a position independent metric tensor, $\mathbf{M}$, defining a flat manifold, upon the statistical model
- Marginal density follows as required

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

- Complete sampler follows as

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

Riemannian Hamiltonian Monte Carlo

- Manifold defined by metric $\mathbf{G}(\theta)$ hence Hamiltonian kinetic energy defined in terms of contraviant form i.e. $\dot{\theta}^{\mathsf{T}}\mathbf{G}(\theta)\dot{\theta} = \mathbf{p}^{\mathsf{T}}\mathbf{G}^{-1}(\theta)\mathbf{p}$
- Hamiltonian defined on Riemann manifold is non-separable

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

- Hamiltonian in HMC artificially imposes a position independent metric tensor, $\mathbf{M}$, defining a flat manifold, upon the statistical model
- Marginal density follows as required

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

- Complete sampler follows as

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

Riemannian Hamiltonian Monte Carlo

- The dynamics of the non-separable Hamiltonian follow as

$$\begin{aligned}
\frac{d\theta_i}{d\tau} &= \frac{\partial H}{\partial p_i} = \left(\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\right)_i \\
\frac{dp_i}{d\tau} &= -\frac{\partial H}{\partial \theta_i} = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} - \frac{1}{2}\mathrm{Tr}\left[\mathbf{G}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{G}(\boldsymbol{\theta})}{\partial \theta_i}\right] + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\frac{\partial \mathbf{G}^{-1}(\boldsymbol{\theta})}{\partial \theta_i}\mathbf{p}
\end{aligned}$$

Riemannian Hamiltonian Monte Carlo

- The dynamics of the non-separable Hamiltonian follow as

$$\begin{aligned}
\frac{d\theta_i}{d\tau} &= \frac{\partial H}{\partial p_i} = \left(\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\right)_i \\
\frac{dp_i}{d\tau} &= -\frac{\partial H}{\partial \theta_i} = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} - \frac{1}{2}\mathrm{Tr}\left[\mathbf{G}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{G}(\boldsymbol{\theta})}{\partial \theta_i}\right] + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\frac{\partial \mathbf{G}^{-1}(\boldsymbol{\theta})}{\partial \theta_i}\mathbf{p}
\end{aligned}$$

- Require reversible, volume preserving integrator - 2'nd order semi-implicit integrator

Riemannian Hamiltonian Monte Carlo

- The dynamics of the non-separable Hamiltonian follow as

$$\begin{aligned}
\frac{d\theta_i}{d\tau} &= \frac{\partial H}{\partial p_i} = \left(\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\right)_i \\
\frac{dp_i}{d\tau} &= -\frac{\partial H}{\partial \theta_i} = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} - \frac{1}{2}\mathrm{Tr}\left[\mathbf{G}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{G}(\boldsymbol{\theta})}{\partial \theta_i}\right] + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\frac{\partial \mathbf{G}^{-1}(\boldsymbol{\theta})}{\partial \theta_i}\mathbf{p}
\end{aligned}$$

- Require reversible, volume preserving integrator - 2'nd order semi-implicit integrator

$$\begin{aligned}
\mathbf{p}^{\tau+\frac{\epsilon}{2}} &= \mathbf{p}^{n} - \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^{\tau}, \mathbf{p}^{\tau+\frac{\epsilon}{2}}) && (1)\\
\boldsymbol{\theta}^{\tau+\epsilon} &= \boldsymbol{\theta}^{\tau} + \frac{\epsilon}{2}\left[\nabla_{\mathbf{p}} H(\boldsymbol{\theta}^{\tau}, \mathbf{p}^{\tau+\frac{\epsilon}{2}}) + \nabla_{\mathbf{p}} H(\boldsymbol{\theta}^{\tau+\epsilon}, \mathbf{p}^{\tau+\frac{\epsilon}{2}})\right] && (2)\\
\mathbf{p}^{\tau+\epsilon} &= \mathbf{p}^{\tau+\frac{\epsilon}{2}} - \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^{\tau+\epsilon}, \mathbf{p}^{\tau+\frac{\epsilon}{2}}) && (3)
\end{aligned}$$

Step (3) is explicit so require to solve (1) and (2) for $\mathbf{p}^{n+\frac{1}{2}}$ and $\boldsymbol{\theta}^{n+1}$ using e.g. fixed point iteration found to be very efficient in practice.

Illustrative Example - Bivariate Gaussian

- Target density $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where

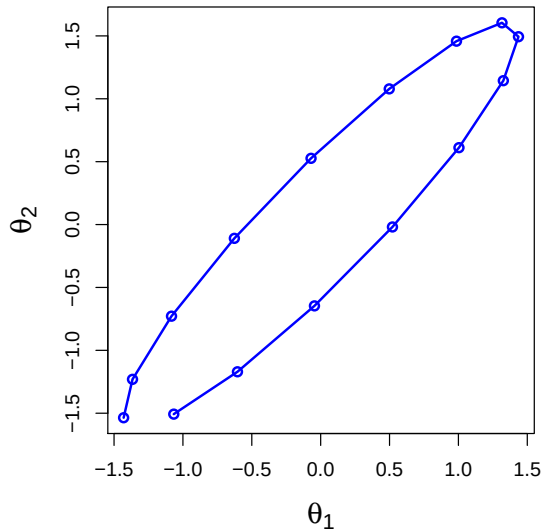
$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

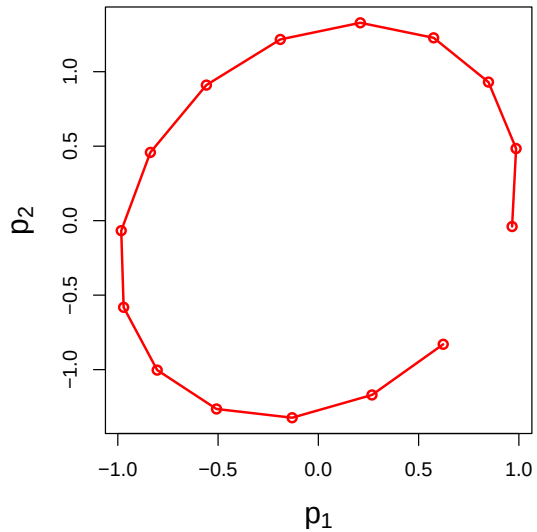
- For ρ large e.g. 0.98 sampling from this distribution is challenging
- Metric tensor defines flat manifold $\mathbf{\Sigma}^{-1}$
- Overall Hamiltonian

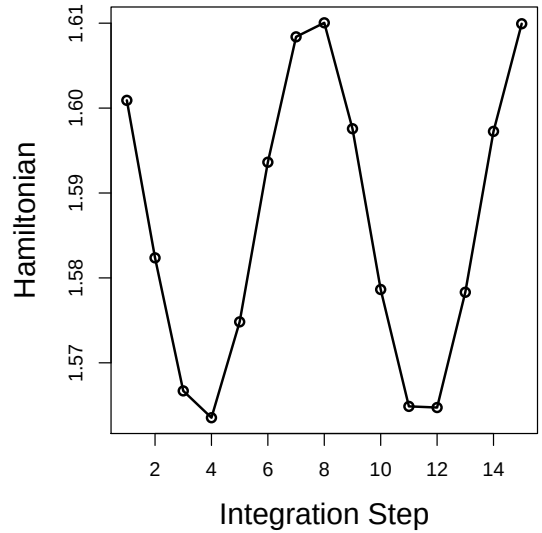
$$\frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{\Sigma}^{-1}\mathbf{x} + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{p}$$

- Stormer Verlet integrator is required

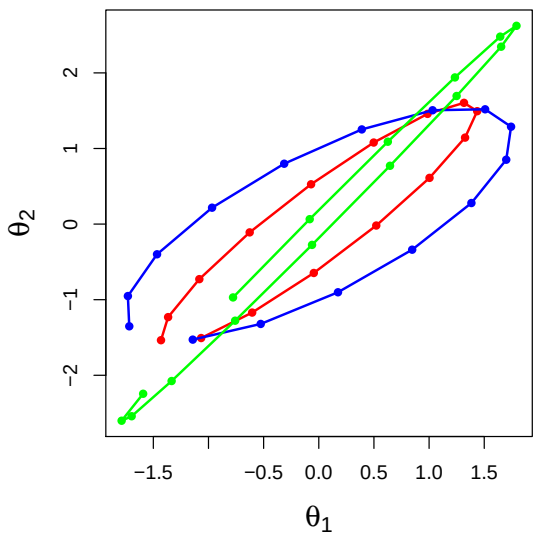
Illustrative Example - $\rho = 0.6$, RMHMC Integration $\epsilon = 0.8$, $L = 15$

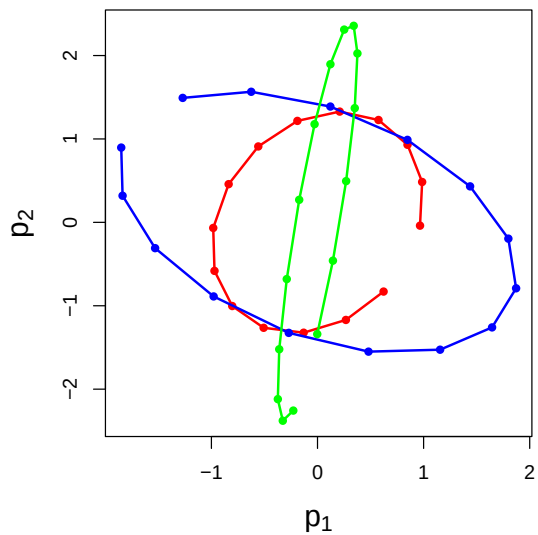


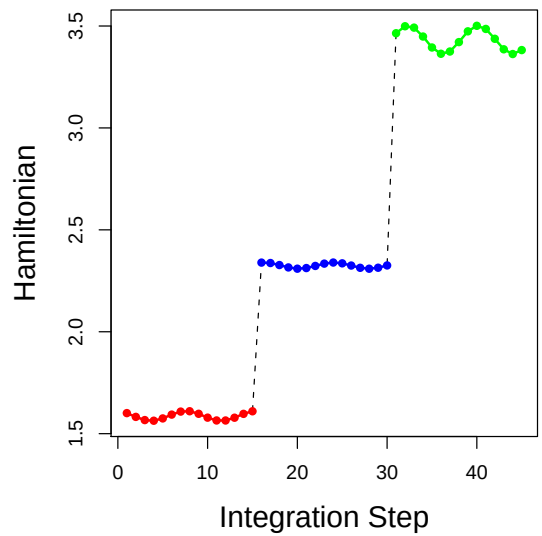




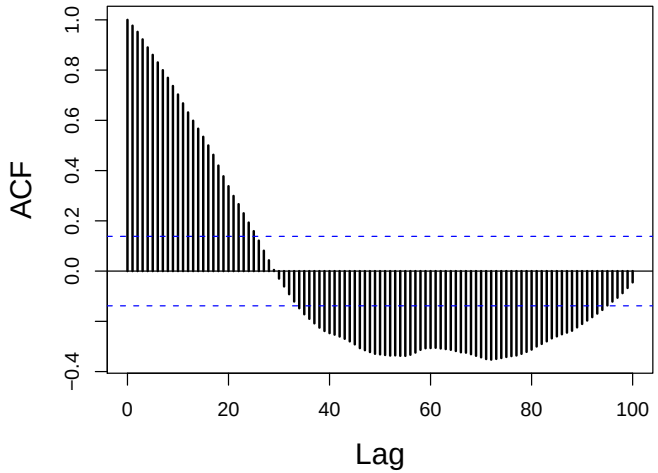
RMHMC Integration and 3 $\times$ Auxiliary Variable Sampling

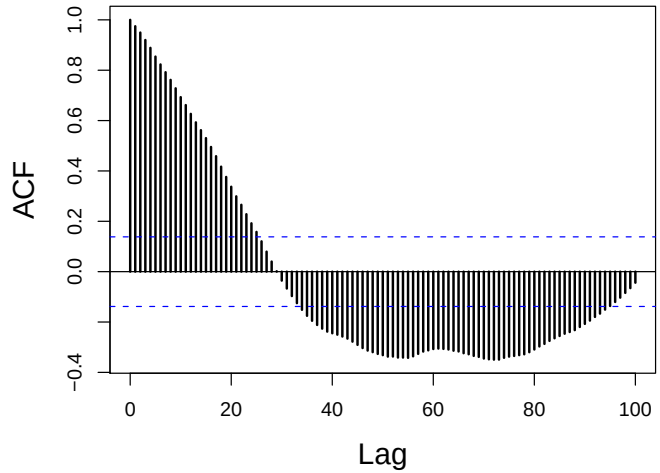


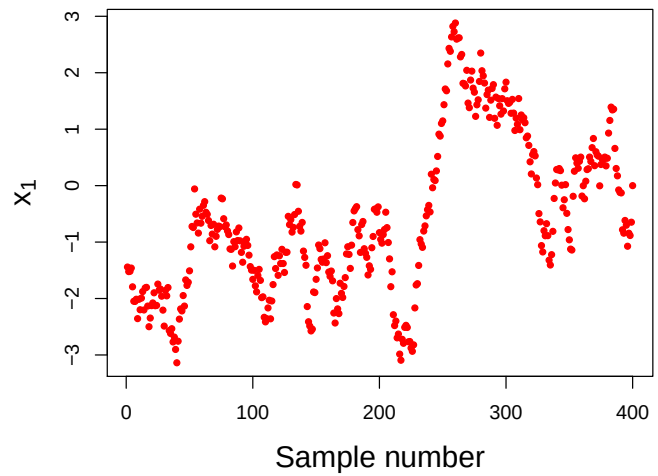


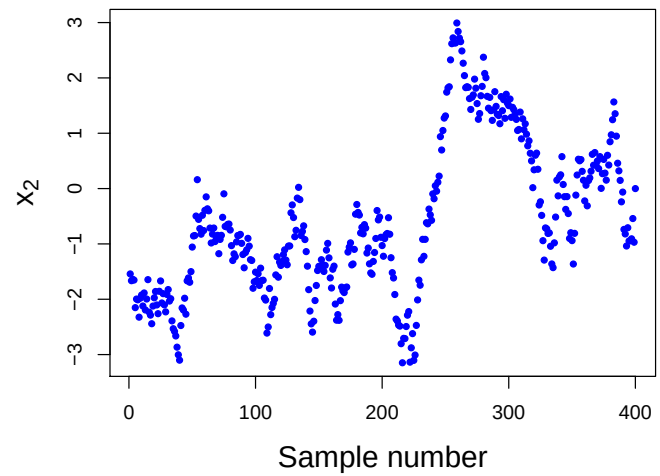


Gibbs Sampler Performance for $\rho = 0.98$

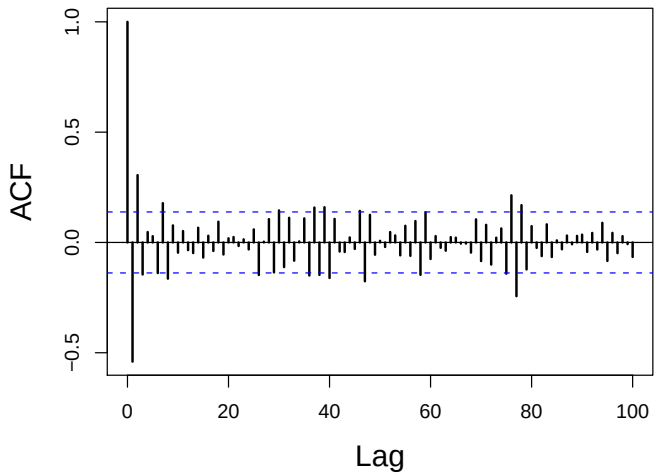


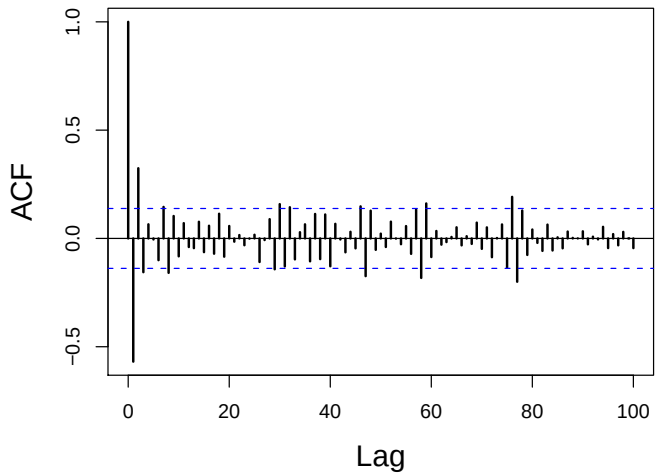


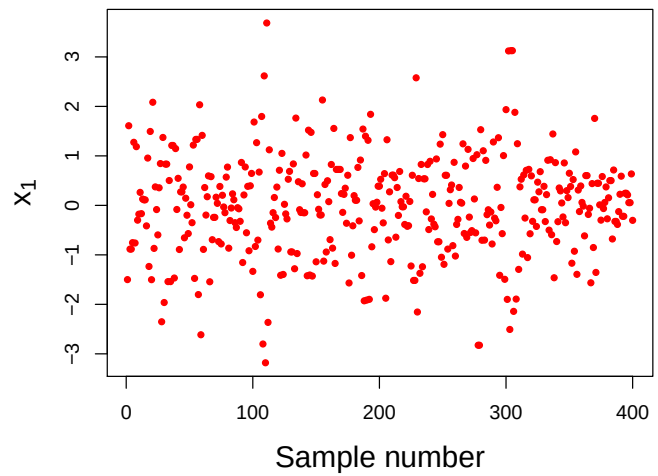


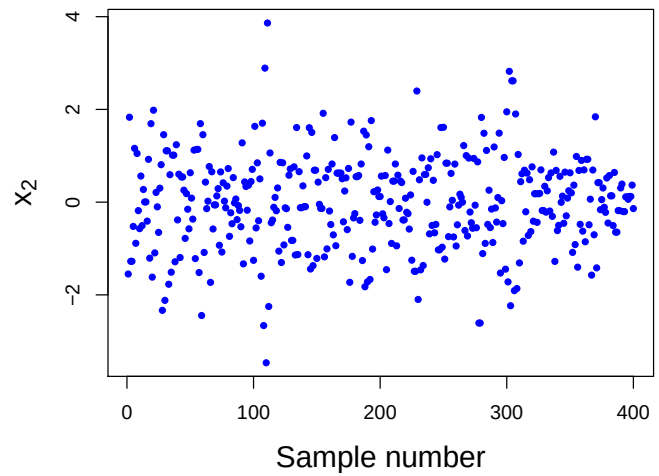


RMHMC Performance, $\rho = 0.98$, $\epsilon = 0.8$, $L = 15$

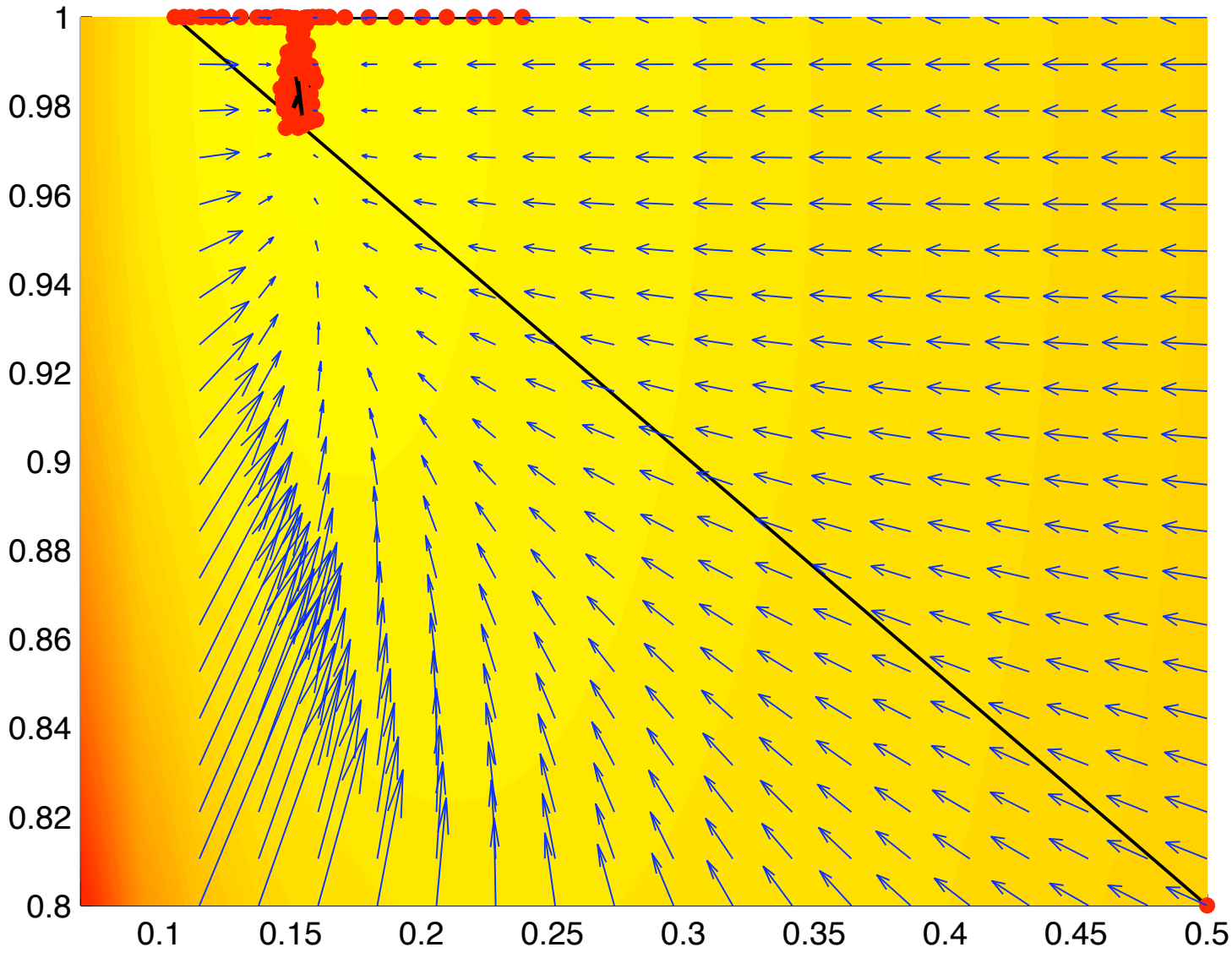




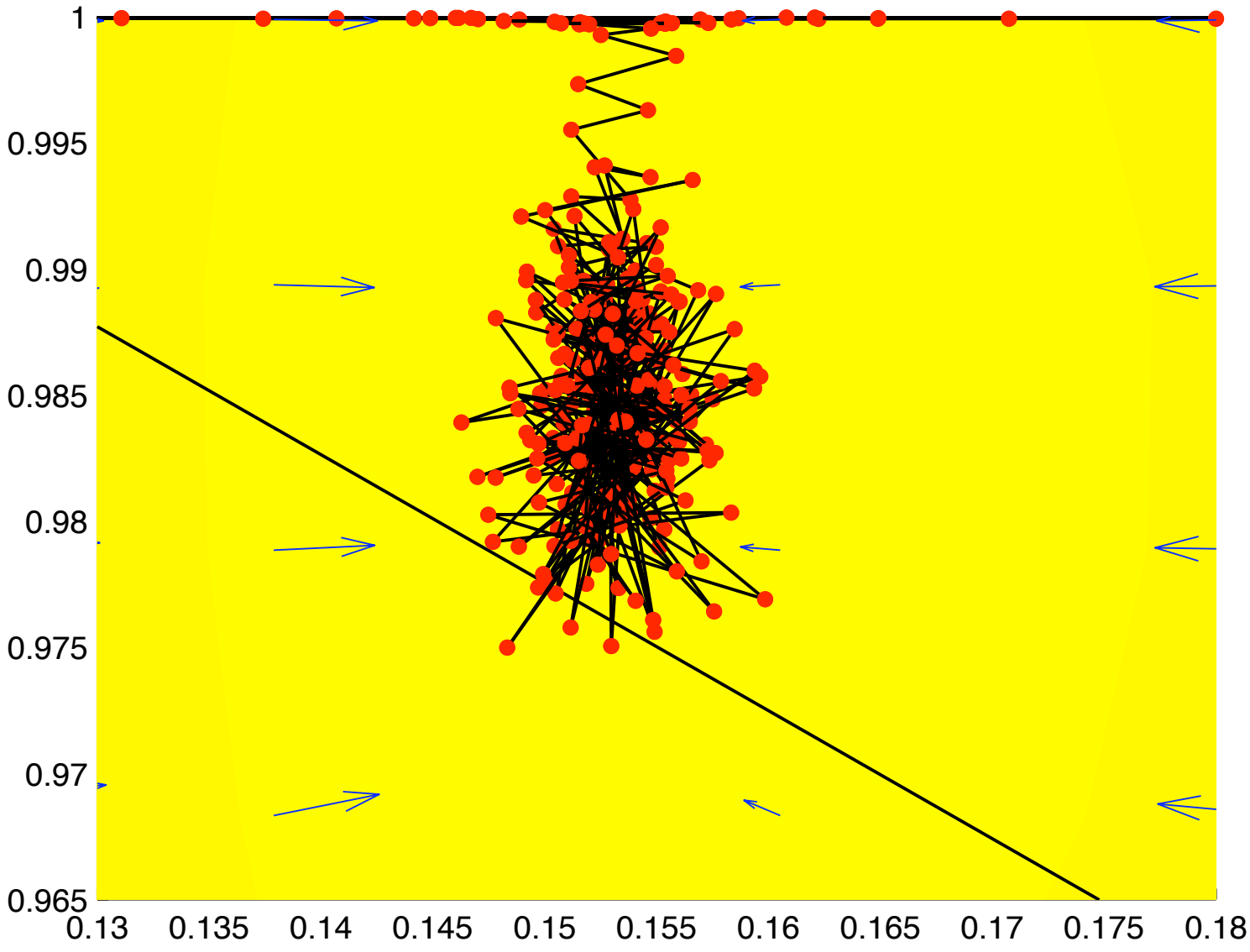




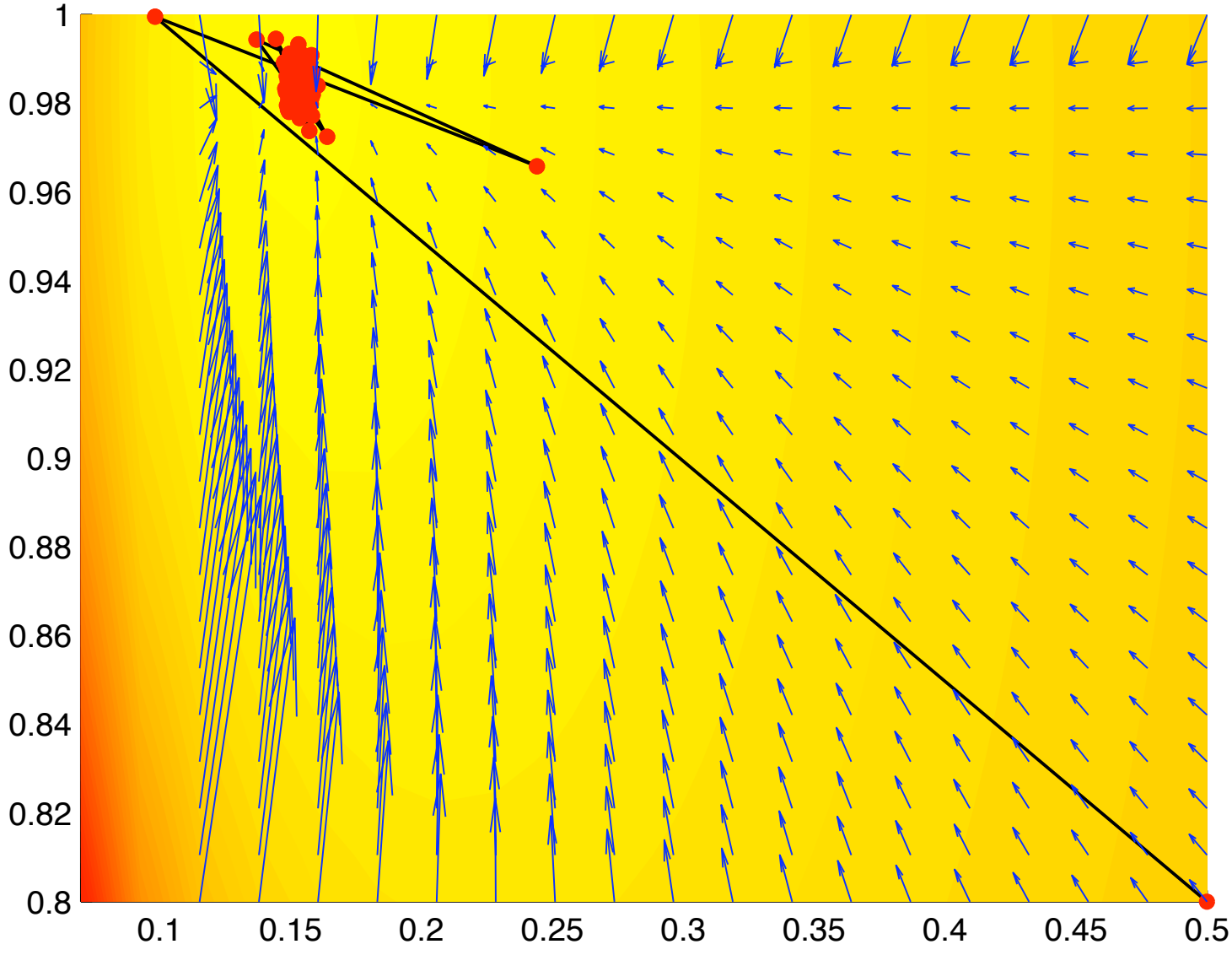
HMC Algorithm, Parameters of Stoch Vol Model, Acc Rate 95%



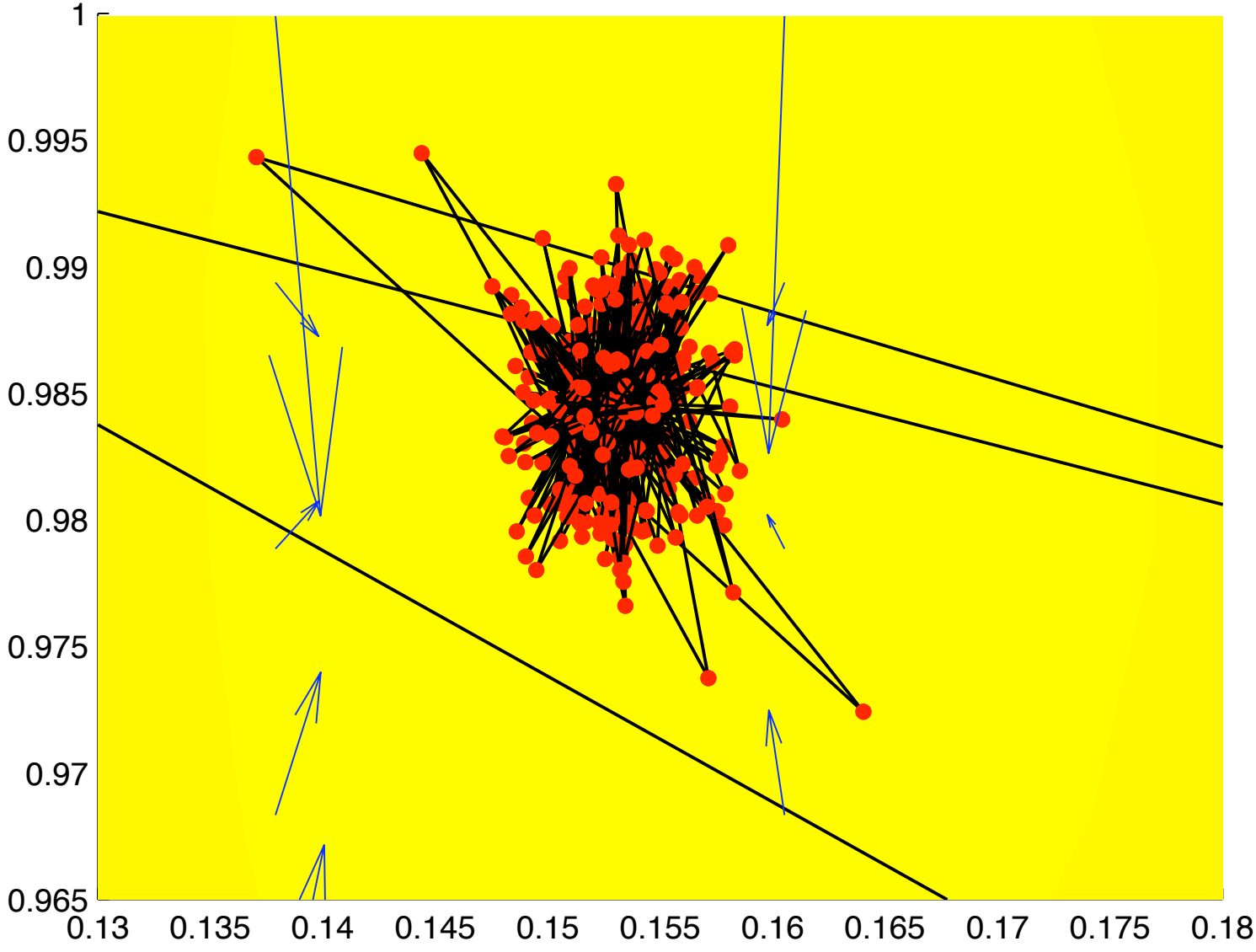
HMC Algorithm, Parameters of Stoch Vol Model, Acc Rate 95%



RMHMC Algorithm, Parameters of Stoch Vol Model, Acc Rate 95%



RMHMC Algorithm, Parameters of Stoch Vol Model, Acc Rate 95%



Log-Gaussian Cox Point Process with Latent Field

- Number of points in cells on 64×64 grid denoted by $\mathbf{Y} = \{Y_{i,j}\}$ conditionally independent and Poisson distributed with means $m\Lambda(i,j) = m\exp(X_{i,j})$

Log-Gaussian Cox Point Process with Latent Field

- Number of points in cells on 64×64 grid denoted by $\mathbf{Y} = \{Y_{i,j}\}$ conditionally independent and Poisson distributed with means $m\Lambda(i,j) = m\exp(X_{i,j})$
- $\mathbf{X} = \{X_{i,j}\} \sim GP$, $E\{\mathbf{x}\} = \mu\mathbf{1}$, $\Sigma_{(i,j),(i',j')} = \sigma^2 \exp(-\delta(i,i',j,j')/64\beta)$, and $\delta(i,i',j,j') = \sqrt{(i-i')^2 + (j-j')^2}$.

Log-Gaussian Cox Point Process with Latent Field

- Number of points in cells on 64×64 grid denoted by $\mathbf{Y} = \{Y_{i,j}\}$ conditionally independent and Poisson distributed with means $m\Lambda(i,j) = m\exp(X_{i,j})$
- $\mathbf{X} = \{X_{i,j}\} \sim GP$, $E\{\mathbf{x}\} = \mu\mathbf{1}$, $\Sigma_{(i,j),(i',j')} = \sigma^2 \exp(-\delta(i,i',j,j')/64\beta)$, and $\delta(i,i',j,j') = \sqrt{(i-i')^2 + (j-j')^2}$.
- The joint density is

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

Log-Gaussian Cox Point Process with Latent Field

- Number of points in cells on 64×64 grid denoted by $\mathbf{Y} = \{Y_{i,j}\}$ conditionally independent and Poisson distributed with means $m\Lambda(i,j) = m\exp(X_{i,j})$
- $\mathbf{X} = \{X_{i,j}\} \sim GP$, $E\{\mathbf{x}\} = \mu\mathbf{1}$, $\Sigma_{(i,j),(i',j')} = \sigma^2 \exp(-\delta(i,i',j,j')/64\beta)$, and $\delta(i,i',j,j') = \sqrt{(i-i')^2 + (j-j')^2}$.
- The joint density is

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

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

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

Log-Gaussian Cox Point Process with Latent Field

- Number of points in cells on 64×64 grid denoted by $\mathbf{Y} = \{Y_{i,j}\}$ conditionally independent and Poisson distributed with means $m\Lambda(i,j) = m\exp(X_{i,j})$
- $\mathbf{X} = \{X_{i,j}\} \sim GP$, $E\{\mathbf{x}\} = \mu\mathbf{1}$, $\Sigma_{(i,j),(i',j')} = \sigma^2 \exp(-\delta(i,i',j,j')/64\beta)$, and $\delta(i,i',j,j') = \sqrt{(i-i')^2 + (j-j')^2}$.
- The joint density is

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

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

  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

- Number of points in cells on 64×64 grid denoted by $\mathbf{Y} = \{Y_{i,j}\}$ conditionally independent and Poisson distributed with means $m\Lambda(i,j) = m\exp(X_{i,j})$
- $\mathbf{X} = \{X_{i,j}\} \sim GP$, $E\{\mathbf{x}\} = \mu\mathbf{1}$, $\Sigma_{(i,j),(i',j')} = \sigma^2 \exp(-\delta(i,i',j,j')/64\beta)$, and $\delta(i,i',j,j') = \sqrt{(i-i')^2 + (j-j')^2}$.
- The joint density is

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

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

  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
- Metropolis and HMC fails.... completely. MALA requires transformation of latent field to sample - with separate tuning in transient and stationary phases - RMHMC requires no pilot tuning or additional transformations

RMHMC for Log-Gaussian Cox Point Processes

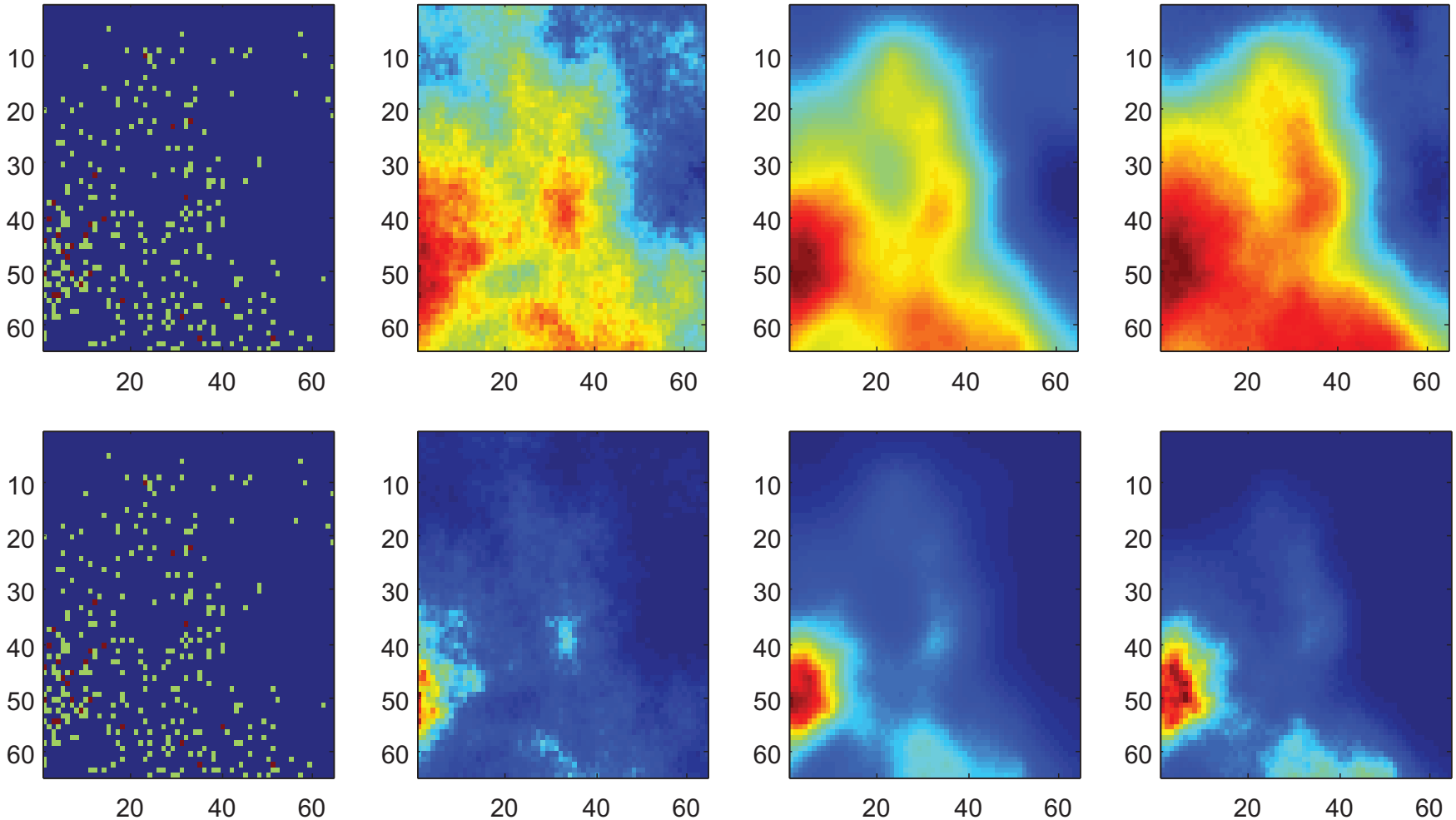
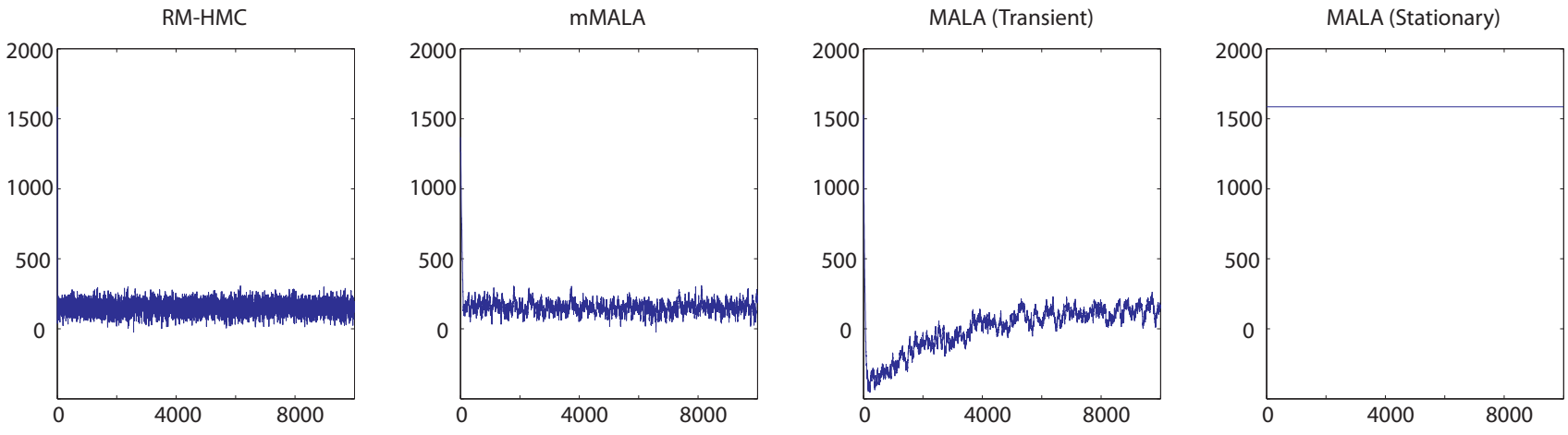


Figure: Data, Latent Field, Poisson Mean Field

RMHMC for Log-Gaussian Cox Point Processes



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	$\times 7070$

RMHMC for Log-Gaussian Cox Point Processes

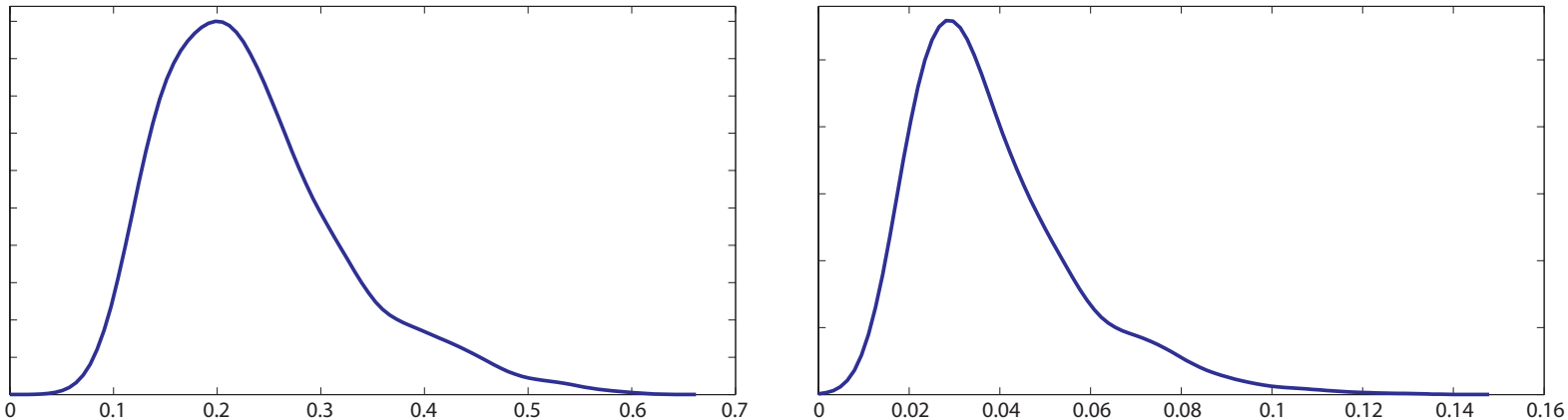


Figure: Kernel density estimates of the hyperparameter samples obtained from Gibbs style sampling from the Log-Gaussian Cox model. The true values are $\sigma = 0.19$ (left hand plot) and $\beta = 0.03$ (right hand plot).

Conclusion and Discussion

- HMC implicitly defines a flat manifold upon which the statistical model resides - manual tuning challenging in complex models

Conclusion and Discussion

- HMC implicitly defines a flat manifold upon which the statistical model resides - manual tuning challenging in complex models
- Exploiting Riemannian structure of statistical models to devise Hamiltonian on manifold

Conclusion and Discussion

- HMC implicitly defines a flat manifold upon which the statistical model resides - manual tuning challenging in complex models
- Exploiting Riemannian structure of statistical models to devise Hamiltonian on manifold
- No requirement for costly pilot runs and estimates of posterior covariance

Conclusion and Discussion

- HMC implicitly defines a flat manifold upon which the statistical model resides - manual tuning challenging in complex models
- Exploiting Riemannian structure of statistical models to devise Hamiltonian on manifold
- No requirement for costly pilot runs and estimates of posterior covariance
- No requirement for costly tuning in transient and stationary phases of Markov chain

Conclusion and Discussion

- HMC implicitly defines a flat manifold upon which the statistical model resides - manual tuning challenging in complex models
- Exploiting Riemannian structure of statistical models to devise Hamiltonian on manifold
- No requirement for costly pilot runs and estimates of posterior covariance
- No requirement for costly tuning in transient and stationary phases of Markov chain
- Assessed on complex high-dimensional latent variable models

Conclusion and Discussion

- HMC implicitly defines a flat manifold upon which the statistical model resides - manual tuning challenging in complex models
- Exploiting Riemannian structure of statistical models to devise Hamiltonian on manifold
- No requirement for costly pilot runs and estimates of posterior covariance
- No requirement for costly tuning in transient and stationary phases of Markov chain
- Assessed on complex high-dimensional latent variable models
- Theoretical analysis of effect of integration error, design of metric, theoretical convergence rate of sampler, Optimality of Hamiltonian flows as local geodesics,

Conclusion and Discussion

- HMC implicitly defines a flat manifold upon which the statistical model resides - manual tuning challenging in complex models
- Exploiting Riemannian structure of statistical models to devise Hamiltonian on manifold
- No requirement for costly pilot runs and estimates of posterior covariance
- No requirement for costly tuning in transient and stationary phases of Markov chain
- Assessed on complex high-dimensional latent variable models
- Theoretical analysis of effect of integration error, design of metric, theoretical convergence rate of sampler, Optimality of Hamiltonian flows as local geodesics,
- Transition operators that DO NOT rely on implicit integrator desirable

Conclusion and Discussion

- HMC implicitly defines a flat manifold upon which the statistical model resides - manual tuning challenging in complex models
- Exploiting Riemannian structure of statistical models to devise Hamiltonian on manifold
- No requirement for costly pilot runs and estimates of posterior covariance
- No requirement for costly tuning in transient and stationary phases of Markov chain
- Assessed on complex high-dimensional latent variable models
- Theoretical analysis of effect of integration error, design of metric, theoretical convergence rate of sampler, Optimality of Hamiltonian flows as local geodesics,
- Transition operators that DO NOT rely on implicit integrator desirable

Codes to Replicate Results in Paper

- `http://www.dcs.gla.ac.uk/inference/rmhmc`

Codes to Replicate Results in Paper

- `http://www.dcs.gla.ac.uk/inference/rmhmc`
- Work funded by EPSRC Advanced Research Fellowship (Girolami), BBSRC project grant and Microsoft PhD Scholarship (Calderhead)

Funding and Research Group







http://www.dcs.gla.ac.uk/inference