Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods

Mark Girolami and Ben Calderhead

Paper reviewed by Hui Li

July 8, 2013

- Standard Metropolis-Hastings Sampling (M-H)
- Metropolis Adjusted Langevin Algorithm (MALA)
- Riemann Manifold Metropolis Adjusted Langevin Algorithm (RMMALA)
- Simplified Manifold Metropolis Adjusted Langevin Algorithm (Simplified mMALA)
- Riemann Manifold Hamiltonian Monte Carlo Method (RMHMC)

Standard Metropolis-Hastings Sampling

- Target distribution $p(\boldsymbol{\theta})$
- Proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$
- Draw a new sample sample $\boldsymbol{\theta}^*$ from proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$
- Accept the new sample with probability $min\{1, \frac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\}$
- Success of MCMC relies upon appropriate proposal design q

- Most-used choice of $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ includes Normal distribution $\mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \sigma^2\mathrm{I})$, a D-dimensional norm distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\sigma^2\mathrm{I}$
- Random walk
- Need a long time to travel the whole parameter space
- Low acceptance rate
- Poor mixing of the chain and highly-correlated samples
- No information of target distribution needed

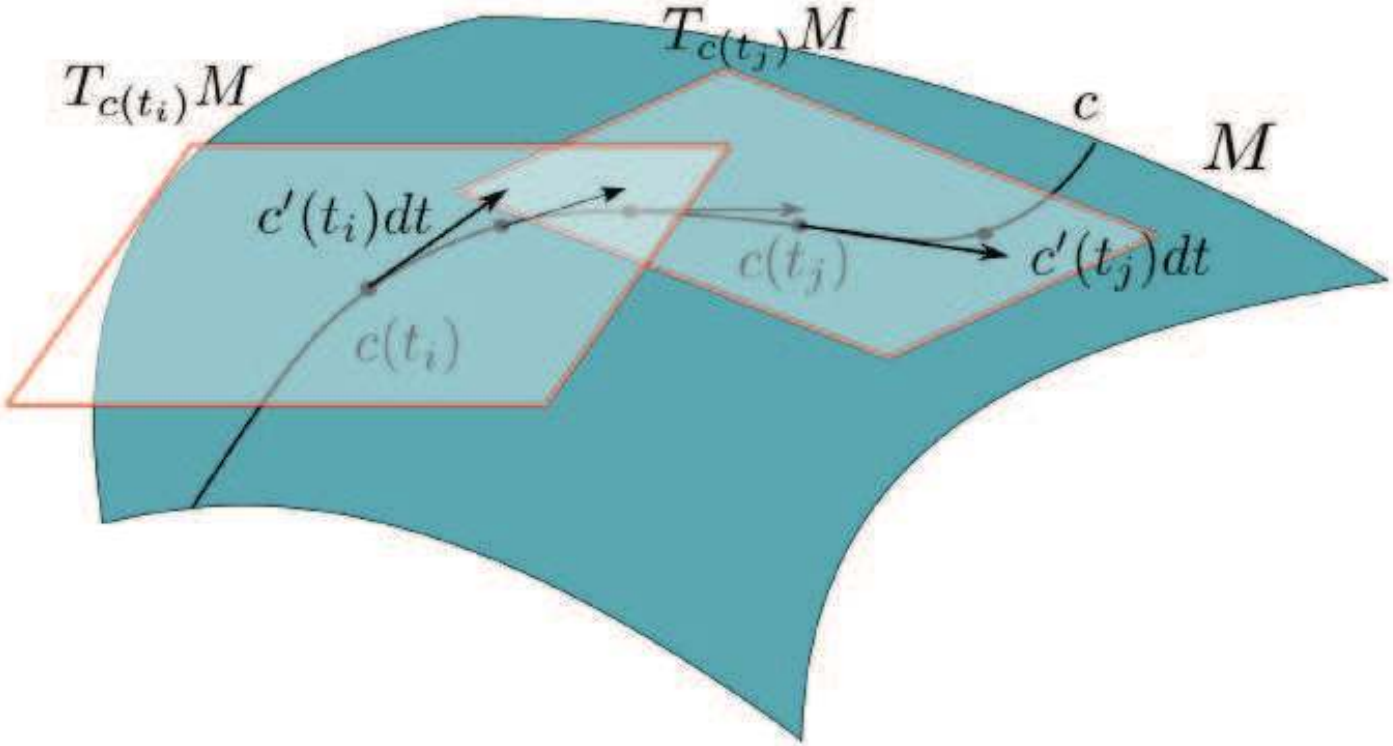
Metropolis Adjusted Langevin Algorithm (MALA)

- Stochastic differential equation (SDE) in Langevin diffusion: $d\boldsymbol{\theta} = \nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\}dt + d\mathbf{b}(t)$
- If $\mathcal{L}\{\boldsymbol{\theta}(t)$ is defined as log density of $p(\boldsymbol{\theta})$, then the solution of SDE is the stationary distribution $p(\boldsymbol{\theta})$
- First order Euler discretization of SDE: $\boldsymbol{\theta}^* = \boldsymbol{\theta}^n + \varepsilon^2\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}^n\}/2 + \varepsilon\mathbf{z^n}$
- Proposal distribution can be defined as $\mathcal{N}(\boldsymbol{\theta}^*|\mu(\boldsymbol{\theta}^n, \varepsilon), \mathrm{I})$ with $\mu(\boldsymbol{\theta}^n, \varepsilon) = \boldsymbol{\theta}^n + \varepsilon^2\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}^n\}/2$

Properties of MALA

- Instead of random walking, MALA tries to sample a new value along the direction which maximizes the log density function
- Isotropic diffusion which forces the choice of step size to accommodate variate with smallest variance
- This can be circumvented by employing a pre-conditioning matrix $\mathbf{M}$, $\mathcal{N}(\boldsymbol{\theta}^*|\mu(\boldsymbol{\theta}^n, \varepsilon, M), \mathbf{M}^{1/2})$ with $\mu(\boldsymbol{\theta}^n, \varepsilon, M) = \boldsymbol{\theta}^n + \varepsilon^2 \mathbf{M} \nabla_{\boldsymbol{\theta}} \mathcal{L}\{\boldsymbol{\theta}^n\}/2$
- Problem: How to select the pre-conditioning matrix $\mathbf{M}$

A manifold:



- A two-dimensional surface embedded in 3D ambient space;
- Euclidean geometry;
- Non-Euclidean geometry;

How to measure the distance of two points

- Euclidean geometry measures a distance between two points $\boldsymbol{\theta} = c(t_i)$ and $\boldsymbol{\theta} + \delta\boldsymbol{\theta} = c(t_j)$ as follows

$$D_E(c(t_i), c(t_j)) = \sqrt{\delta\boldsymbol{\theta}^T \delta\boldsymbol{\theta}} \tag{1}$$

- Non-Euclidean geometry measures a distance between two points by considering the manifold to which the two points belong

Riemann manifold

- Tangent space $T_{\boldsymbol{\theta}}M$ is a linear approximation of the manifold and is spanned by the differential operator $[\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_n}]$
- Riemann manifold is a differential manifold in which the tangent space $T_{\theta}M$ at each point has an inner product defined via a metric tensor $G_{\boldsymbol{\theta}}$.
- The metric tensor is a function which takes two vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ as input and outputs a real-valued scalar $G_{\boldsymbol{\theta}} : T_{\boldsymbol{\theta}}M \times T_{\boldsymbol{\theta}}M \mapsto \mathcal{R}$
- The distance of two points $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \delta\boldsymbol{\theta}$ in Rieman manifold is defined as $D_{G_{\boldsymbol{\theta}}}(\boldsymbol{\theta}, \boldsymbol{\theta} + \delta\boldsymbol{\theta}) = \sqrt{\delta\boldsymbol{\theta}^T G_{\boldsymbol{\theta}} \delta\boldsymbol{\theta}}$

Properties of Riemannian metric tensor

- Symmetric: $G_{\boldsymbol{\theta}}(t_1, t_2) = G_{\boldsymbol{\theta}}(t_2, t_1)$;
- Bilinear: $G_{\boldsymbol{\theta}}(t_1 + t_2, t_3) = G_{\boldsymbol{\theta}}(t_1, t_3) + G_{\boldsymbol{\theta}}(t2, t_3)$;
- Positive definite: $G_{\boldsymbol{\theta}}(t_1, t_1) > 0$

Geometric concept in MCMC

- Denote the expected Fisher information as $G_{\boldsymbol{\theta}} = cov(\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}))$
- The first order approximation of KL distance between $p(y;\boldsymbol{\theta})$ and $p(y;\boldsymbol{\theta}+\delta\boldsymbol{\theta})$

$$\begin{aligned} KL(\boldsymbol{\theta}, \boldsymbol{\theta}+\delta\boldsymbol{\theta}) &= \int p(y;\boldsymbol{\theta}+\delta\boldsymbol{\theta}) log \frac{p(y;\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{p(y;\boldsymbol{\theta})} dy \\ &\approx \delta\boldsymbol{\theta}^T G_\theta \delta\boldsymbol{\theta} \end{aligned} \tag{2}$$

- The expected Fisher information is a Reimannian metric tensor.

Riemannian manifold MALA (RMMALA)

- The SDE defining the Langevin diffusion in Riemann manifold $d\boldsymbol{\theta} = \tilde{\nabla}_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\}dt + d\tilde{\mathbf{b}}(t)$
- The natural gradient $\tilde{\nabla}_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\} = \mathbf{G}^{-1}(\boldsymbol{\theta}(t))\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\}$
- The Brownian motion in Riemann manifold $d\tilde{\mathbf{b}}_i(t) = |\mathbf{G}^{-1/2}(\boldsymbol{\theta}(t))| + \sum_{j=1}^{D} \frac{\partial}{\partial \boldsymbol{\theta}_j}[\mathbf{G}^{-1}(\boldsymbol{\theta}(t))_{ij}\mathbf{G}^{-1/2}(\boldsymbol{\theta}(t))]dt + [\sqrt{\mathbf{G}^{-1}(\boldsymbol{\theta}(t))}d\mathbf{b}(t)]_i$
- proposal distribution $\mathcal{N}(\boldsymbol{\theta}^*|\mu(\boldsymbol{\theta}^n, \varepsilon, \mathbf{G}), \sqrt{\mathbf{G}^{-1}}\mathrm{I})$ with $\mu(\boldsymbol{\theta}^n, \varepsilon, \mathbf{G}) = \boldsymbol{\theta}^n + \mathbf{G}^{-1}(\boldsymbol{\theta}(t))\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\} + |\mathbf{G}^{-1/2}(\boldsymbol{\theta}(t))| + \sum_{j=1}^{D} \frac{\partial}{\partial \boldsymbol{\theta}_j}[\mathbf{G}^{-1}(\boldsymbol{\theta}(t))_{ij}\mathbf{G}^{-1/2}(\boldsymbol{\theta}(t))]dt$

Illustrative example

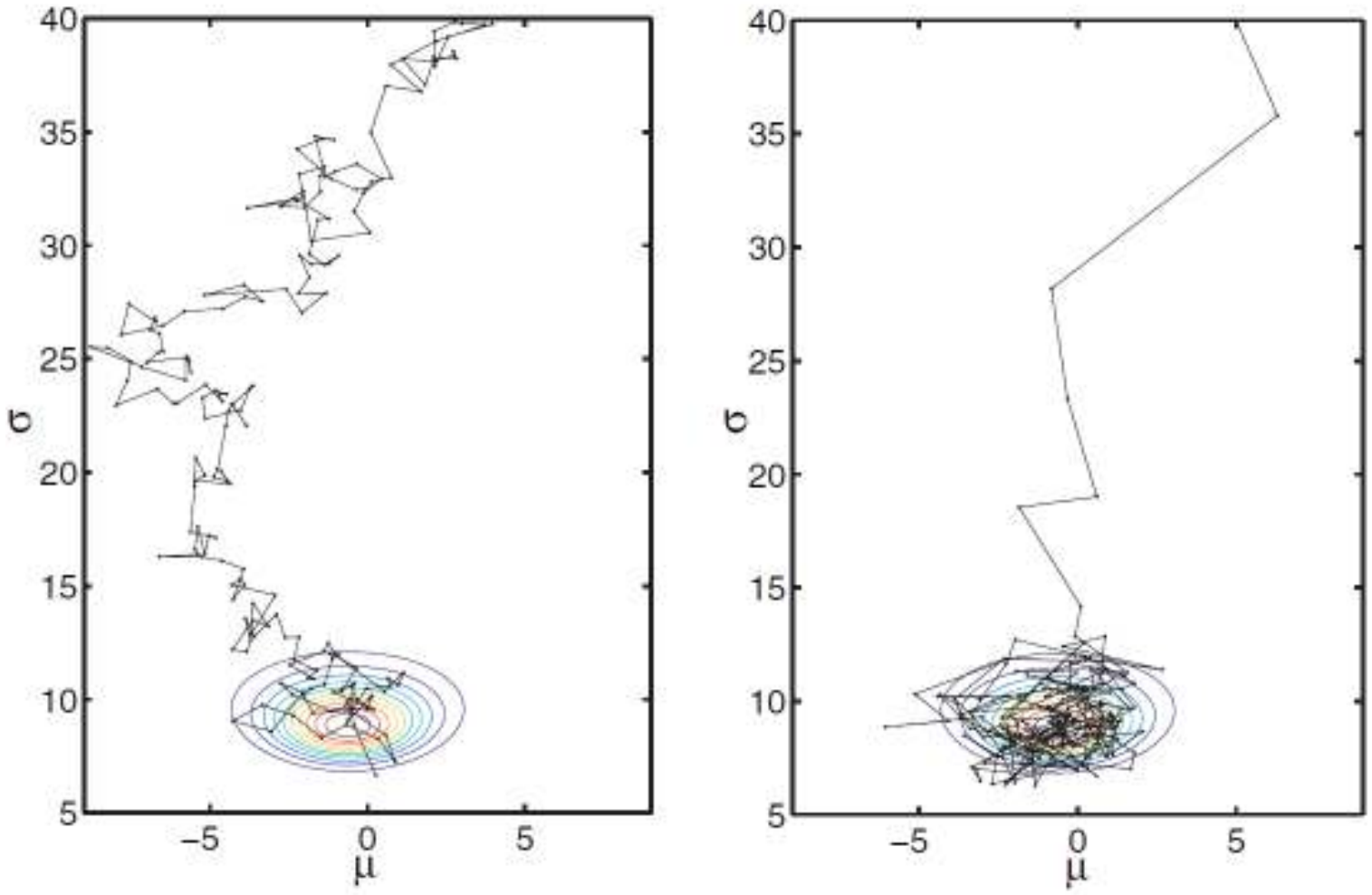
- Consider normal density $p(x|\mu,\sigma) = N_x(\mu,\sigma)$
- Infer the distribution of parameters μ and σ with mMala
- Local inner product on tangent space defined by a metric tensor, i.e. $\delta\boldsymbol{\theta}^T G_{\boldsymbol{\theta}} \delta\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\mu,\sigma)^T$
- Metric $G_{\boldsymbol{\theta}}$ is the expected Fisher information matrix

$$G(\mu,\sigma) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{bmatrix} \tag{3}$$

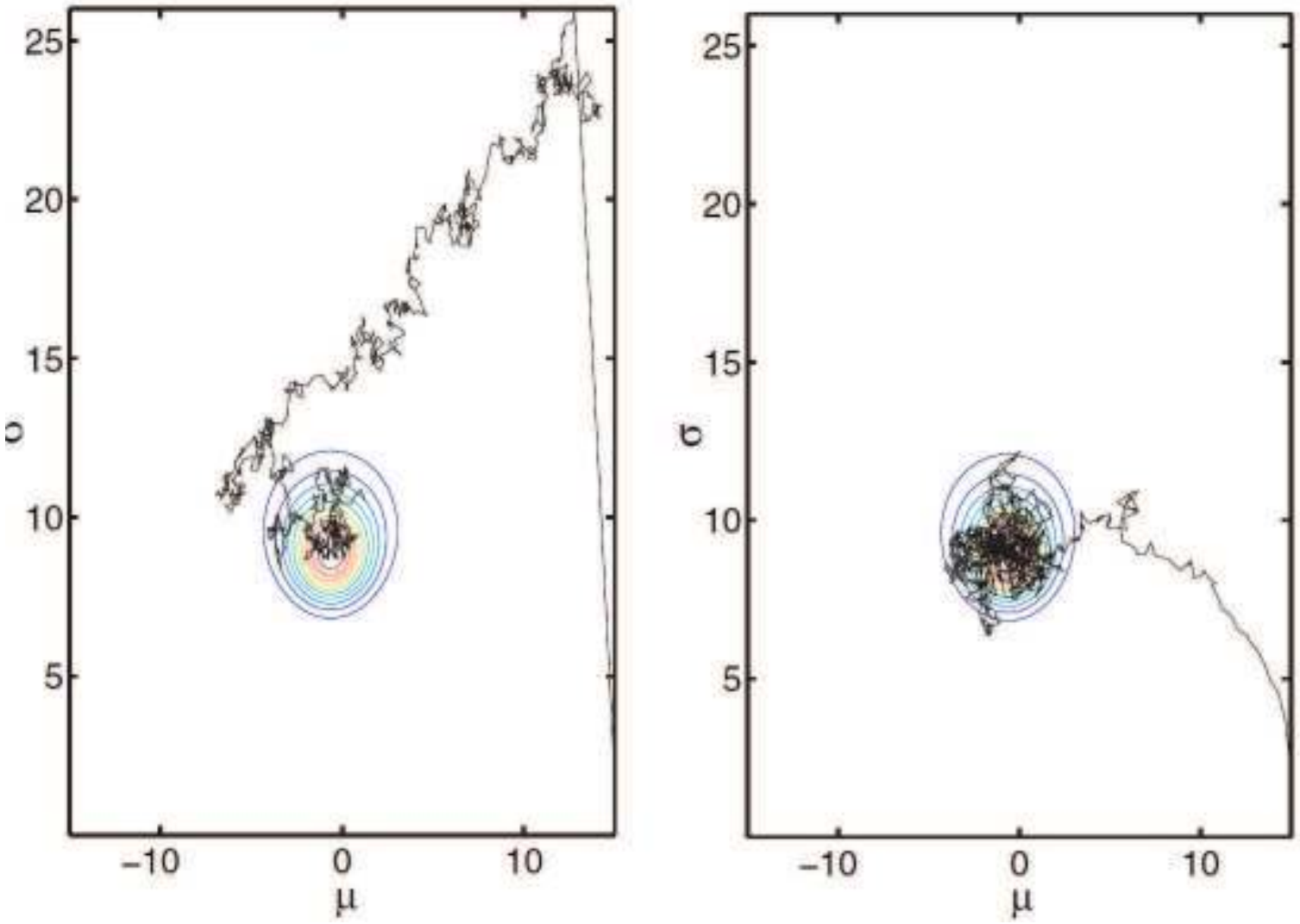
- Metric on tangent space

$$\delta\boldsymbol{\theta}^T G_{\boldsymbol{\theta}} \delta\boldsymbol{\theta} = \frac{\delta\mu^2 + 2\delta\sigma^2}{\sigma^2} \tag{4}$$

- A sample of size $N = 30$ was drawn from $N_x(\mu = 0, \sigma = 10)$
- Starting point is $\mu_0 = 0, \sigma_0 = 40$



- A sample of size $N = 30$ was drawn from $N_x(\mu = 0, \sigma = 10)$
- Starting point is $\mu_0 = 15, \sigma_0 = 2$



Conclusions of RMMALA

- Make moves in a Riemann metric rather than according to the standard Euclidean metric
- Utilize the information of the curvature of the manifold

Hamitanian Monte Carlo Method (HMC)

- A joint density $p(\boldsymbol{\theta}, \mathbf{p})$ is factorized as $p(\boldsymbol{\theta}, \mathbf{p}) = p(\boldsymbol{\theta})p(\mathbf{p})$
- $p(\boldsymbol{\theta})$ is the target density function
- $p(\mathbf{p}) = \mathcal{N}(0, \mathbf{M})$ is an independent auxiliary density function
- The Hamiltanian is defined as the negative of the log joint density $H(\boldsymbol{\theta}, \mathbf{p}) = -L(\boldsymbol{\theta}) + \frac{1}{2}\log\{(2\pi)^D|\mathbf{M}|\} + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}$
- The Hamiltanian equations:

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \\ \frac{d\mathbf{p}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \end{aligned} \tag{5}$$

Properties of HMC

- Preserve the total energy: $H(\boldsymbol{\theta}(t), \mathbf{p}(t)) = H(\boldsymbol{\theta}(0), \mathbf{p}(0))$, and hence $p(\boldsymbol{\theta}(t), \mathbf{p}(t)) = p(\boldsymbol{\theta}(0), \mathbf{p}(0))$
- Preserve the volume: $d\boldsymbol{\theta}(t)d\mathbf{p}(t) = d\boldsymbol{\theta}(0)d\mathbf{p}(0)$
- Time reversal

Euler first-order discretization and Leapfrog method

$$\begin{aligned} \mathbf{p}(t+\frac{\varepsilon}{2}) &= \mathbf{p}(t) + \varepsilon \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}(t))/2 \\ \boldsymbol{\theta}(t+\varepsilon) &= \boldsymbol{\theta}(t) + \varepsilon \mathbf{M}^{-1}\mathbf{p}(t+\frac{\varepsilon}{2}) \\ \mathbf{p}(t+\varepsilon) &= \mathbf{p}(t+\frac{\varepsilon}{2}) + \varepsilon \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}(t+\varepsilon))/2 \end{aligned} \tag{6}$$

- Probability of accepting a new sample $(\boldsymbol{\theta}^*, \mathbf{p}^*)$ is $min\{1, \exp(-H(\boldsymbol{\theta}^*, \mathbf{p}^*) + H(\boldsymbol{\theta}^n, \mathbf{p}^{n+1}))\}$

Riemann manifold HMC

- Considering the geometric information, a tensor metric $\mathbf{G}(\theta)$ defined at a point θ is used instead of $\mathbf{M}$
- Hamiltanian equations become

$$\begin{aligned}
\frac{d\theta}{dt} &= \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} = \mathbf{G}(\theta)^{-1}\mathbf{p} \\
\frac{d\mathbf{p}}{dt} &= -\frac{\partial H}{\partial \theta} = \nabla_{\theta} L(\theta) - \frac{1}{2} tr\{\mathbf{G}(\theta)^{-1}\frac{\partial \mathbf{G}(\theta)}{\partial \theta}\} \\
&+ \frac{1}{2}\mathbf{p}^T \mathbf{G}(\theta)^{-1}\frac{\partial \mathbf{G}(\theta)}{\partial \theta}\mathbf{G}(\theta)^{-1}\mathbf{p}
\end{aligned} \tag{7}$$

- Probability of accepting a new sample $(\theta^*, \mathbf{p}^*)$ is $min\{1, \exp(-H(\theta^*, \mathbf{p}^*) + H(\theta^n, \mathbf{p}^{n+1}))\}$

Example – Stochastic volatility model

Stochastic volatility model is defined with the latent volatilities taking the form of an AR(1) process such that

$$y_t = \epsilon_t \beta \exp(x_t/2) \tag{8}$$

with

$$x_{t+1} = \phi x_t + \eta_{t+1} \tag{9}$$

where

$$\begin{aligned} \epsilon_t &\sim N(0,1) \\ \eta_t &\sim N(0,\sigma^2) \end{aligned} \tag{10}$$

and

$$x_1 \sim N(0, \sigma^2/(1-\phi)) \tag{11}$$

Stochastic volatility model

- Joint density

$$p(\mathbf{y}, \mathbf{x}, \beta, \phi, \sigma) = \quad p(x_1) \prod_{t=1}^{T} p(y_t | x_t, \beta) \prod_{t+2}^{T} p(x_t | x_t - 1, \phi, \sigma) \\ \pi(\beta)\pi(\sigma)\pi(\phi) \tag{12}$$

- Split up the sampling procedure in two steps, simulate from

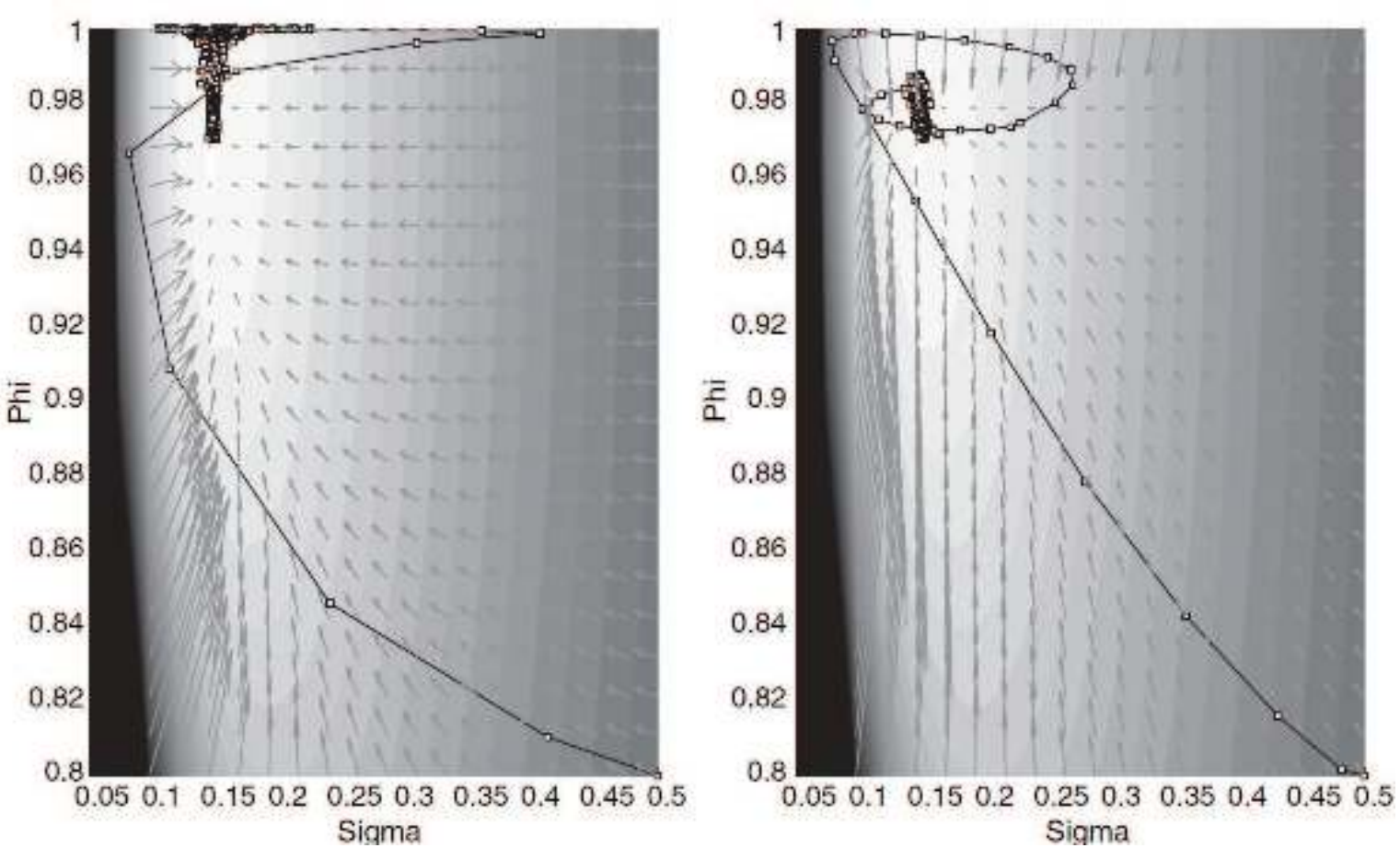
$$\begin{aligned} \beta, \phi, \sigma | \mathbf{y}, \mathbf{x} &\sim p(\beta, \phi, \sigma | \mathbf{y}, \mathbf{x}) \\ \mathbf{x} | \mathbf{y}, \beta, \phi, \sigma &\sim p(\mathbf{x} | \mathbf{y}, \beta, \phi, \sigma) \end{aligned} \tag{13}$$

- Metric tensor for parameters

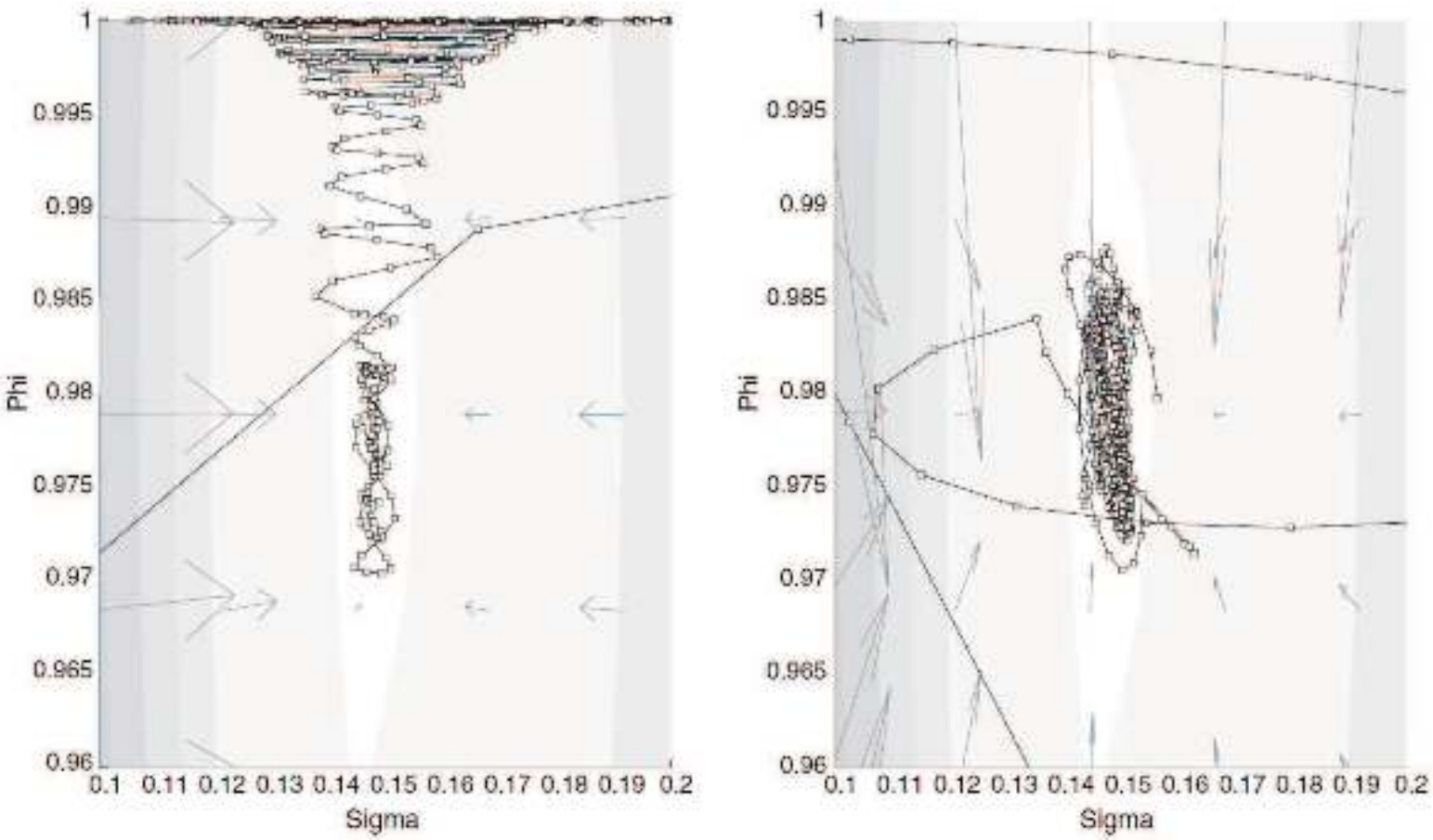
$$G(\beta, \phi, \sigma) = \begin{bmatrix} \frac{2T}{\beta^2} & 0 & 0 \\ 0 & 2T & 2\phi \\ 0 & 2\phi & 2\phi^2 + (T-1)(1-\phi^2) \end{bmatrix} \tag{14}$$

- Metric tensor for latent volatilites $G(\mathbf{x}) = \frac{\mathbf{I}}{2} + \mathbf{C}^{-1}$

- The value of β is set to the true value
- The log-joint-probability given different values of σ and ϕ is shown by the contour plot



Zoom up the plot



- RMHMC sampling effectively normalizes the gradient in each direction;
- HMC sampling with a unit mass matrix exhibits stronger gradients in horizontal direction than vertical direction and therefore takes much longer to converge to the target density.

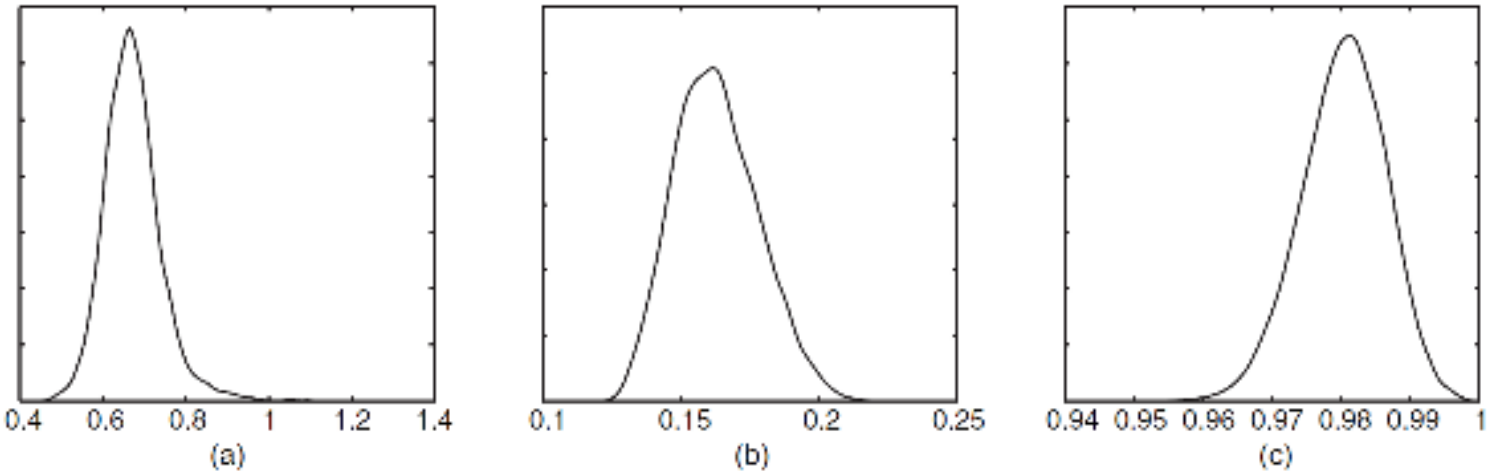
Stochastic Volatility Model - Performance

Method	*Mean time (s)*	*ESS* (β, σ, ϕ)	*Standard error* (β, σ, ϕ)	*s/minimum ESS*	*Relative speed*
MALA	44.0	(19.1, 11.3, 30.1)	(1.9,0.8,2.1)	3.89	36.7
HMC	424.8	(117, 81, 198)	(9.3, 3.1, 10.3)	5.24	27.3
MMALA	2455.9	(17.2, 17.4, 44.5)	(2.8, 2.4, 9.2)	142.8	1
RMHMC	329.4	(325, 139, 344)	(19.0, 7.3, 25.2)	2.37	60.3

Method	*Mean time (s)*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
MALA	44.0	(9.7, 16.7, 28.4)	4.53	7.5
HMC	424.8	(409, 624, 1239)	1.04	32.9
MMALA	2455.9	(71.8, 131.0, 329.8)	34.2	1
RMHMC	329.4	(977, 1689, 3376)	0.34	100.6

- 2000 simulated observations with $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$
- 20000 posterior samples averaged over 10 runs.

Stochastic Volatility Model - Performance



- Posterior marginal density for β, σ and ϕ