Hamiltonian Monte Carlo

Shiwei Lan

Department of Statistics,
University of California, Irvine, USA

March 12, 2014

Motivation

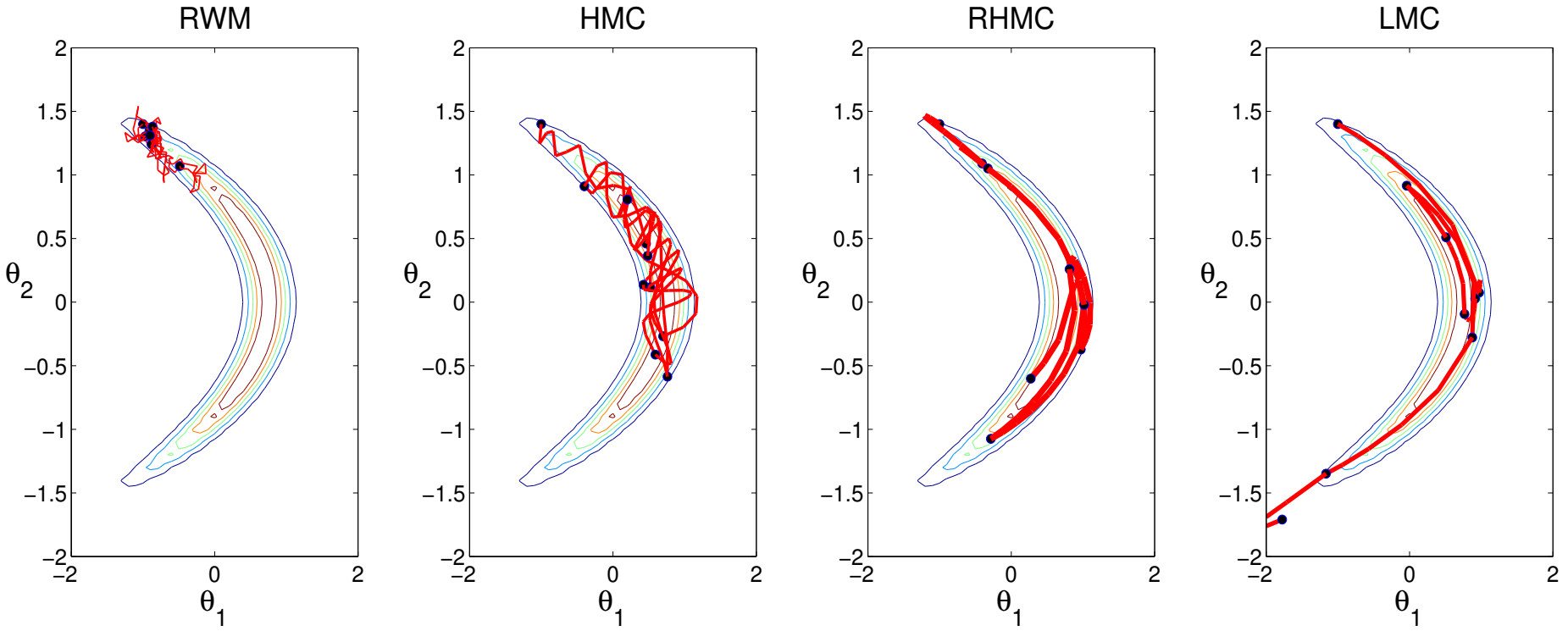


Figure 1: Comparing RWM, HMC, RHMC and LMC in sampling from a banana shaped distribution. Trajectory length is set to 1.5, the acceptance rates are 0.725, 0.9, 0.7, 0.8 respectively for the first 10 iterations. Blue dots are accepted proposals and red lines are sampling paths, with the thickness indicating the computational cost per iteration.

- In Bayesian statistics, given data and prior, we are interested in the posterior distribution of model parameters.

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{P(\mathcal{D})} P(\mathcal{D}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \propto P(\mathcal{D}|\boldsymbol{\theta}) P(\boldsymbol{\theta})$$

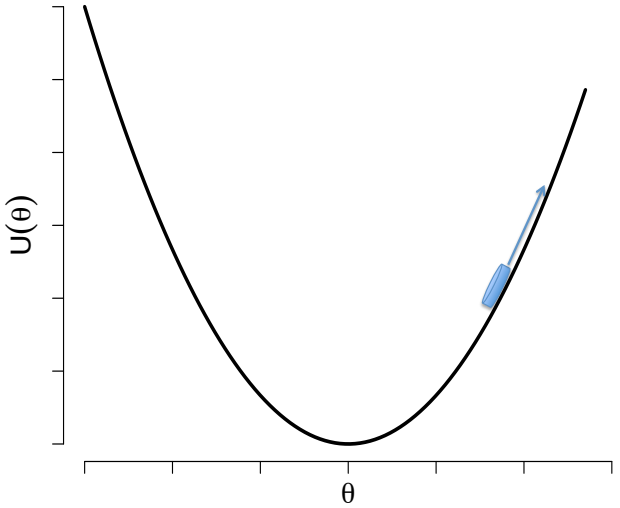
- The integral can be approximated by samples from Markov chain

$$P(y^*|\mathcal{D}) = \int P(y^*|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \approx \frac{1}{S} \sum_{s=1}^{S} P(y^*|\boldsymbol{\theta}^{(s)}),\ \boldsymbol{\theta}^{(s)} \sim P(\boldsymbol{\theta}|\mathcal{D})$$

Theorem 1 (Ergodicity)

An irreducible, aperiodic Markov chain that has a stationary distribution $\pi(\cdot)$ uniquely converges to $\pi(\cdot)$.

Hamiltonian Dynamics



$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \frac{\partial H}{\partial \mathbf{p}} \\ \dot{\mathbf{p}} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} \end{aligned}$$

- Position $\boldsymbol{\theta} \in \mathbb{R}^D \Longleftarrow$ variables of interest
- Momentum $\mathbf{p} \in \mathbb{R}^D \Longleftarrow$ fictitious, usually $\sim \mathcal{N}(\mathbf{0}, \mathbf{M})$
- Potential energy $U(\boldsymbol{\theta}) \Longleftarrow$ minus log of target density $\pi(\cdot)$
- Kinetic energy $K(\mathbf{p}) \Longleftarrow$ minus log of momentum density
- Hamiltonian $H(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p}) \Longleftarrow$ constant.

Bayesian Posterior Sampling

- We are interested in Posterior sampling $f(\theta|D) \propto f(\theta)L(\theta|D)$. Set

$$U(\theta) = -\log f(\theta|D) = -[\log f(\theta) + \sum_{i=1}^{N} \log f(x_i|\theta)] \quad \boxed{+C}$$

- Sample $p \sim \mathcal{N}(0, M^1)$, then set

$$K(p) = -\log f(p) = \frac{1}{2} p^{\mathsf{T}} M^{-1} p \quad \boxed{+C}$$

- Thus Hamiltonian $H(\theta, p) = U(\theta) + K(p)$ and joint density of (θ, p) is

$$f(\theta, p) \propto \exp\{-H(\theta, p)\} = \exp\{-U(\theta)\} \exp\{-K(p)\}$$

[1] Often set $M = I_d$ for simplicity, but more informative M works better.

Hamilton Dynamics

- With Hamiltonian defined as above, we have Hamilton Dynamic (HD)

Hamilton Dynamic

$$\begin{aligned} \dot{\theta} &= \frac{\partial}{\partial p} H(\theta, p) &= M^{-1} p \\ \dot{p} &= -\frac{\partial}{\partial \theta} H(\theta, p) &= -\frac{\partial U(\theta)}{\partial \theta} \end{aligned}$$

- If HD is analytically solvable, e.g. $U(\theta) \propto \theta^{\mathsf{T}} C \theta$, we directly get next sample θ^* after evolving for some time T.
- Important properties:
  - **Reversibility**: the target distribution remains invariant.
  - **Volume preservation**: the Jacobian determinant of HD mapping is 1.
  - **Conservation of Hamiltonian**: the acceptance probability is one.
  - See Neal (2010) for detailed discussion.

Leapfrog

- Numerical integration is used when analytic solution is not available:

Leapfrog

$$
\begin{aligned}
p(t+\varepsilon/2) &= p(t) - (\varepsilon/2)\frac{\partial U}{\partial \theta}(\theta(t)) \\
\theta(t+\varepsilon) &= \theta(t) + \varepsilon\frac{\partial K}{\partial p}(p(t+\varepsilon/2)) \\
p(t+\varepsilon) &= p(t+\varepsilon/2) - (\varepsilon/2)\frac{\partial U}{\partial \theta}(\theta(t+\varepsilon))
\end{aligned}
$$

- Important properties:
  - **Stability**: numerically stable
  - **Reversibility and Volume preservation**: still hold.
  - **Conservation of Hamiltonian**: broken, but corrected with Acceptance/Rejection step as follows.

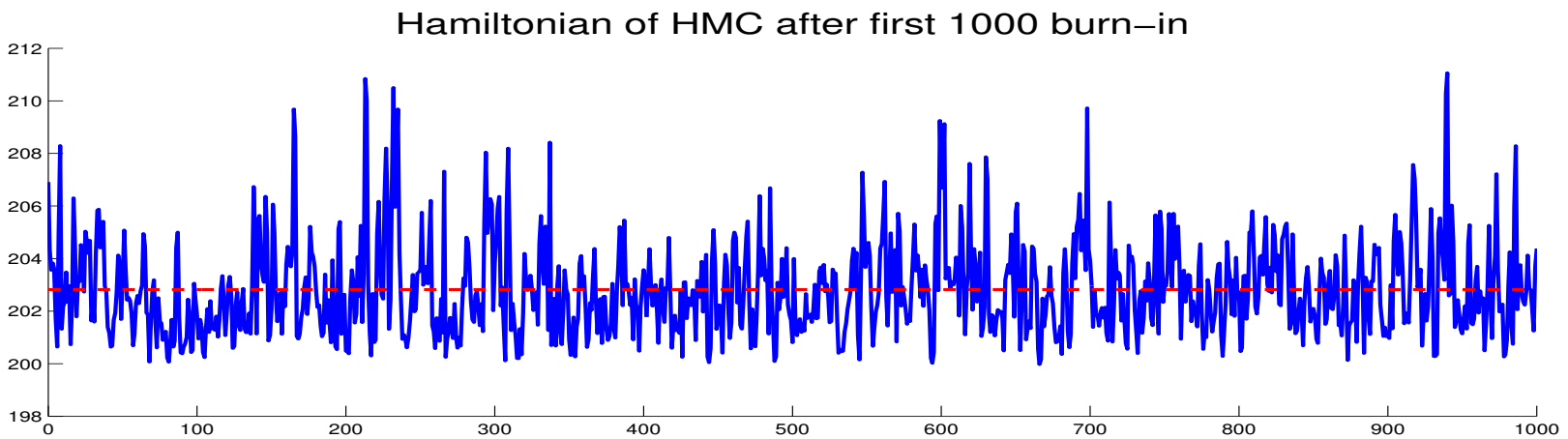
Acceptance/Rejection

Denote $z := (\theta, p)$. At the end of L-th step accept the proposal θ^* with

Metropolis Hastings Acceptance Probability

$$\alpha = \min\left\{1, \frac{f(z^*)g(z^* \to z)}{f(z)g(z \to z^*)}\right\} = \min\left\{1, \frac{\exp(-H(z^*))\delta(z^*, z)}{\exp(-H(z))\delta(z, z^*)}\right\}$$
$$= \min\{1, \exp(-H(z^*) + H(z))\}$$

Figure 2: Hamiltonian varies around its true value in a simulated LR problem



Algorithm 1 HMC algorithm (one iteration)

Initialize $\theta^{(1)} = \text{current } \theta$

Sample new momentum $p^{(1)} \sim \mathcal{N}(0, M)$

Calculate current $H(\theta^{(1)}, p^{(1)}) = U(\theta^{(1)}) + \frac{1}{2}(p^{(1)})^{\mathsf{T}} M^{-1} p^{(1)}$

for $\ell = 1$ to L (leapfrog steps) **do**

% Update momentum for half step

$p^{(\ell+\frac{1}{2})} = p^{(\ell)} - \varepsilon/2 \nabla_\theta U(\theta^{(\ell)})$

% Update position for full step

$\theta^{(\ell+1)} = \theta^{(\ell)} + \varepsilon M^{-1} p^{(\ell+\frac{1}{2})}$

% Update momentum for half step

$p^{(\ell+1)} = p^{(\ell+\frac{1}{2})} - \varepsilon/2 \nabla_\theta U(\theta^{(\ell+1)})$

end for

Calculate proposed $H(\theta^{(L+1)}, p^{(L+1)}) = U(\theta^{(L+1)}) + \frac{1}{2}(p^{(L+1)})^{\mathsf{T}} M^{-1} p^{(L+1)}$

logRatio $= -\text{ProposedH} + \text{CurrentH}$

Accept or reject according to the Metropolis ratio

Challenges of HMC

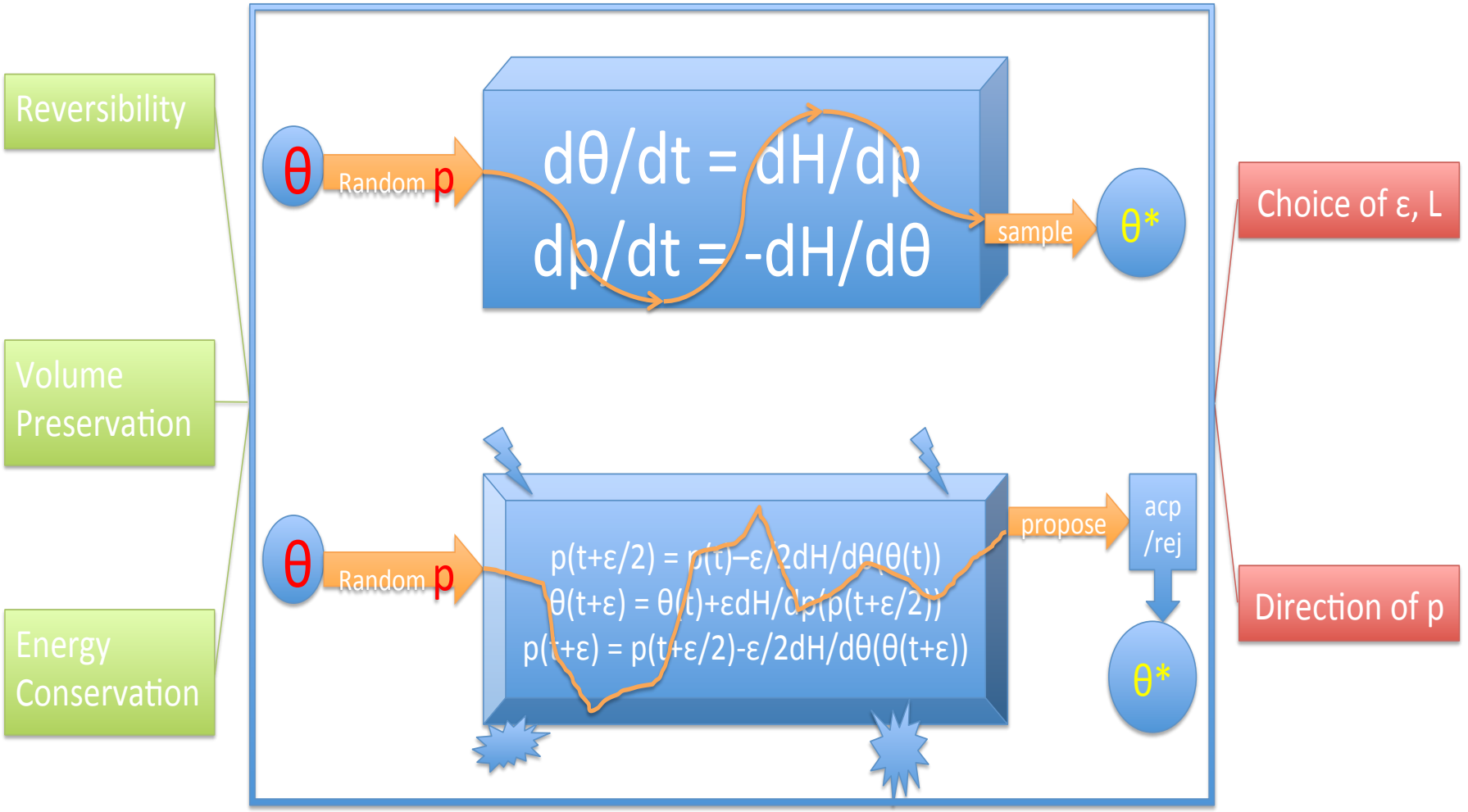
Note, the choice of step size ε is crucial:

- too large $\varepsilon \Longrightarrow$ large discretization errors $\Longrightarrow$ low acceptance rate;
- too small $\varepsilon \Longrightarrow$ waste computational time $\Longrightarrow$ slow exploration of posterior distribution.

Besides, direction of momentum p is also important:

- Want it always pointing at high density region, otherwise waste time in low density region or in a 'zig-zag' way.
- Aid from Geometry can suppress such 'zig-zag' behavior. Refer to Girolami and Calderhead (2011).

HMC Sampling



Simulated Logistic Regression

$$\begin{aligned} \Pr[y_i = 1|X, \theta] &= \frac{\exp(x_i\theta)}{1+\exp(x_i\theta)},\ i = 1, \cdots, N(=1000) \\ \theta &\sim \mathcal{N}(0, \sigma^2 I_3) \end{aligned}$$

where $\theta = (\theta_0, \theta_1, \theta_2)$. We simulate data x 1000 items from 2d Gaussian with covariance matrix $\begin{bmatrix} 1 & .5 \\ 5. & 1 \end{bmatrix}$, and then set $X = [1\ x]$. Then simulate $y \sim \text{Binom}(1, \text{expit}(X\theta'))$ with true $\theta' = [0\ 1\ 2]^\text{T}$.

Now we want to posterior sample θ. Compute

$$\begin{aligned} U(\theta) &= -y^\text{T} X\theta + 1_N^\text{T} \log(1+\exp(X\theta)) + \frac{1}{2}\theta^2/\sigma^2 \\ H(\theta, p) &= U(\theta) + \frac{1}{2}p^2 \\ \nabla U(\theta) &= -X^\text{T}[\theta - 1/(1+\exp(-X\theta))] - \theta/\sigma^2 \end{aligned}$$

Posterior Sampling

We sample 2000 iterations and burn in the first 100, and set trajectory length to be $T = \varepsilon L = 0.3$. For RWM, we set $\varepsilon = T/15$, record the result for every $L = 15$ jumps, and for HMC, we set $\varepsilon = T/5$ and $L = 5$. The following figure 4 shows the first 10 sampling iterations:

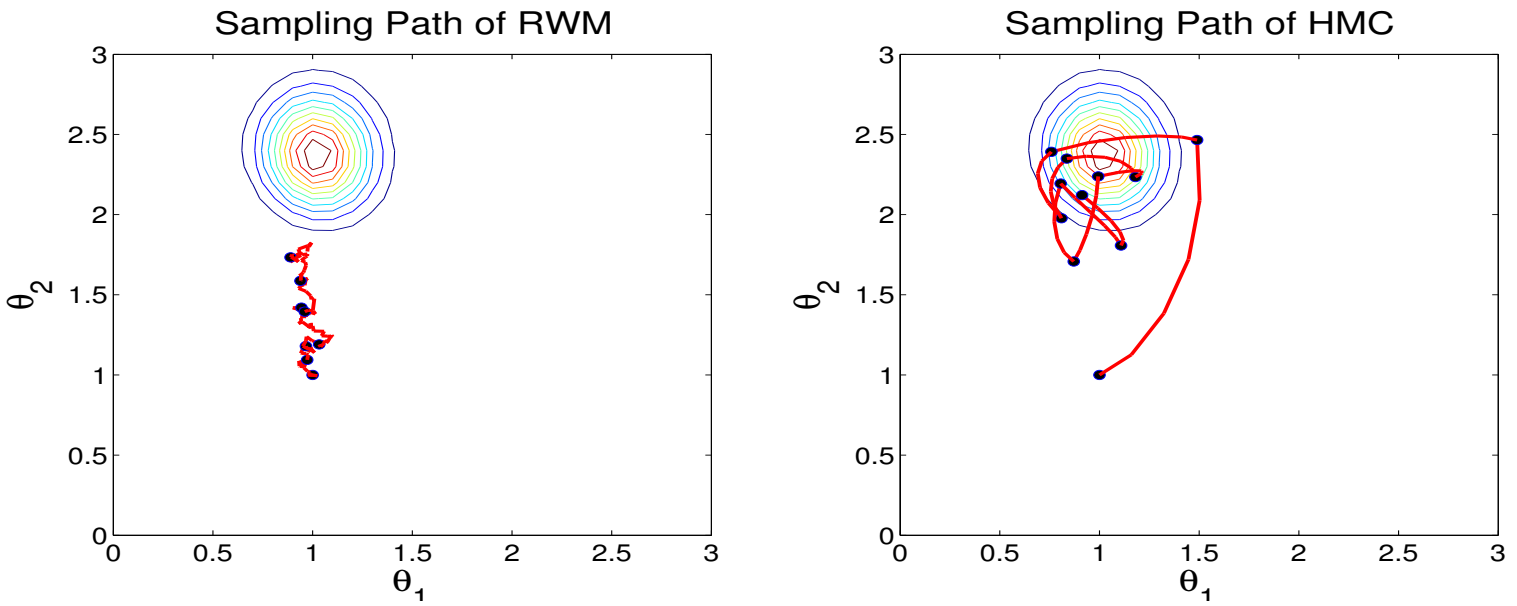


Figure 3: Sampling path for the first 10 iterations: RWM (left) vs HMC (right)

Posterior Sampling

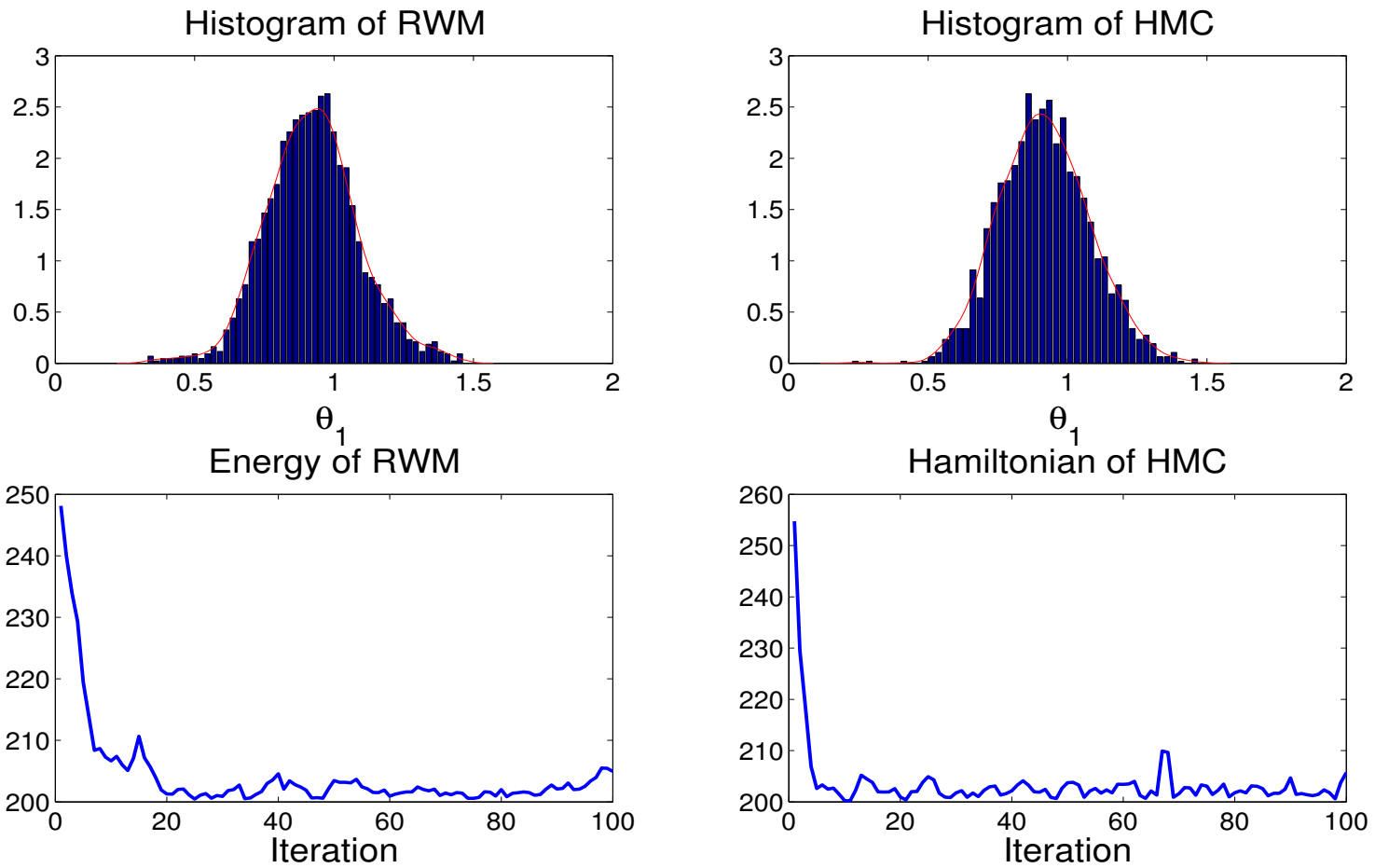
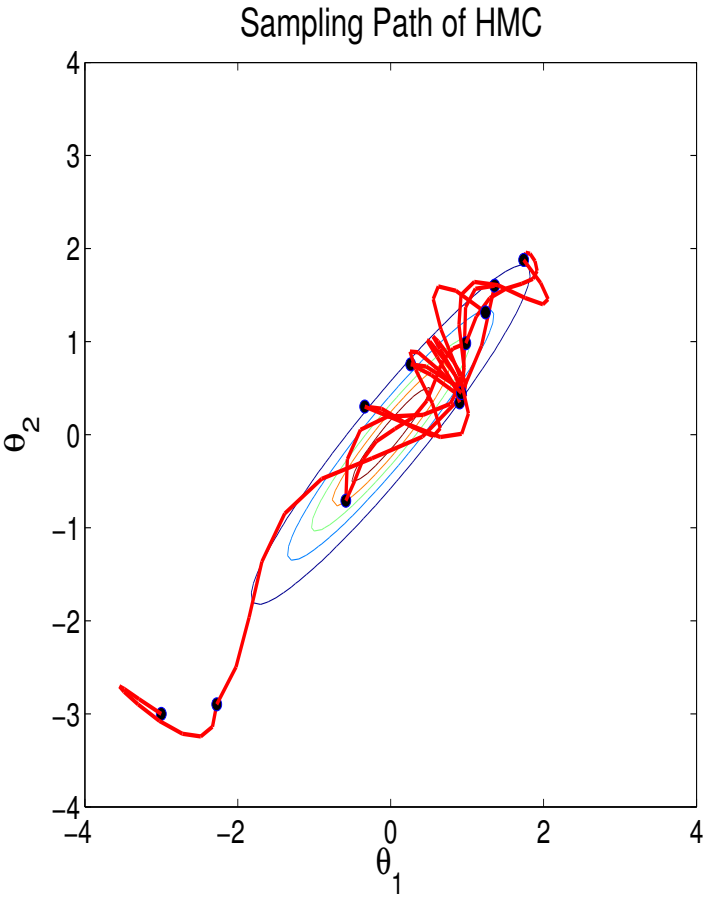
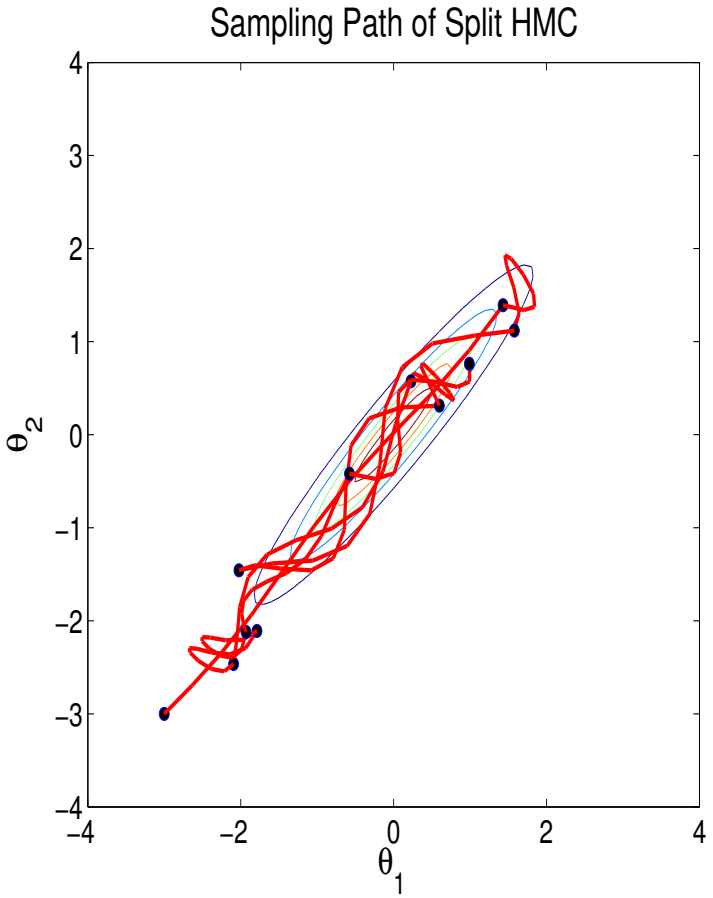


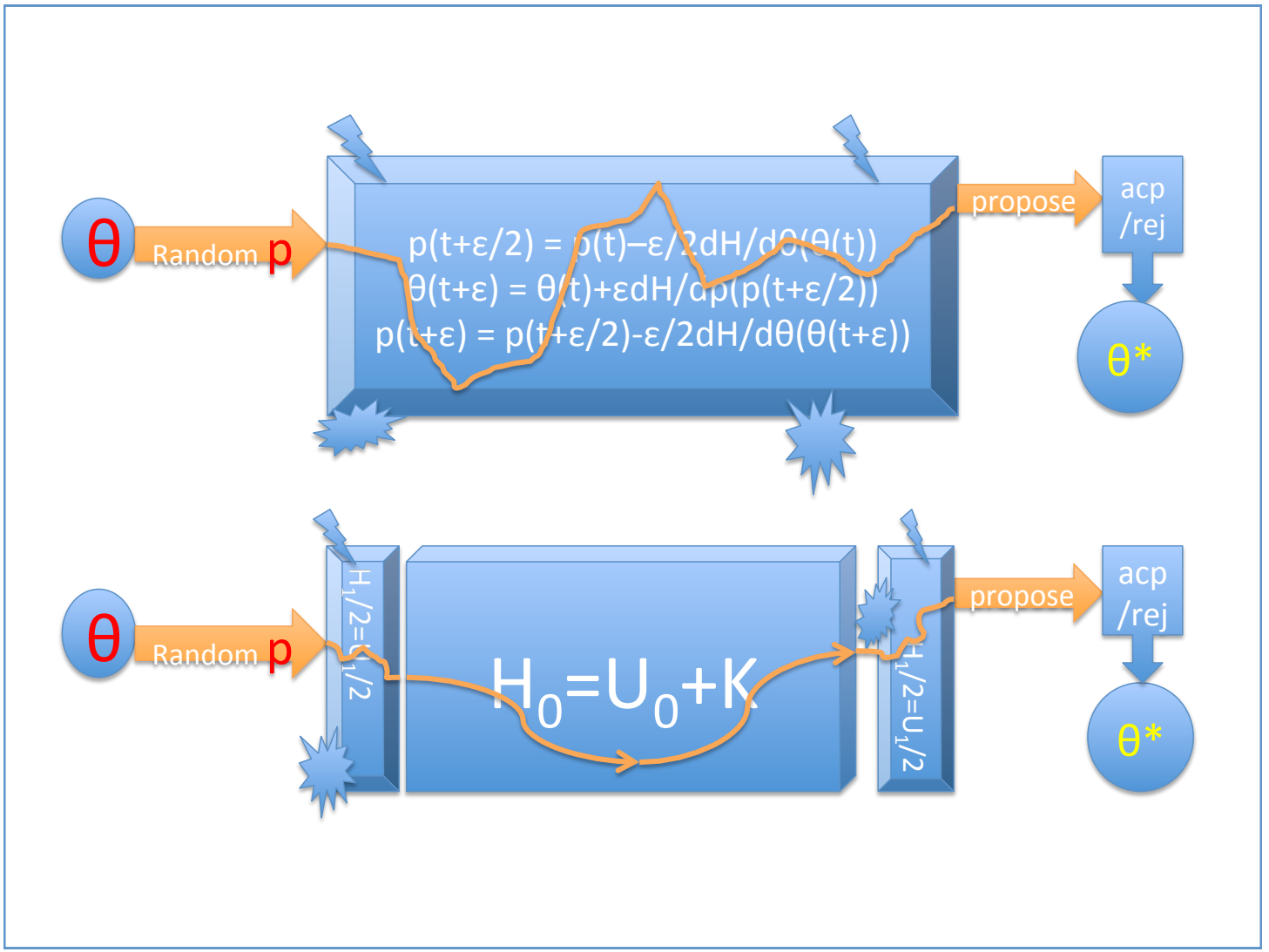
Figure 4: Histogram of θ_1 and energy plot for burn-in phase: RWM (left) vs HMC (right)

Split HMC

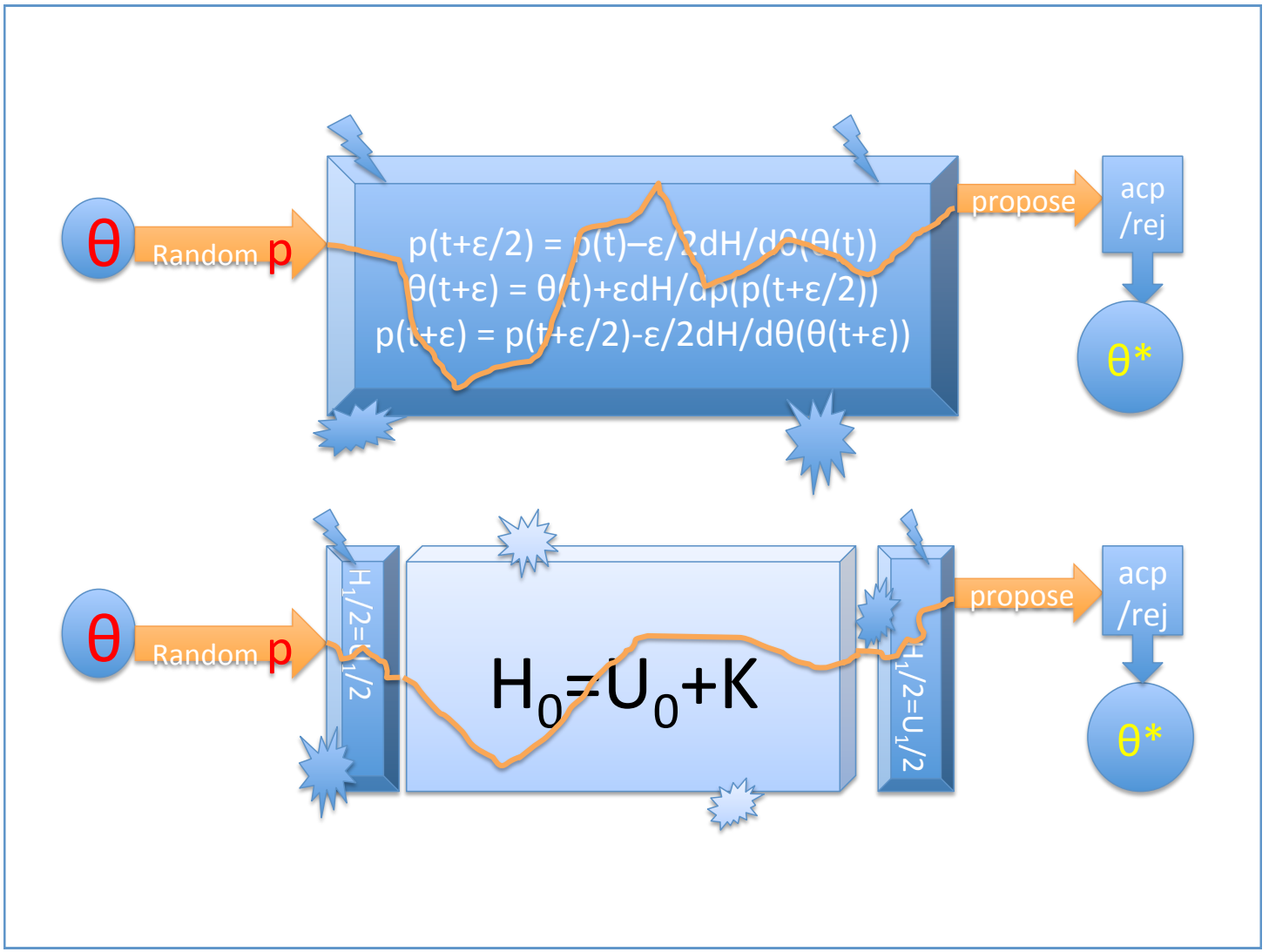

Sampling Path of HMC
θ2
θ1
4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4


Sampling Path of Split HMC
θ2
θ1
4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4

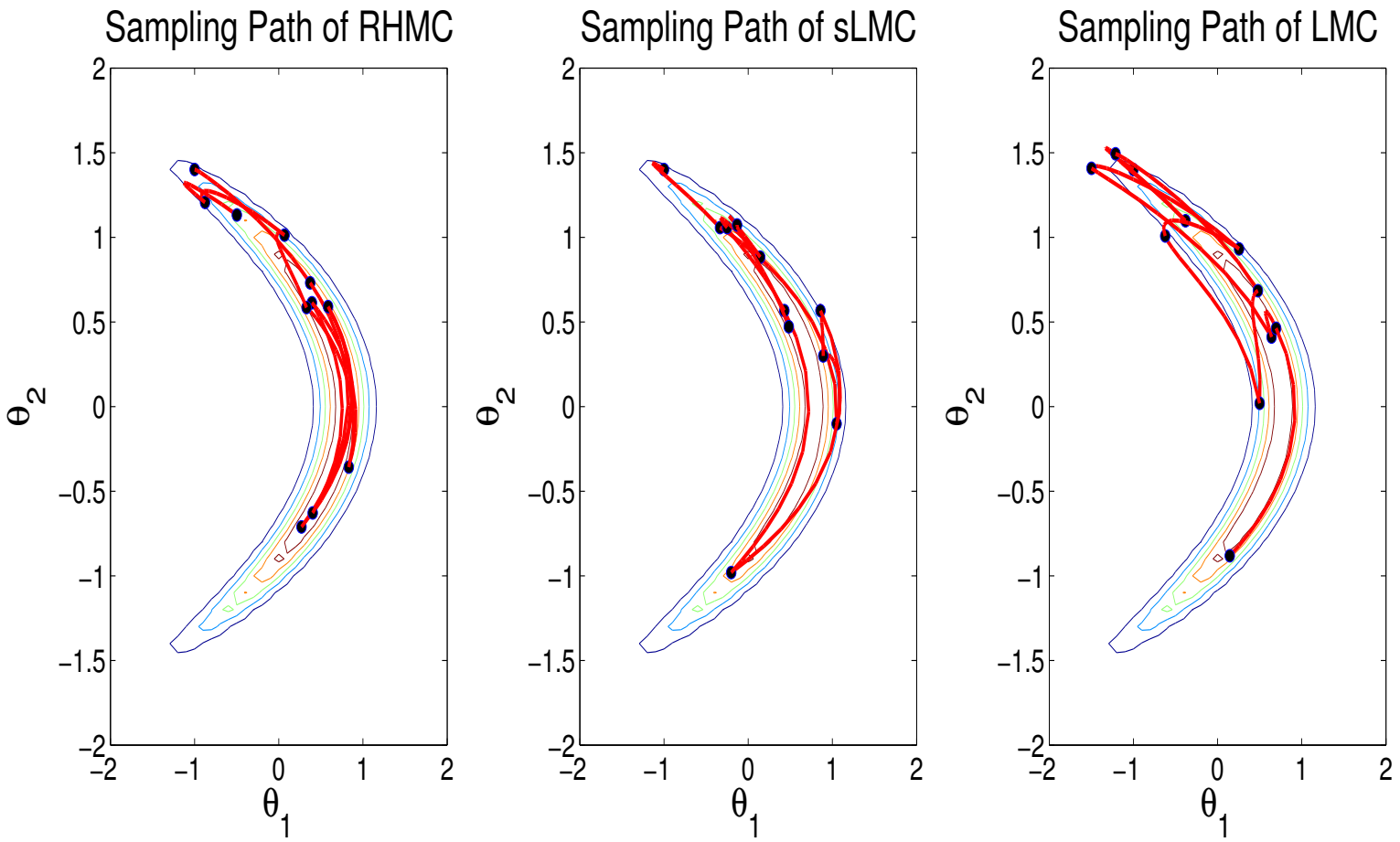
Split HMC with Partial Analytic Solution



Split HMC by Splitting Data



Riemannian HMC and Lagrangian Monte Carlo



Riemannian Hamiltonian Dynamics

On the manifold $\{\pi(\cdot|\boldsymbol{\theta})\}$ with metric $G(\boldsymbol{\theta}) = \mathrm{E}[-\nabla^2_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta})]$:

$$\begin{aligned} H(\boldsymbol{\theta}, \mathbf{p}) &= U(\boldsymbol{\theta}) + K(\mathbf{p}, \boldsymbol{\theta}) \\ &= -\log \pi(\boldsymbol{\theta}) + \frac{1}{2} \log \det \mathbf{G}(\boldsymbol{\theta}) + \frac{1}{2} \mathbf{p}^{\mathsf{T}} \mathbf{G}(\boldsymbol{\theta})^{-1} \mathbf{p} \\ &\equiv \phi(\boldsymbol{\theta}) + \frac{1}{2} \mathbf{p}^{\mathsf{T}} \mathbf{G}(\boldsymbol{\theta})^{-1} \mathbf{p} \end{aligned}$$

where $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}))$. Girolami and Calderhead (2011) propose:

$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \frac{\partial}{\partial p} H(\boldsymbol{\theta}, \mathbf{p}) = \mathbf{G}(\boldsymbol{\theta})^{-1} \mathbf{p} \\ \dot{\mathbf{p}} &= -\frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p}) = -\nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}) + \frac{1}{2} \mathbf{p}^{\mathsf{T}} \mathbf{G}(\boldsymbol{\theta})^{-1} \partial \mathbf{G}(\boldsymbol{\theta}) \mathbf{G}(\boldsymbol{\theta})^{-1} \mathbf{p} \end{aligned}$$

Generilized Leapfrog

$$\mathbf{p}^{(n+\frac{1}{2})} = \mathbf{p}^{(n)} - \frac{\varepsilon}{2}\left[\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}^{(n)}) - (\mathbf{p}^{(n+\frac{1}{2})})^{\mathsf{T}}\nabla_{\boldsymbol{\theta}}\mathbf{G}^{-1}(\boldsymbol{\theta}^{(n)})\mathbf{p}^{(n+\frac{1}{2})}\right]$$

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} + \frac{\varepsilon}{2}\left[\mathbf{G}^{-1}(\boldsymbol{\theta}^{(n)}) + \mathbf{G}^{-1}(\boldsymbol{\theta}^{(n+1)})\right]\mathbf{p}^{(n+\frac{1}{2})}$$

$$\mathbf{p}^{(n+1)} = \mathbf{p}^{(n+\frac{1}{2})} - \frac{\varepsilon}{2}\left[\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}^{(n+1)}) - (\mathbf{p}^{(n+\frac{1}{2})})^{\mathsf{T}}\nabla_{\boldsymbol{\theta}}\mathbf{G}^{-1}(\boldsymbol{\theta}^{(n+1)})\mathbf{p}^{(n+\frac{1}{2})}\right]$$

- Time reversible
- Volume preserving
- But $\cdots$ time consuming, and occasionally un-stable $\cdots$

Lagrangian Monte Carlo

$$\dot{\theta} = \mathbf{G}(\theta)^{-1}\mathbf{p}$$

$$\dot{\mathbf{p}} = -\nabla_{\theta}\phi(\theta) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\theta)^{-1}\partial\mathbf{G}(\theta)\mathbf{G}(\theta)^{-1}\mathbf{p}$$

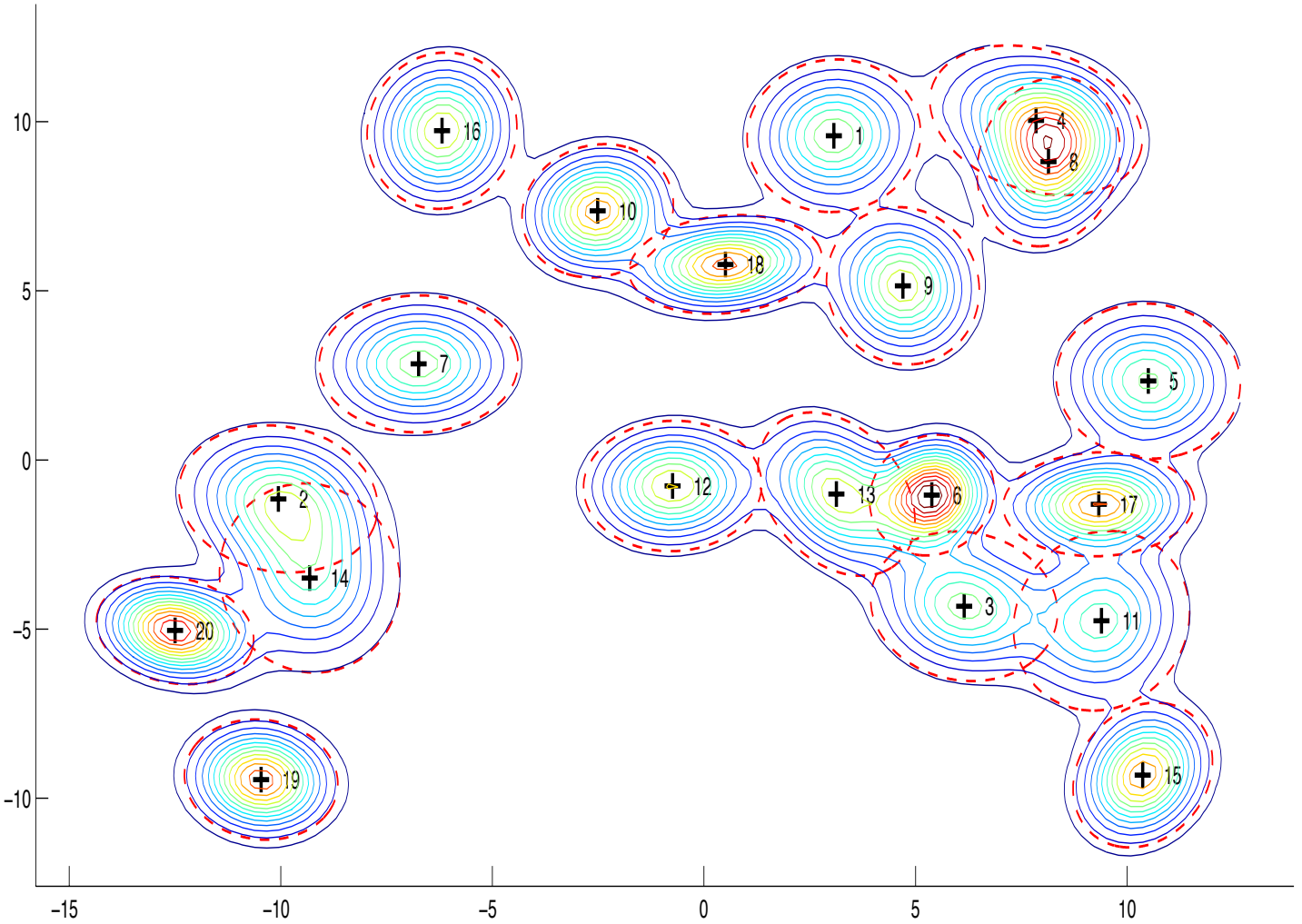
$\boxed{\mathbf{p} \to \mathbf{v}}$ $\Downarrow$ **Lagrangian Dynamics**

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

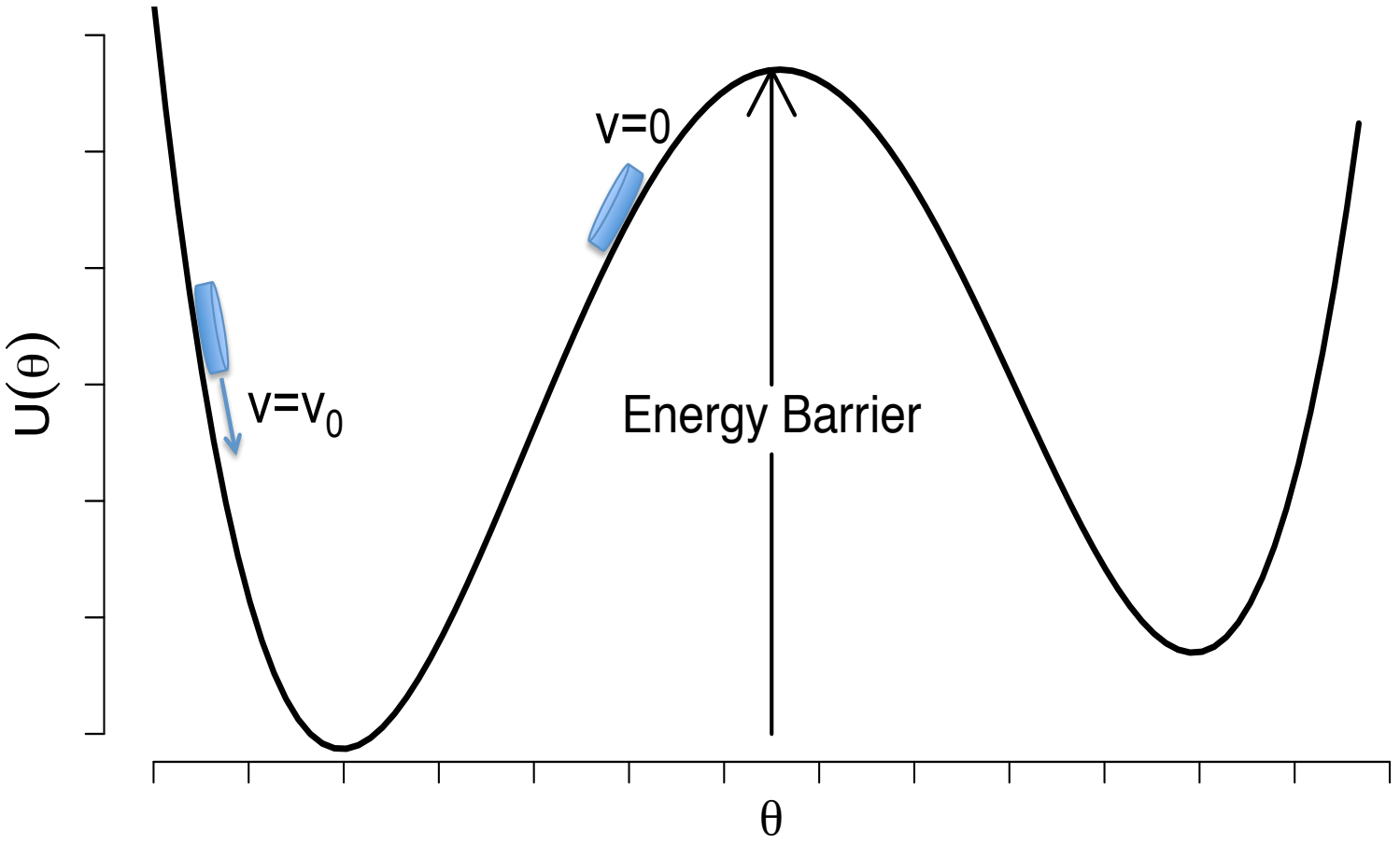
$$\dot{\mathbf{v}} = -\mathbf{v}^{\mathsf{T}}\boldsymbol{\Gamma}(\theta)\mathbf{v} - \mathbf{G}(\theta)^{-1}\nabla_{\theta}\phi(\theta)$$

- Not Hamiltonian dynamics of $(\theta, \mathbf{v})$!
- Computational time is mainly spent on finding direction $\mathbf{v}$.

Wormhole HMC



Energy Barrier



Tunnel Metric

Denote $\mathbf{v_T} := \hat{\theta}_2 - \hat{\theta}_1$ and refer to *tunnel* as a small neighborhood (tube) of $\mathbf{v_T}$. Let $\mathbf{v_T^*} = \mathbf{v_T}/\|\mathbf{v_T}\|$.

Definition 1 (Tunnel Metric)

Given tangent vectors $\mathbf{u}, \mathbf{w}$, a *pseudo tunnel metric* $\mathbf{G_T^*}$ is defined as

$$\mathbf{G_T^*}(\mathbf{u}, \mathbf{w}) := \langle \mathbf{u} - \langle \mathbf{u}, \mathbf{v_T^*} \rangle \mathbf{v_T^*}, \mathbf{w} - \langle \mathbf{w}, \mathbf{v_T^*} \rangle \mathbf{v_T^*} \rangle = \mathbf{u}^\mathsf{T}[\mathbf{I} - \mathbf{v_T^*}(\mathbf{v_T^*})^\mathsf{T}]\mathbf{w} \quad (4.1)$$

Modifying it to be positive definite, we define *tunnel metric* $\mathbf{G_T}$ as

$$\mathbf{G_T} = \mathbf{G_T^*} + \varepsilon \mathbf{v_T^*}(\mathbf{v_T^*})^\mathsf{T} = \mathbf{I} - (1 - \varepsilon)\mathbf{v_T^*}(\mathbf{v_T^*})^\mathsf{T} \quad (4.2)$$

Wind Tunnel and Wormhole

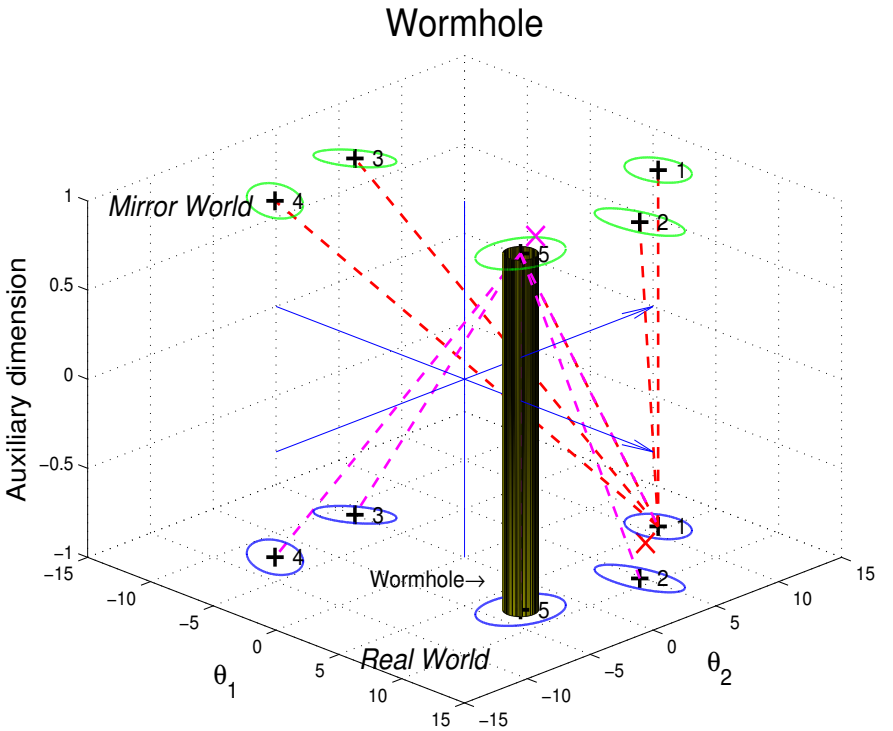
Effect of tunnel metric diminishes

Tunnels interfere with each other

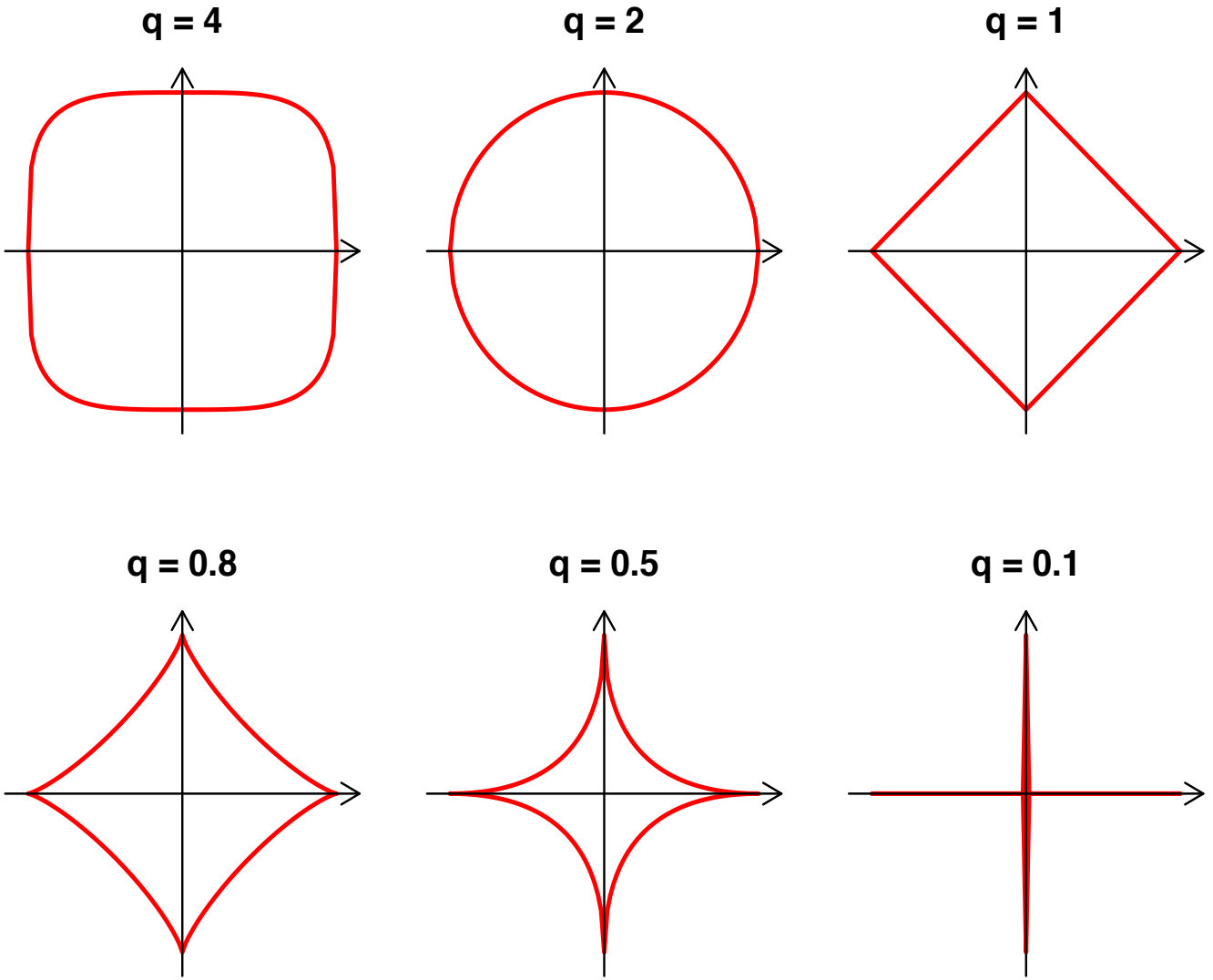
$$\dot{\theta} = \mathbf{v}$$

$$\boxed{\textit{wind vector}} \Rightarrow + \mathbf{f}(\theta, \mathbf{v})$$

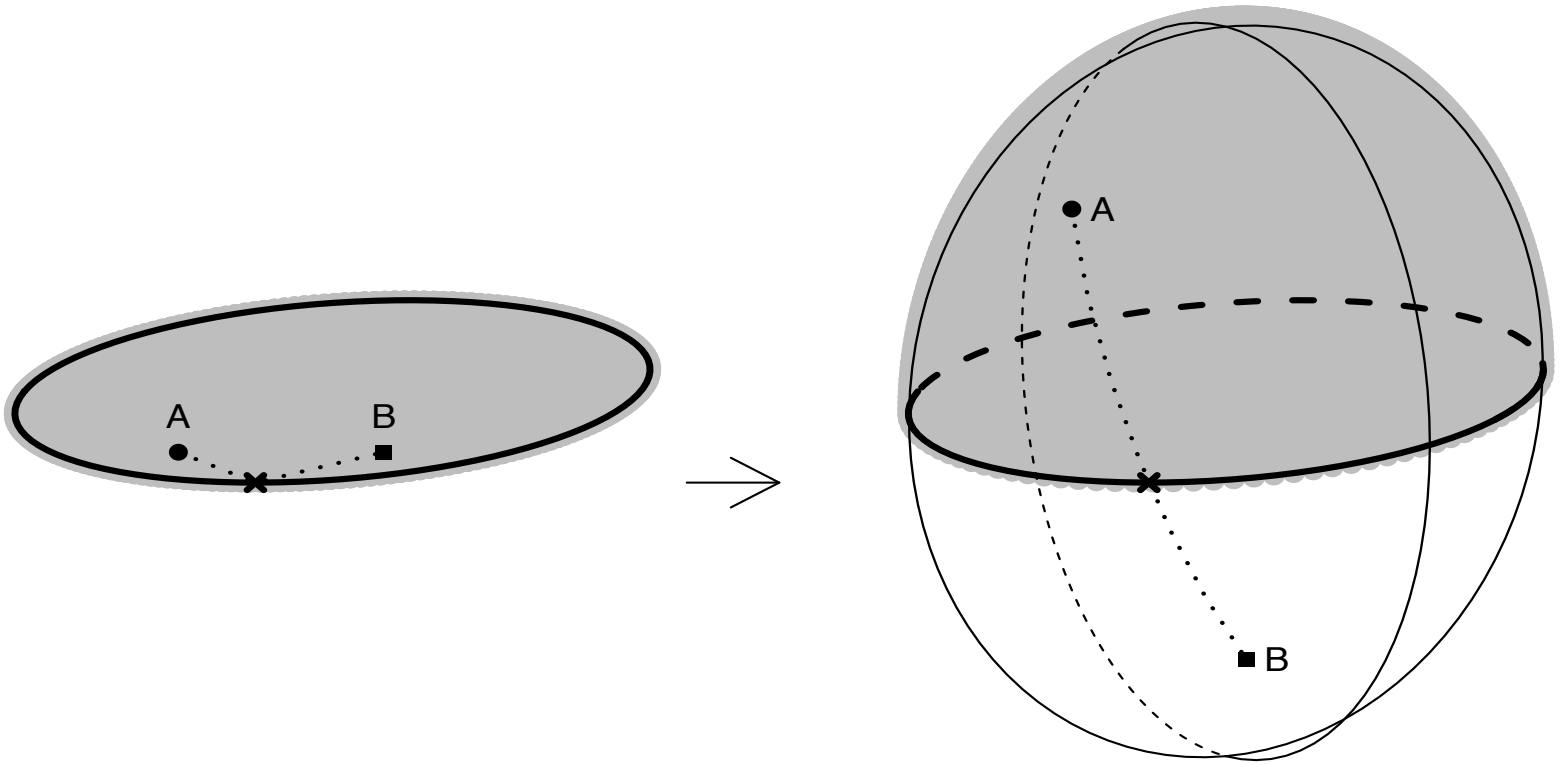
$$\dot{\theta} = \mathbf{v} + \mathbf{f}(\theta, \mathbf{v})$$



Spherical HMC



Change of the domain: from $\mathcal{B}_0^D(1)$ to $\mathcal{S}^D$



Geometry will change: $\mathbf{I} \longrightarrow \mathbf{G}_{\mathcal{S}}(\boldsymbol{\theta})$

Change of variable

$\mathcal{B}_0^D(1) := \{\boldsymbol{\theta} \in \mathbb{R}^D : \|\boldsymbol{\theta}\|_2 = \sqrt{\sum_{i=1}^D \theta_i^2} \leq 1\}$

$\xrightarrow[\theta_{D+1} = \pm\sqrt{1-\|\boldsymbol{\theta}\|_2^2}]{\boldsymbol{\theta} \mapsto \tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \theta_{D+1})}$

$\mathcal{S}^D := \{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{D+1} : \|\tilde{\boldsymbol{\theta}}\|_2 = 1\}$

Change of Variable

$$\int_{\mathcal{B}_0^D(1)} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{B}} = \int_{\mathcal{S}_+^D} \pi(\tilde{\boldsymbol{\theta}}) \left| \frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\tilde{\boldsymbol{\theta}}_{\mathcal{S}}} \right| d\tilde{\boldsymbol{\theta}}_{\mathcal{S}} = \int_{\mathcal{S}_+^D} \pi(\tilde{\boldsymbol{\theta}}) |\theta_{D+1}| d\tilde{\boldsymbol{\theta}}_{\mathcal{S}} \qquad (4.3)$$

where $\pi(\tilde{\boldsymbol{\theta}}) \equiv \pi(\boldsymbol{\theta})$.

What We Want:
$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{B}}$

$\xleftarrow[\text{weigh it by } |\theta_{D+1}|]{\text{drop } \theta_{D+1}}$

What We Sample:
$\tilde{\boldsymbol{\theta}} \sim \pi(\tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}}_{\mathcal{S}}$

Hamiltonian/Lagrangian dynamics on sphere

On $\mathcal{B}_0^D(1)$

$$H(\theta, \mathbf{v}) = U(\theta) + K(\mathbf{v})$$
$$= -\log \pi(\theta) + \frac{1}{2}\mathbf{v}^T \mathbf{I} \mathbf{v}$$

$\boldsymbol{\theta} \mapsto \tilde{\boldsymbol{\theta}}$

On $\mathcal{S}^D$

$$H^*(\tilde{\theta}, \tilde{\mathbf{v}}) = U(\tilde{\theta}) + K(\tilde{\mathbf{v}})$$
$$= -\log \pi(\theta) + \frac{1}{2}\mathbf{v}^T \mathbf{G}_{\mathcal{S}}(\theta)\mathbf{v}$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_D})$$

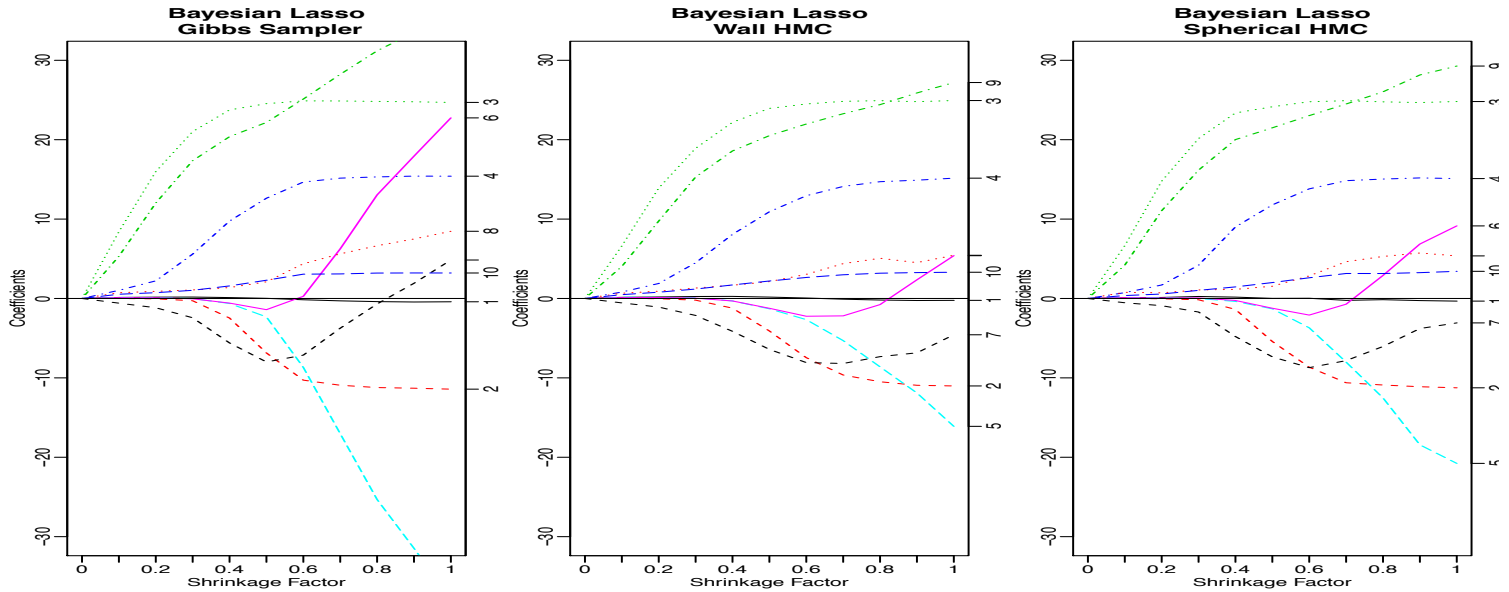
$\mathbf{v} \mapsto \tilde{\mathbf{v}}$

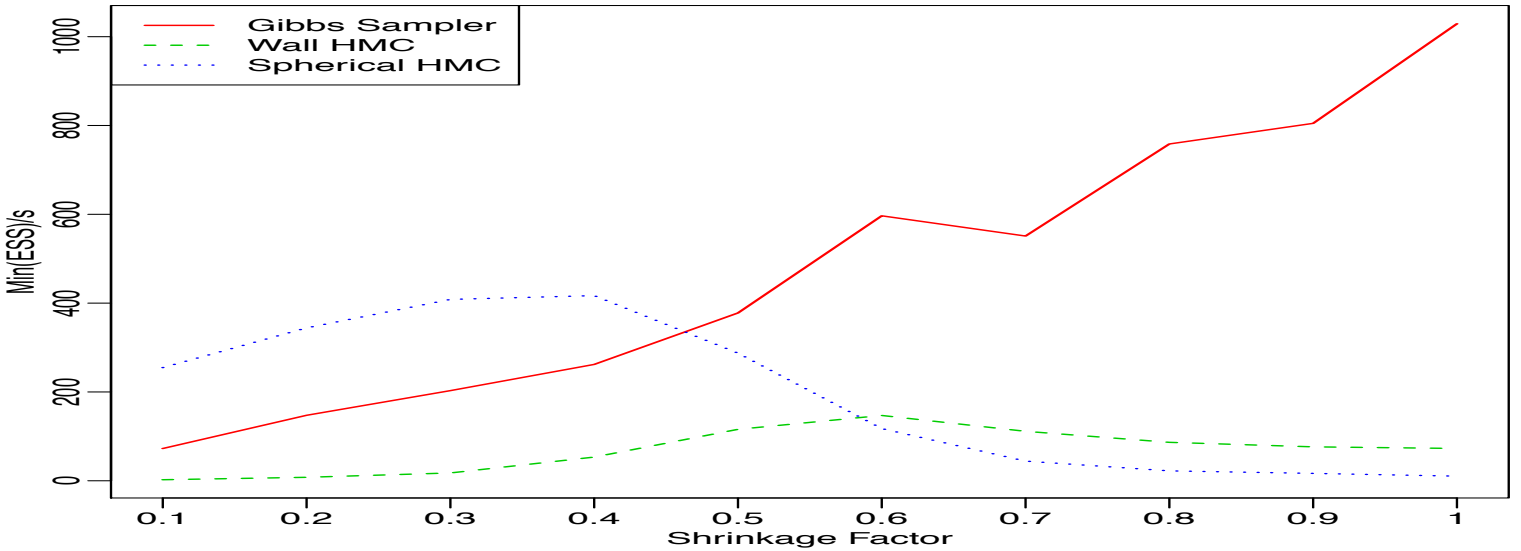
$$\tilde{\mathbf{v}} \sim (\mathbf{I}_{D+1} - \tilde{\theta}\tilde{\theta}^T)\mathcal{N}(\mathbf{0}, \mathbf{I}_{D+1})$$

$$\dot{\theta} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\nabla_\theta U(\theta)$$
$$\|\theta\|_2^2 \leq 1$$

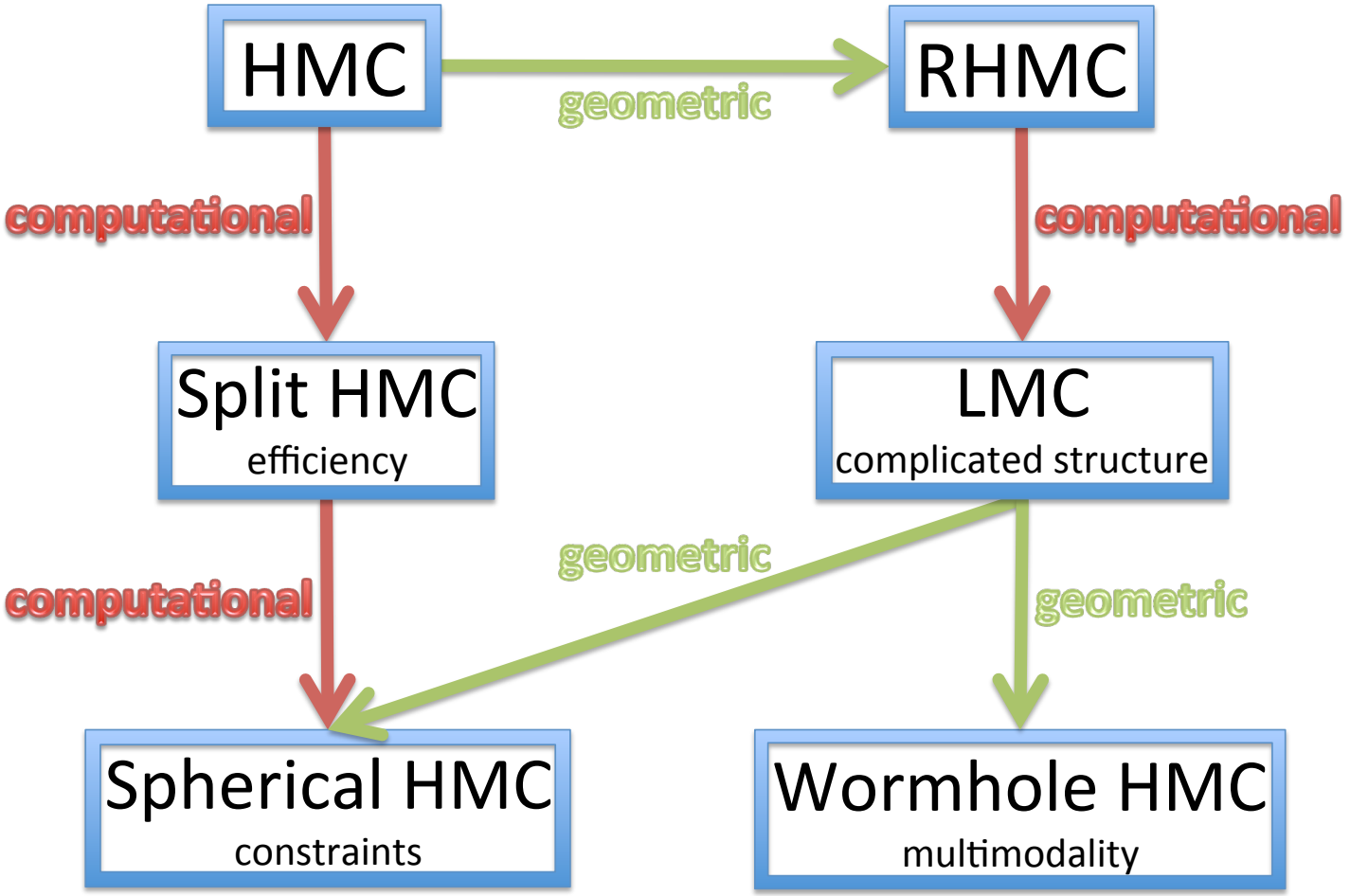
$$\dot{\theta} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\mathbf{v}^T \Gamma(\theta)\mathbf{v} - \mathbf{G}_{\mathcal{S}}(\theta)^{-1}\nabla_\theta U(\theta)$$
$$\theta_{D+1} = \sqrt{1 - \|\theta\|_2^2}, v_{D+1} = -\theta^T \mathbf{v}/\theta_{D+1}$$

Bayesian Lasso





Connections



Reference

1. **Split Hamiltonian Monte Carlo** (2013)
   - Babak Shahbaba, Shiwei Lan, Wesley O. Johnson and Radford M. Neal
   - *Statistics and Computing*, DOI: 10.1007/s11222-012-9373-1.
2. **Lagrangian Dynamical Monte Carlo** (2012)
   - Shiwei Lan, Vassilios Stathopoulos, Babak Shahbaba, and Mark Girolami
   - http://arxiv.org/abs/1211.3759 (accepted by JCGS)
3. **Wormhole Hamiltonian Monte Carlo** (2013)
   - Shiwei Lan, Jeffrey Streets, and Babak Shahbaba
   - http://arxiv.org/abs/1306.0063
4. **Spherical HMC for Constrained Target Distributions** (2013)
   - Shiwei Lan, Bo Zhou, and Babak Shahbaba
   - http://jmlr.org/proceedings/papers/v32/lan14.pdf (ICML 2014)

Thank you !