Hamiltonian Monte Carlo for high dimensional problems

Tijana Radivojević[1] and **Elena Akhmatskaya**[1,2]

[1]BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[2]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

BCAM-IMUVA Summer School on Uncertainty Quantification for Applied Problems,

Bilbao, Spain
July 7, 2016

Preview

- We introduce **an alternative to Hamiltonian Monte Carlo** (HMC) for efficient sampling in computational statistics

- The new method called Mix & Match Hamiltonian Monte Carlo (MMHMC) has been inspired by **Generalized Shadow Hybrid Monte Carlo** (GSHMC) by Akhmatskaya & Reich

- MMHMC:
  - ✓ is **generalized HMC**
  - ✓ is **HMC importance sampler** that samples with **modified Hamiltonians**
  - ✓ offers computationally effective **Metropolis test for momentum update**
  - ✓ **reduces** potential negative effects of **momentum flips**
  - ✓ relies on **method- and system- specific adaptive integrators** and compatible modified Hamiltonians

Behind the scenes

- 2008 – 2011: GSHMC was
  - ✓ **introduced** for sampling in molecular simulation
  - ✓ **published**: *Akhmatskaya, Reich (2008), JCOMP, 227, 4934; Akhmatskaya, Bou-Rabee, Reich (2009), JCOMP, 228 (6), 2256*
  - ✓ **patented**: GB patent (2009), US patent (2011) [*Fujitsu, Authors: Akhmatskaya, Reich*]
  - ✓ **proved** to be successful in simulations of complex molecular systems in Biology and Chemistry (6 publications)

✗ No implementation and testing in statistical computation till 2015

✗ Never been implemented in open source software due to patent restrictions

- **November 2015**: Fujitsu issued the license giving a permission

  (i) to use the patented method in open source software

  (ii) to EA to implement and use know-how

- **Current status**: GSHMC has been modified and adapted to computational statistics, leading to MMHMC. Implemented in BCAM in-house software *HAICS* [*Radivojević, Akhmatskaya, preprint*]

MMHMC – features

- For the **target density** $\pi(\boldsymbol{\theta}) \propto L(\boldsymbol{\theta}|\mathbf{y})p(\boldsymbol{\theta})$ of position vector $\boldsymbol{\theta}$ given data $\mathbf{y}$, where $L(\boldsymbol{\theta}|\mathbf{y})$ is the likelihood and $p(\boldsymbol{\theta})$ the prior, momenta $\mathbf{p}$ conjugate to $\boldsymbol{\theta}$, with a 'mass' matrix M (a preconditioner) construct a **potential function**

$$U(\boldsymbol{\theta}) = -\log L(\boldsymbol{\theta}|\mathbf{y}) - \log p(\boldsymbol{\theta})$$

  and a **Hamiltonian**

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

- Sampling is performed with respect to an **importance sampling density**

$$\tilde{\pi}(\boldsymbol{\theta}, \mathbf{p}) \propto \exp(-\tilde{H}(\boldsymbol{\theta}, \mathbf{p}))$$

- E. g., the **modified Hamiltonian** of order $m = 4$ for the Verlet method

$$\tilde{H}(\boldsymbol{\theta}, \mathbf{p}) = H(\boldsymbol{\theta}, \mathbf{p}) + \frac{h^2}{24}\left(2\mathbf{p}^{\intercal} M^{-1} U_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}) M^{-1}\mathbf{p} - U_{\boldsymbol{\theta}}(\boldsymbol{\theta})^{\intercal} M^{-1} U_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right)$$

Modified / Shadow Hamiltonians

- Comparatively cheap and arbitrary accurate approximations of Hamiltonians

$$\tilde{H} = H + \mathcal{O}(h^m), m \geq 4$$

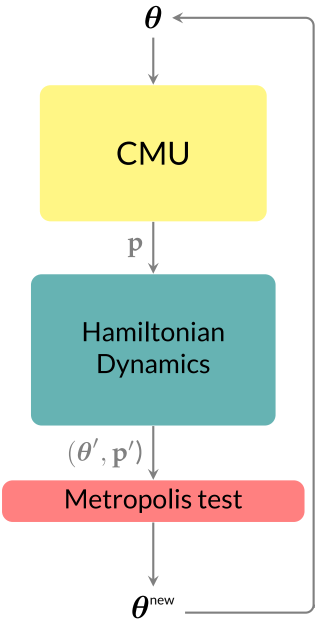
- Defined by an asymptotic expansion in powers of the discretization parameter h
- Exact (by construction) for quadratic Hamiltonians
- Stay close to true Hamiltonians for short HD simulations (such as in HMC)
- Conserved by symplectic integrators to higher accuracy than true Hamiltonian

  $\mathbb{E}(\Delta H) = \mathcal{O}(Dh^{2p})$, p is an order of an integrator ($p = 2$ for Verlet), D is dimension

  $\mathbb{E}(\Delta \tilde{H}) = \mathcal{O}(Dh^{2m})$, $m \geq 4$ is the order of modified Hamiltonian

- The ways to construct approximations to modified Hamiltonians are defined by a choice of the integrator

Recap: HMC algorithm



Complete Momentum Update (CMU)

Randomly assigns momentum from Gaussian distribution

Hamiltonian Dynamics

Integrates for L steps the Hamiltonian equations using a symplectic numerical integrator with a step size h generating proposal

$$(\boldsymbol{\theta}', \mathbf{p}') = \Psi_\tau(\boldsymbol{\theta}, \mathbf{p}),$$

Ψ_τ is the time-reversible symplectic map, $\tau = Lh$

Metropolis test

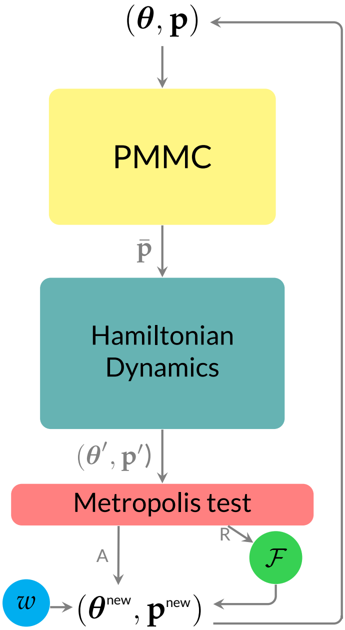
$$\boldsymbol{\theta}^{new} = \begin{cases} \boldsymbol{\theta}' & \text{with probability } \alpha \\ \boldsymbol{\theta} & \text{otherwise} \end{cases}$$

$\alpha = \min\{1, \exp(-\Delta H)\}$

$\Delta H = H(\boldsymbol{\theta}', \mathbf{p}') - H(\boldsymbol{\theta}, \mathbf{p}) \neq 0$

	HMC	MMHMC
Momentum update	complete	partial
Momentum Metropolis test	✗	✓
Metropolis test	H	$\tilde{H}$
Momentum flip	✗	✓
Re-weighting	✗	✓

MMHMC – algorithm



Partial Momentum Monte Carlo (PMMC)

$$\bar{\mathbf{p}} = \begin{cases} \sqrt{1-\varphi}\,\mathbf{p} + \sqrt{\varphi}\,\mathbf{u} & \text{with probability } \mathcal{P} = \min\{1, \exp(-\Delta\hat{H})\} \\ \mathbf{p} & \text{otherwise} \end{cases}$$

$\mathbf{u} \sim \mathcal{N}(0, M)$ is the noise, $\varphi \in (0, 1]$

$$\Delta\hat{H} = \frac{h^2(6b-1)}{24}\left(\varphi A + 2\sqrt{\varphi(1-\varphi)}B\right)$$

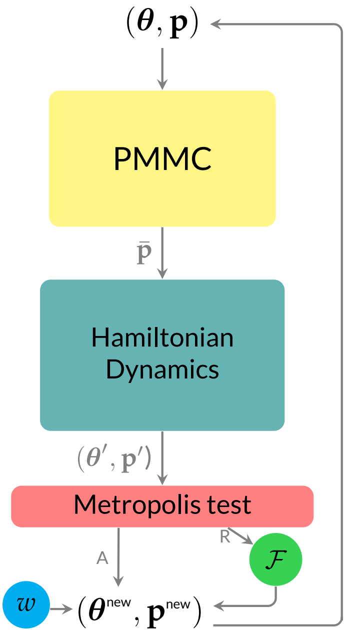
$A = \mathbf{u}^\intercal M^{-1} U_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}) M^{-1}\mathbf{u} - \mathbf{p}^\intercal M^{-1} U_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}) M^{-1}\mathbf{p}$
$B = \mathbf{u}^\intercal M^{-1} U_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}) M^{-1}\mathbf{p}$
b is the integrator's parameter

The extended "Hamiltonian"

$$\hat{H}(\boldsymbol{\theta}, \mathbf{p}, \mathbf{u}) = \tilde{H}(\boldsymbol{\theta}, \mathbf{p}) + \frac{1}{2}\mathbf{u}^\intercal M^{-1}\mathbf{u}$$

defines the extended reference density
$\hat{\pi}(\boldsymbol{\theta}, \mathbf{p}, \mathbf{u}) \propto \exp(-\hat{H}(\boldsymbol{\theta}, \mathbf{p}, \mathbf{u}))$

MMHMC – algorithm



Metropolis test

$$(\boldsymbol{\theta}^{\text{new}}, \mathbf{p}^{\text{new}}) = \begin{cases} (\boldsymbol{\theta}', \mathbf{p}') & \text{accept with prob. } \alpha \\ \mathcal{F}(\boldsymbol{\theta}, \mathbf{p}^*) & \text{reject otherwise} \end{cases}$$

$$\alpha = \min\left\{1, \exp(-\Delta \tilde{H})\right\}$$

$$\text{Flip } \mathcal{F}(\boldsymbol{\theta}, \mathbf{p}) = \begin{cases} (\boldsymbol{\theta}, -\mathbf{p}) \\ \text{reduced flip (optionally)} \end{cases}$$

Re-weighting (w)

For every $n = 1, \ldots, N$ stores

$$w_n = \exp\left(\tilde{H}(\boldsymbol{\theta}_n, \mathbf{p}_n) - H(\boldsymbol{\theta}_n, \mathbf{p}_n)\right)$$

$$\hat{f} = \frac{\sum_{n=1}^{N} f(\boldsymbol{\theta}_n) w_n}{\sum_{n=1}^{N} w_n} \xrightarrow[N \to \infty]{} \mathbb{E}(f)$$

What to expect?

Pros – enhanced sampling

- High acceptance rates – $\tilde{H}$ conserved by symplectic integrators better than H
- Access to second-order information about the target distribution
- Extra parameter for performance enhancing

Cons – extra computational cost

- Computation of $\tilde{H}$ for each proposal (higher orders $\tilde{H}$ are more expensive)
- Extra Metropolis test for momentum update
- Accurate numerical integrators required to use low orders $\tilde{H}$ for systems with highly oscillatory H

Our strategy

To find the numerical integrator that provides the best conservation of modified Hamiltonian.
Search within the family of 2-stage splitting integrating schemes.

Modified Hamiltonians for splitting integrators

4-th order modified Hamiltonian for **2-stage splitting integrators** were derived in terms of **quantities available during simulation**

$$\tilde{H}(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p}) + h^2\Big(\alpha \mathbf{p}^\intercal M^{-1} \dot{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \beta U_{\boldsymbol{\theta}}(\boldsymbol{\theta})^\intercal M^{-1} U_{\boldsymbol{\theta}}(\boldsymbol{\theta})\Big)$$

$\dot{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ - numerical time-derivative of $U_{\boldsymbol{\theta}}(\boldsymbol{\theta})$

$$\alpha = \frac{6b-1}{24}, \quad \beta = \frac{6b^2-6b+1}{12}$$

b - parameter of an integrator

Adaptive integrators for optimal conservation of modified Hamiltonian

Consider one parameter **2-stage splitting integrator** of Hamiltonian system with $H(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + U(\boldsymbol{\theta}) = A + B$:

$$\psi_h = \varphi^B{}_{bh} \circ \varphi^A{}_{\frac{h}{2}} \circ \varphi^B{}_{(1-2b)h} \circ \varphi^A{}_{\frac{h}{2}} \circ \varphi^B{}_{bh}$$

We extended Adaptive Integration Approach (AIA) [*Fernandez-Pendas, Akhmatskaya, Sanz-Serna*] to statistical applications and to sampling with **modified** Hamiltonians

Given step size h find

$$b^*(h, f) \ (\equiv \text{system specific integrator})$$

that minimizes expectation of the modified energy error

$$\Delta = \tilde{H}^{[4]}(\Psi_{h,L}(\boldsymbol{\theta}, \mathbf{p})) - \tilde{H}^{[4]}(\boldsymbol{\theta}, \mathbf{p}), \quad \text{i. e.}$$

$$0 \leq \mathbb{E}_{\tilde{\pi}}(\Delta) \leq \rho(h, b^*) \to \text{minimal}$$

where

$$f = f\left(\langle H \rangle_V, \langle \Delta H \rangle_V, h_V\right)$$

are emulated highest frequencies obtained using a warm-up HMC (Verlet) and the underlying AIA analysis

Adaptive integrators for optimal conservation of modified Hamiltonian

$$\rho(h, b) = \frac{(SB_h + C_h)^2}{2S\left(1 - A_h^2\right)}$$ (specially derived for MMHMC)

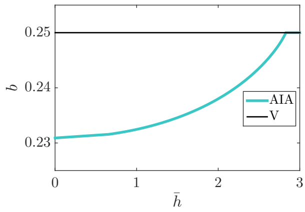
$$\begin{aligned} A_h &= \frac{h^4 b(1-2b)}{4} - \frac{h^2}{2} + 1 \\ B_h &= -\frac{h^3(1-2b)}{4} + h \\ C_h &= -\frac{h^5 b^2(1-2b)}{4} + h^3 b(1-b) - h \\ S &= \frac{1+2h^2\beta}{1+2h^2\alpha} \end{aligned}$$

Then:

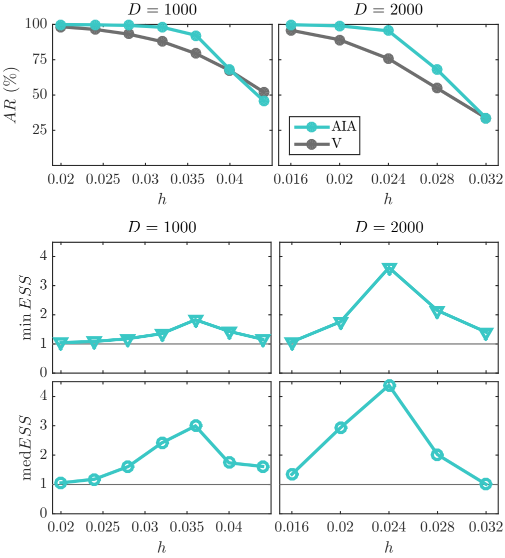
$$b^*(h, f) = \arg\min_{b \in B} \max_{0<h<\bar{h}} \rho(h, b)$$

where $\bar{h}$ (reduced step size) is a function of f

AIA in action



V – velocity Verlet, $b = 0.25$



Efficiency improvement over Verlet in sampling a multivariate Gaussian distribution

(AIA up to **4.5x** more efficient than Verlet)

Benchmark Models

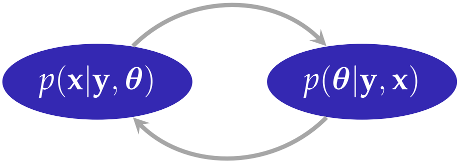
- Gaussian distribution

  $D = 100, 1000, 2000$

- Bayesian Logistic Regression

Dataset	# of param (D)	# of obs (K)
german	25	1000
sonar	61	208
musk	167	476
secom	444	1567

- Stochastic Volatility



Simulated data
$d = 2000$

Lichman, M. (2013), UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

Testing

Methods
MMHMC
HMC
GHMC
RMHMC

Criteria
Sampling efficiency

Metrics

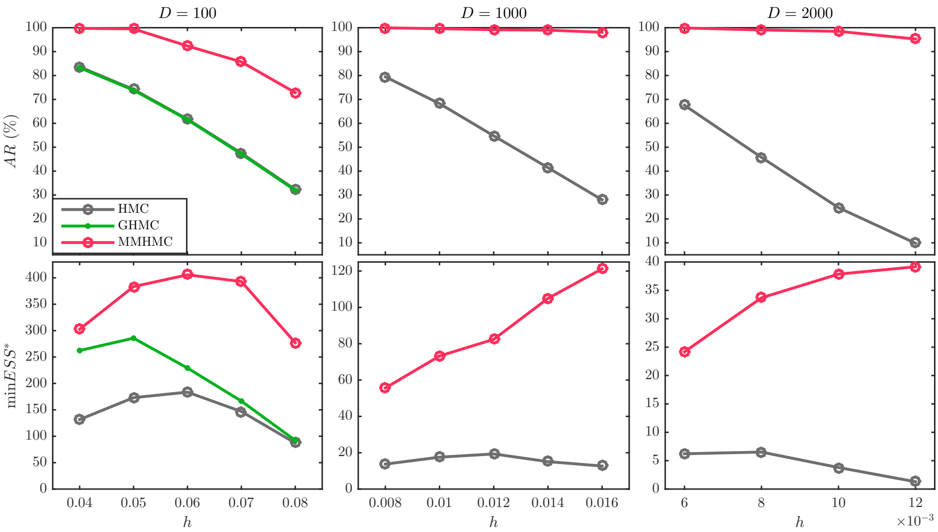
- AR – acceptance rate
- ESS* - time normalized effective sample size
- EF – efficiency factor (relative ESS* w.r.t HMC)

- Results averaged over 10 runs
- Choice of simulation parameters – each method tuned for the best performance

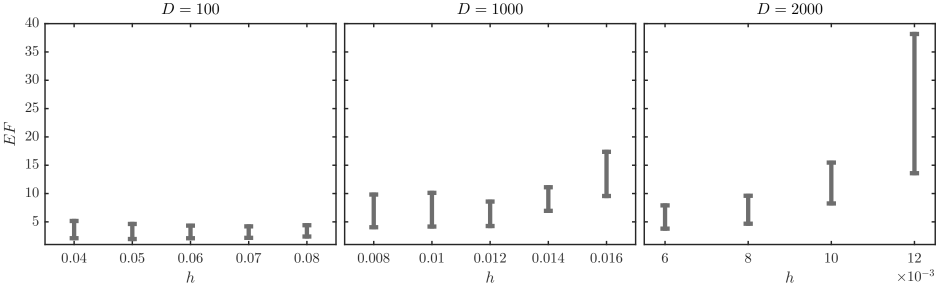
A. D. Kennedy and B. Pendleton (2001), Cost of the Generalised Hybrid Monte Carlo Algorithm for Free Field Theory, Nuclear Physics B, 607: 456–510

M. Girolami, B. Calderhead (2011), Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods, Journal of the Royal Statistical Society: Series B, 73(2):123–214

Gaussian distribution

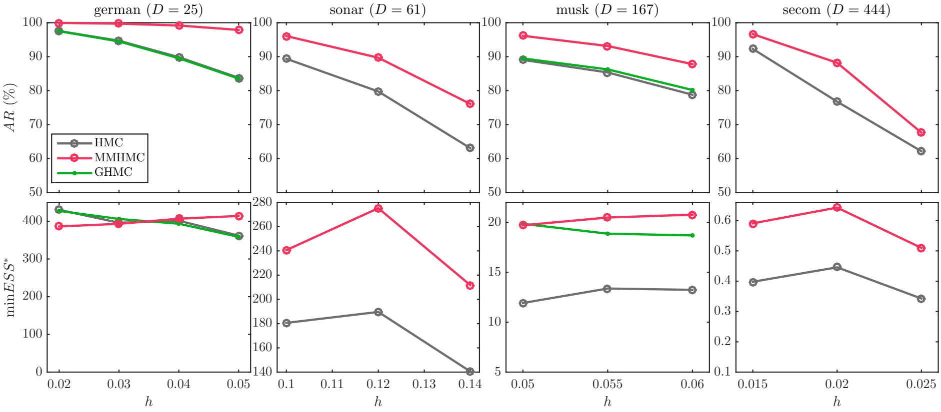


Gaussian distribution – MMHMC vs HMC



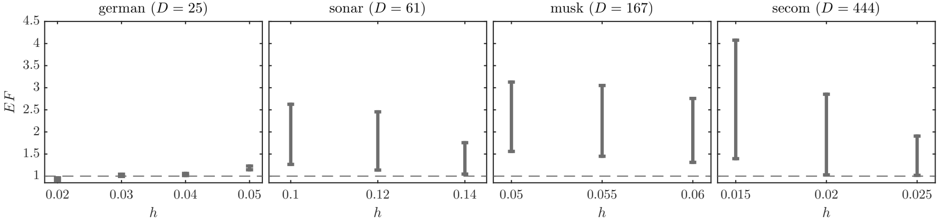
- MMHMC demonstrates from $\mathbf{2}\times$ up to $\mathbf{40}\times$ higher sampling efficiency compared to HMC
- Improvement increases with dimension

Bayesian Logistic Regression (BLR)



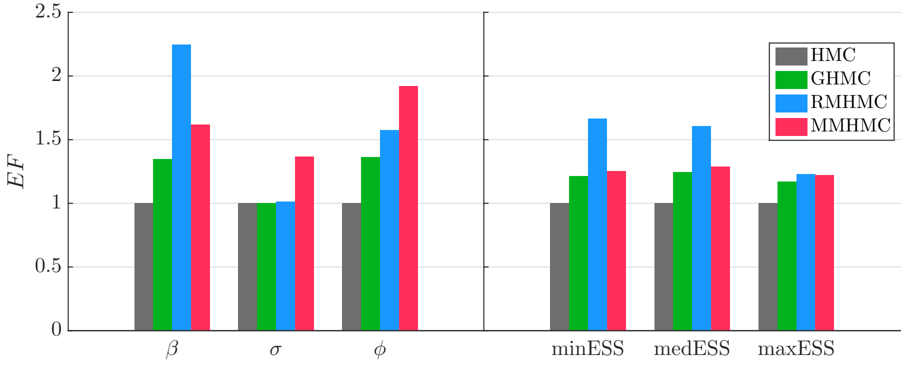
- MMHMC does not use AIA in these tests (testing in progress)

BLR – MMHMC vs HMC



- Comparable performance for the smallest dimension
- Improvement with MMHMC increases with dimension (up to $\mathbf{4}\times$)

Stochastic volatility (SV): d=2000



- All GHMC, RMHMC, MMHMC outperform HMC
- MMHMC and RMHMC are comparable
  (MMHMC is not more than 28% less efficient than RMHMC or up to 35% more efficient than RMHMC)
- MMHMC does not use AIA in these tests (testing in progress)

MMHMC for Uncertainty Quantification (UQ)

- We have learned this week that
  - for highly non-linear and non-Gaussian data assimilation problems
  - when an accurate characterisation of the posterior pdf is needed

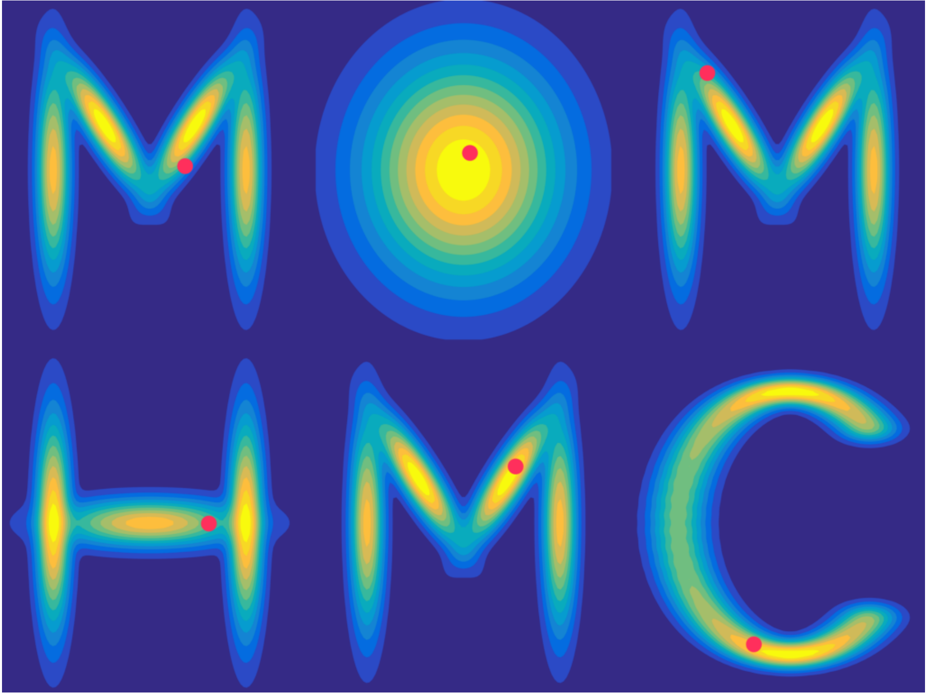
  HMC is a good candidate for UQ problems
- MMHMC, being more efficient than HMC, is also a good candidate

Summary

- MMHMC demonstrates higher AR, bigger ESS and faster convergence than HMC, GHMC for a range of applications and dimensions.
  The improvements are more dramatic for high dimensional problems

- MMHMC and RMHMC demonstrate comparable sampling performance for the tested system.
  In contrast to RMHMC, MMHMC
  - does not require matrix inversion (computationally less expensive)
  - relies on separable Hamiltonians - allows for use of new, more efficient numerical integrators
  - efficient for high dimensions problems

- MMHMC is a good candidate for UQ problems



Modification of the HarlMCMC code by Broderick&Duvenaud (https://www.youtube.com/watch?v=Vv3f0QNWvWQ)