# Parameterisation and convergence of Langevin sampling algorithms

Benedict Leimkuhler
University of Edinburgh

Mostly work with **C. Matthews,** or with
**Peter Whalley** (UoE PhD Student) and **Daniel Paulin** (UoE Statistics)

Brin MRC "Rare Events", 2023

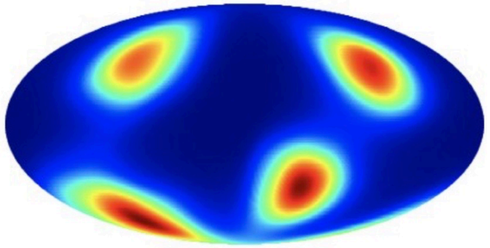
# The sampling problem

How best to calculate averages of nice functions with respect to a well-defined probability distribution (typically in high dimensions).

e.g. in Euclidean space

$$\langle \varphi \rangle = \int_\Omega \varphi(x)\rho(x)\mathrm{d}x, \qquad \Omega \subset \mathbb{R}^d, \qquad \int_\Omega \rho(x)\mathrm{d}x = 1$$

or with respect to a smooth probability distribution on a differentiable manifold $\mathcal{M} \subset \mathbb{R}^d$



**Applications: molecular modelling, statistics, machine learning...**

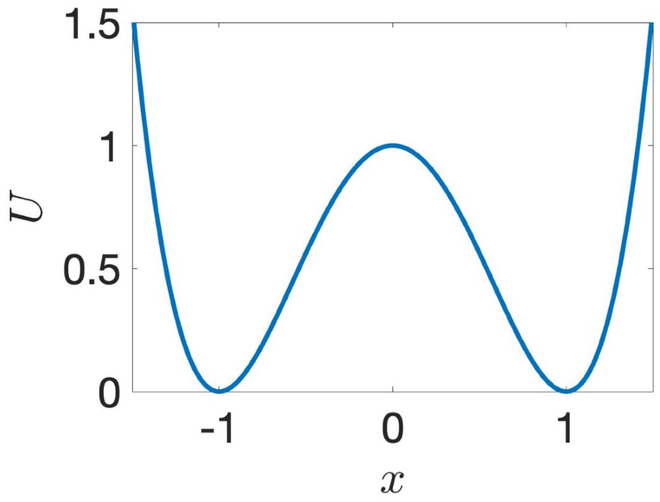
# Overdamped Langevin

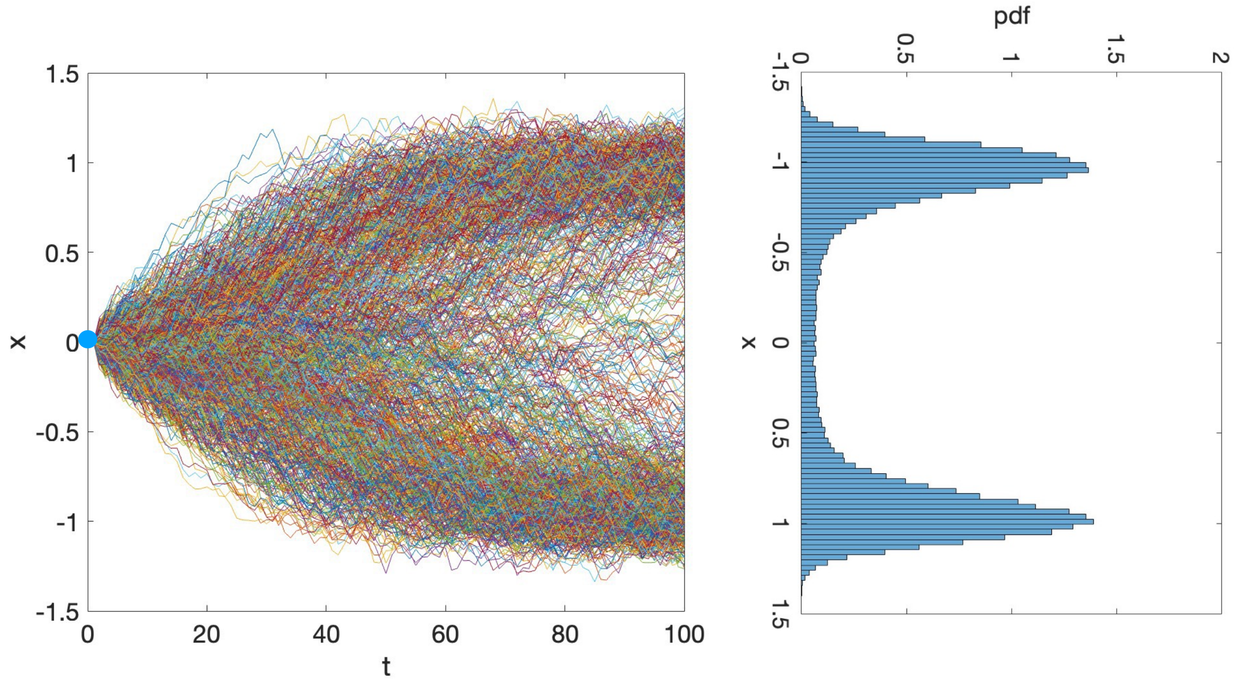
Overdamped Langevin $\mathrm{d}x = -\nabla U(x)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{dW}$

converges to the Gibbs-Boltzmann distribution with density

$$\rho_\beta \propto \exp -\beta U(x)$$

under assumptions of confinement and smoothness on the potential $U$



# SDE-based sampling

$$\mathrm{d}x = -\nabla U(x)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{dW}$$

**Discretization (Euler-Maruyama)**

$$x_{n+1} = x_n - h\nabla U(x_n) + \sqrt{2\beta^{-1}h}R_n, \qquad R_n \sim \mathcal{N}(0,1)$$

**First order weak MCMC procedure ( $O(h)$ perturbed invariant measure)**

**"BAOAB Limit" (Leimkuhler-Matthews, AMRX, 2013):**

$$x_{n+1} = x_n - h\nabla U(x_n) + \sqrt{\beta^{-1}h/2}(R_n + R_{n+1}), \qquad R_n \sim \mathcal{N}(0,1)$$

**successive Gaussian increments**

***2nd order weak*** $\lim_{K\to\infty} \left| K^{-1} \sum_{k=0}^{K-1} \varphi(x_k) - \langle\varphi\rangle \right| = O(h^2)$

# Assumptions

1.**Continuous** system: Markov process with transition kernel $\mathscr{P}_t$ and stationary distribution $\pi$

$$\|\nu\mathscr{P}_t - \pi\|_{\mathrm{TV}} \leq \mathrm{d}(\nu, \pi)e^{-\lambda t}$$

2. **Discrete** system: defined by a numerical method $\hat{\mathscr{P}}_{\Delta t,\gamma,\ldots}$ with corresponding stationary distribution $\pi_{\Delta t,\gamma,\ldots}$

$$\nu\hat{\mathscr{P}}^n_{\Delta t,\gamma,\ldots} \to \pi_{\Delta t,\gamma,\ldots} \quad n \to \infty$$

Both the propagator and the stationary distribution depend on various properties and parameters from the target system, the formulation, and the numerical method

# Outline

Affect of **stepsize and friction** on the ***long run accuracy*** and ***convergence rate*** of Langevin dynamics methods

Practical experience in **constrained** Langevin dynamics simulations

**Removing the duration parameter** in constrained Hamiltonian Monte Carlo using randomization of time

# Langevin dynamics

$$\boxed{\begin{aligned} \mathrm{d}x &= M^{-1}p\mathrm{d}t \\ \mathrm{d}p &= -\nabla U\,\mathrm{d}t \end{aligned}} - \boxed{\gamma M^{-1}p\mathrm{d}t + \sqrt{2\beta^{-1}\gamma}\mathrm{dW}}$$

*Newton's Equations* *Dissipative-Stochastic Perturbation*

With Periodic Boundary Conditions and smooth potential, **ergodic sampling** of the canonical (Gibbs) distribution with density

$$\rho_\beta(x,p) \propto \exp(-\beta H(x,p)); \qquad H(x,p) = p^T M^{-1} p/2 + U(x)$$

# Convergence

Convergence theory $\rho(\,\cdot\,, t) \to \rho_\beta$ due to **Talay**, **Herau**, **Nier**, **Hairer**, **Mattingly**, **Villani**, **Dolbeaut**, **...** (and many others!)

Lots of different concepts for convergence!  It's a "rich area" of mathematics research.

One typically looks for results that will be **general** and **relevant.**  A good compromise: convergence in the sense of "**Wasserstein distance**" to a target measure.

D. Griffeath, Coupling Methods for Markov Processes, 1975

*Over recent years, a "coupling method" for proving ergodicity results has been developed by Vasershtein, and many others. The technique involves constructing two copies of X, which start from different states and evolve simultaneously in such a way that if they ever reach the same state, then they are "pasted together" from that time on.*

# Discretization: splitting methods

$$\mathrm{d}x = \overbrace{p\mathrm{d}t}^{\mathbf{A}}$$

$$\mathrm{d}p = \overbrace{-\nabla U\,\mathrm{d}t}^{\mathbf{B}} \overbrace{- \gamma p \mathrm{d}t + \sqrt{2\gamma\beta^{-1}}\mathrm{dW}}^{\mathbf{O}}$$

Compose steps with different parts of the SDE system, using exact weak sense propagation...

**OBABO - O step (h/2) → B step (h/2) → A step (h) → B step(h/2) → O step (h/2)**

**Bussi and Parrinello, 2007**

**Stochastic Exponential Euler (SES): (B+O)A - solve B and O together, then solve A**

**Ermak and Bucholz, 1980**

# Convergence in Wasserstein distance

$$dX_t = V_t dt,$$
$$dV_t = -\nabla U(X_t)dt - \gamma V_t dt + \sqrt{2\gamma} dW_t \qquad \text{(LD)}$$

$$(M\text{-Lipschitz})\ \exists M : \|\nabla U(x) - \nabla U(y)\| \leq M\|x-y\|$$
$$(m\text{-convex})\ \langle \nabla U(x) - \nabla U(y), x-y\rangle \geq m\|x-y\|^2$$

Theorem (**Wasserstein... Villani, Dalalyan, Durmus, Moulines, Monmarché...**) contractivity of a discretization of (LD) implies convergence in Wasserstein distance.

$$\mathcal{W}_p^2(\nu^{n+1}, \mu^{n+1}) \leq \theta(h)\mathcal{W}_p^2(\nu^n, \mu^n) \qquad \theta(h) < 1$$

# Convergence results for splitting methods

**Leimkuhler, Paulin, Whalley 2023**, Contraction and Convergence Rates for Discretized Kinetic Langevin Dynamics, ArXiv https://arxiv.org/abs/2302.10684

$\gamma \gtrsim \sqrt{M}$

| Algorithm | stepsize restriction | optimal one-step contraction rate |
|---|---|---|
| EM | $\mathcal{O}(1/\gamma)$ | $\mathcal{O}(m/M)$ |
| BAO, OBA, AOB | $\mathcal{O}(1/\sqrt{M})$ | $\mathcal{O}(m/M)$ |
| OAB, ABO, BOA | $\mathcal{O}(1/\gamma)$ | $\mathcal{O}(m/M)$ |
| BAOAB | $\mathcal{O}(1/\sqrt{M})$ | $\mathcal{O}(m/M)$ |
| OBABO | $\mathcal{O}(1/\sqrt{M})$ | $\mathcal{O}(m/M)$ |
| SES | $\mathcal{O}(1/\gamma)$ | $\mathcal{O}(m/M)$ |

| Algorithm | Previous stepsize restriction | previous explicit best rate |
|---|---|---|
| OBABO | $\mathcal{O}(m/\gamma^3)$ | $\mathcal{O}(m^2/M^2)$ [1́] |
| SES | $\mathcal{O}(1/\gamma)$ | $\mathcal{O}(m/M)$ [2́] |

[1] **P. Monmarche** 2021

[2] **J.M. Sanz-Serna** and **K. Zygalakis**, 2021, using **A. S. Dalalyan** and **L. Riou-Durand**, 2020

# idea of proof

Consider "synchronously coupled" discrete paths $(x_k, p_k),\ (\tilde{x}_k, \tilde{p}_k)$

$$\nabla U(\tilde{x}_k) - \nabla U(x_k) = Q\overline{x}$$

*mean value theorem for vector functions*

Define $$P = \begin{pmatrix} I & hI \\ -hQ & (1-\gamma h)\, I \end{pmatrix}$$

*Example for Euler-Maruyama*

Enough to show $(1 - c(h))M - P^T M P$ is positive definite

$$M = \begin{bmatrix} I & bI \\ bI & aI \end{bmatrix}$$

find $a$ and $b$

# $\gamma$-limit convergence (GLC)

One important limit is that of 'high friction', $\gamma \to \infty$

With a fixed stepsize, this limit is well-defined for Langevin integrators, but not all schemes perform well in this limit

(i.e. most do not converge, after $\Delta t \to 0$, to **overdamped** Langevin dynamics)

| method | limiting scheme | GLC? |
|---|---|---|
| **BAO** | $x_{k+1} = x_k - h^2 \nabla U(x_k) + h\xi_k$ | **No** |
| **OAB** | $x_{k+1} = x_k + h\xi_{k+1}$ | **No** |
| **SES** | $x_{k+1} = x_k$ | **No** |
| **OBABO** | $x_{k+1} = x_k - \frac{h^2}{2} \nabla U(x_k) + h\xi_{k+1}$ | **Yes** |
| **BAOAB** | $x_{k+1} = x_k - \frac{h^2}{2} \nabla U(x_k) + \frac{h}{2}(\xi_k + \xi_{k+1})$ | **Yes** |

***But in MD we don't generally want to use very high friction...***

# with constraints

*B.L. and G. Patrick, J. Nonlin. Sci. 1996 (deterministic)*
***B.L. and C. Matthews, Proc Roy Soc A 2016 (stochastic)***

***An alternative to SHAKE discretization***

**Idea:** *preserve the configuration manifold during position moves and the cotangent space during impulse.*

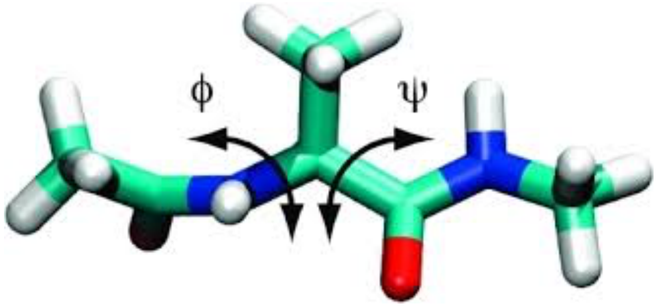
*The natural constrained analogue of* ***Verlet (BAB)*** *or* ***BAOAB***

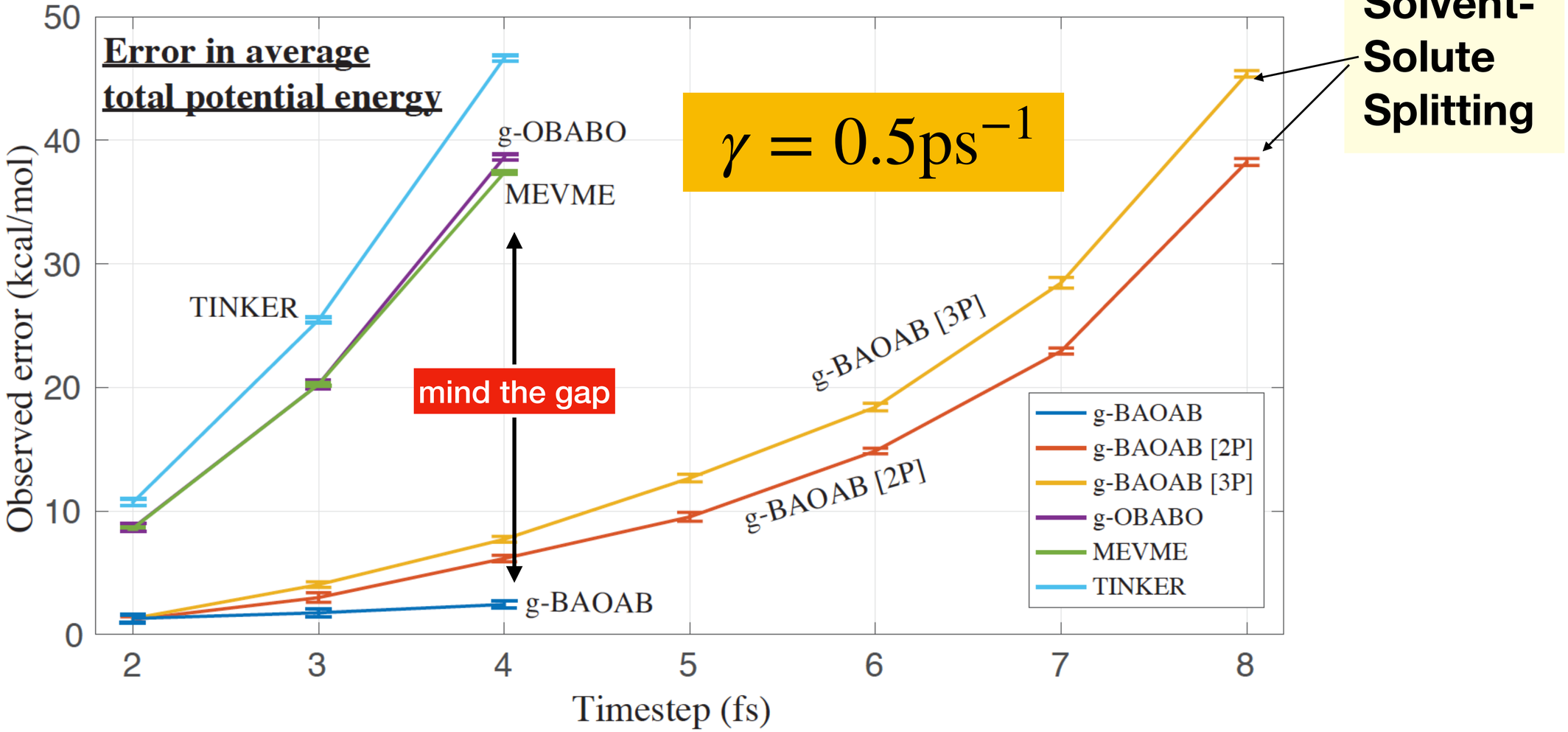
*g-Verlet:* $$\Phi_{h,H_{T^*\mathscr{M}}} \approx \Phi^{\mathcal{B}}_{h/2,U_{T^*\mathscr{M}}} \circ \Phi^{\mathcal{A}}_{h,T_{T^*\mathscr{M}}} \circ \Phi^{\mathcal{B}}_{h/2,U_{T^*\mathscr{M}}}$$

Combines: **geodesic flow** **projected "kicks"**

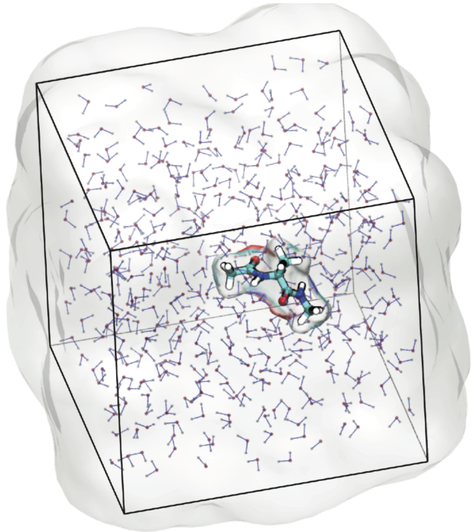
**O-step: projective Ornstein-Uhlenbeck solve**

# Solvated alanine dipeptide (constraints)



# Solute-Solvent Splitting



**multiple timestepping** based on splitting the interaction forces into **PP** (protein-protein), **PS** (protein-solvent), and **SS** (solvent-solvent)

**SS** dominates the computational cost. **PP, PS** determine the stepsize. Therefore, consider two-level multiple timestepping:

$$\exp\left(-\frac{h}{2}U_{\mathrm{SS}}(q)\right)\boxed{\exp\left(-\frac{h}{2}[T(p)+U_{\mathrm{PP}}(q)+U_{\mathrm{PS}}(q)]\right)}\exp\left(-\frac{h}{2}U_{\mathrm{SS}}(q)\right)$$

**m=2 or m=3 iterations of a geodesic Langevin integrator using several RATTLE substeps**

1. **PP+PS** is viewed as a many-body, stochastic system
2. Water motion can be implemented using **SETTLE**.

# Reconstructed free energy barrier height

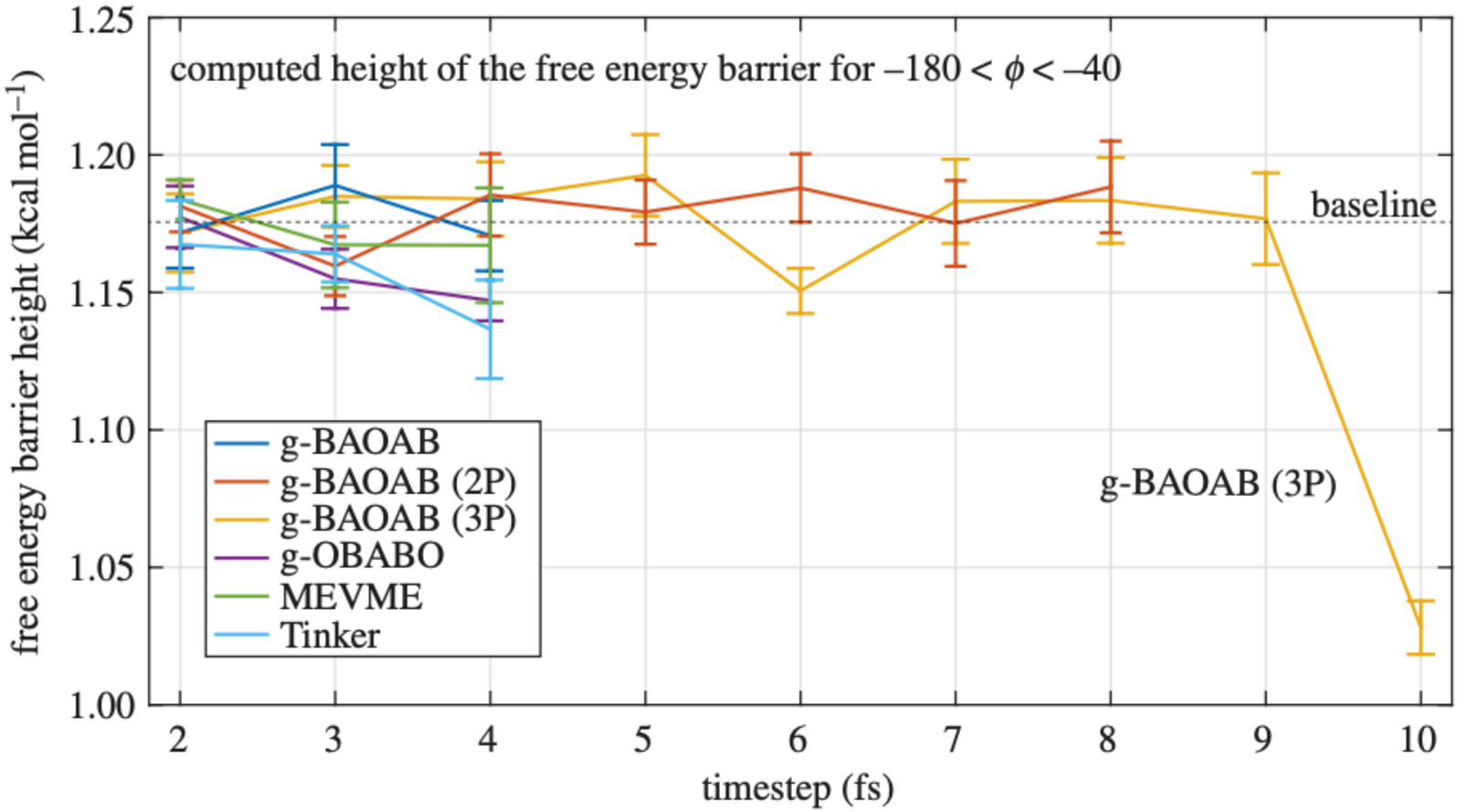


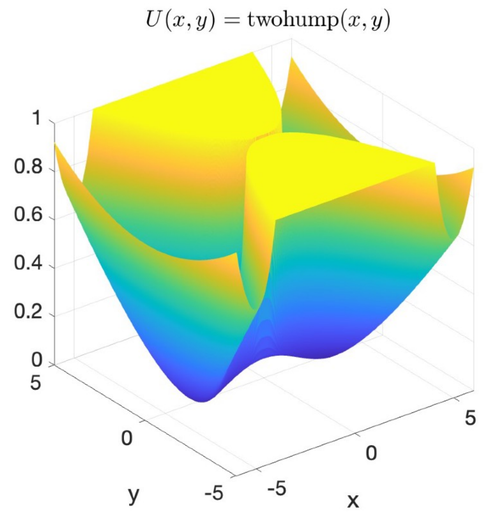

**Figure 7.** The estimated free energy barrier height in the $\phi$ coordinate is plotted as the stepsize is varied for each scheme. The g-BAOAB method is stable and accurate to stepsizes above 8 fs, giving a substantially correct effective free energy barrier height.

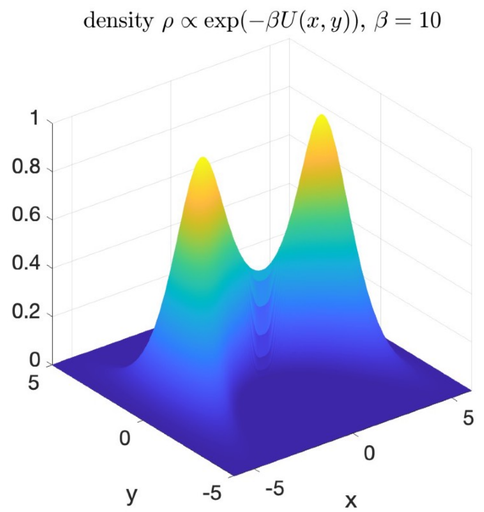
# Small friction regime

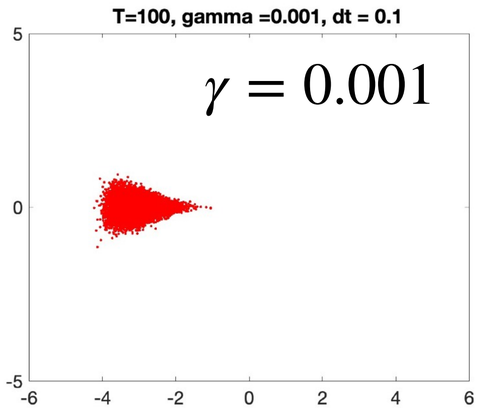
For rare events or dynamical MD calculations, we may wish to use small $\gamma$

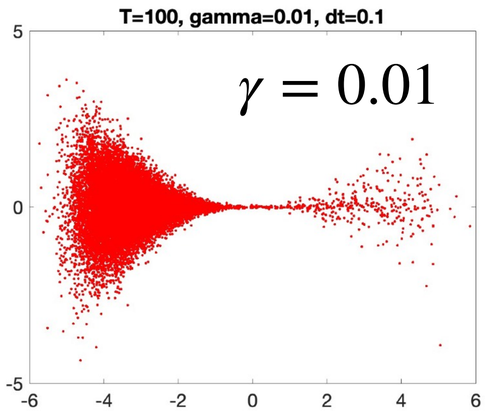
However, Langevin can be unreliable in the regime of small coupling.

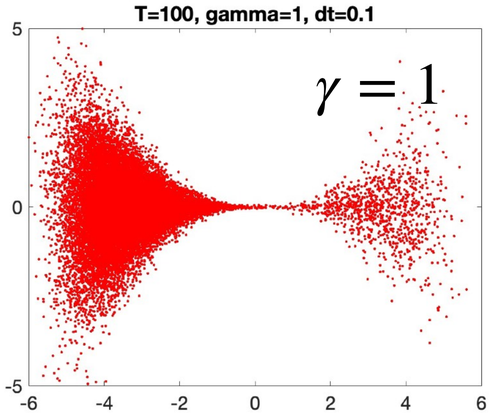
The convergence for numerical methods is poorly understood in this limit

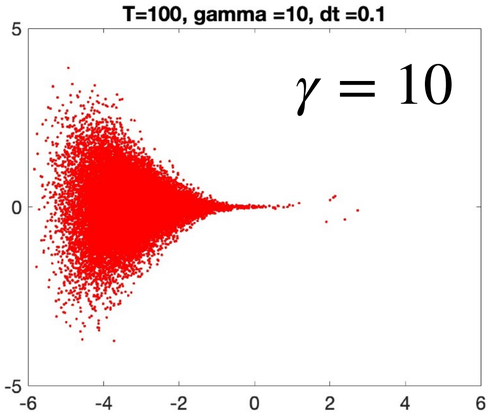


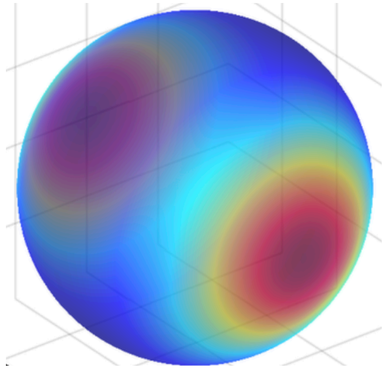



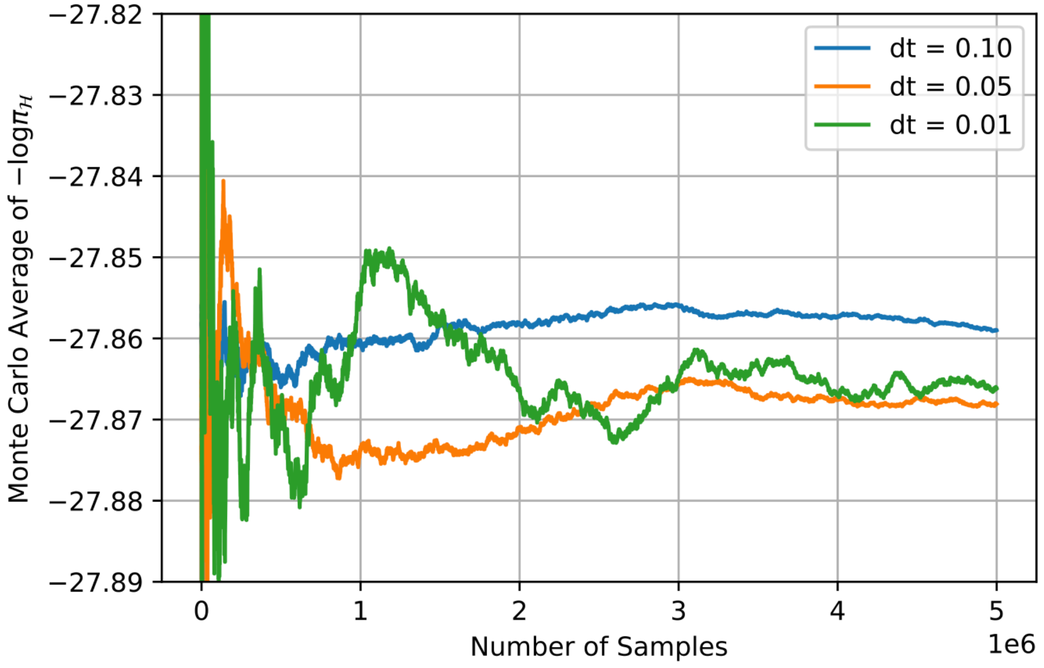

Firing deterministic trajectories might be better than delicately diffusing through the barrier...

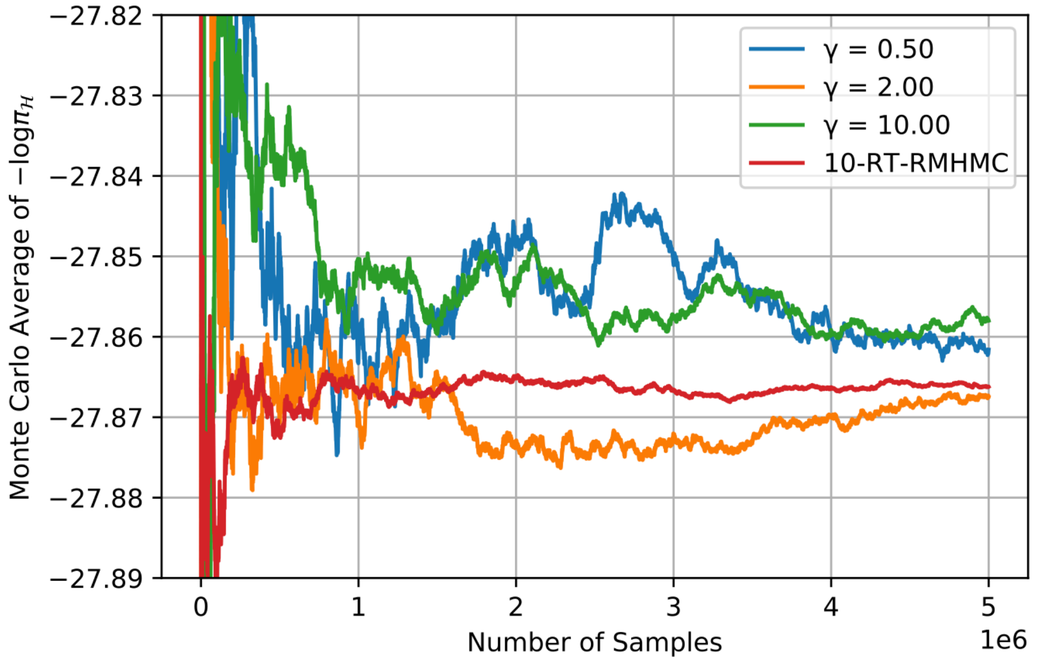
# Slow convergence of Langevin g-BAOAB for a multimodal distribution

Bingham-von Mises-Fisher distribution

$$\pi_{\mathcal{H}}(x) \propto \exp\{c^T x + x^T A x\} \qquad x \in S^4 \subset \mathbb{R}^5$$





# Using HMC to explore complex distributions

Hamiltonian Monte Carlo offers an alternative to Langevin dynamics. Deterministic paths generated from a Hamiltonian may be able to traverse narrow entropic bottlenecks.

Typically HMC is based on a combination of symplectic integration (volume preserving, reversible), stochastic refreshment, and Metropolis correction. Convergence to the invariant distribution follows under similar conditions to Langevin dynamics (analysis as Harris chains).

Ergodicity relies on a **mixing property** in the Hamiltonian system (or its discretization),

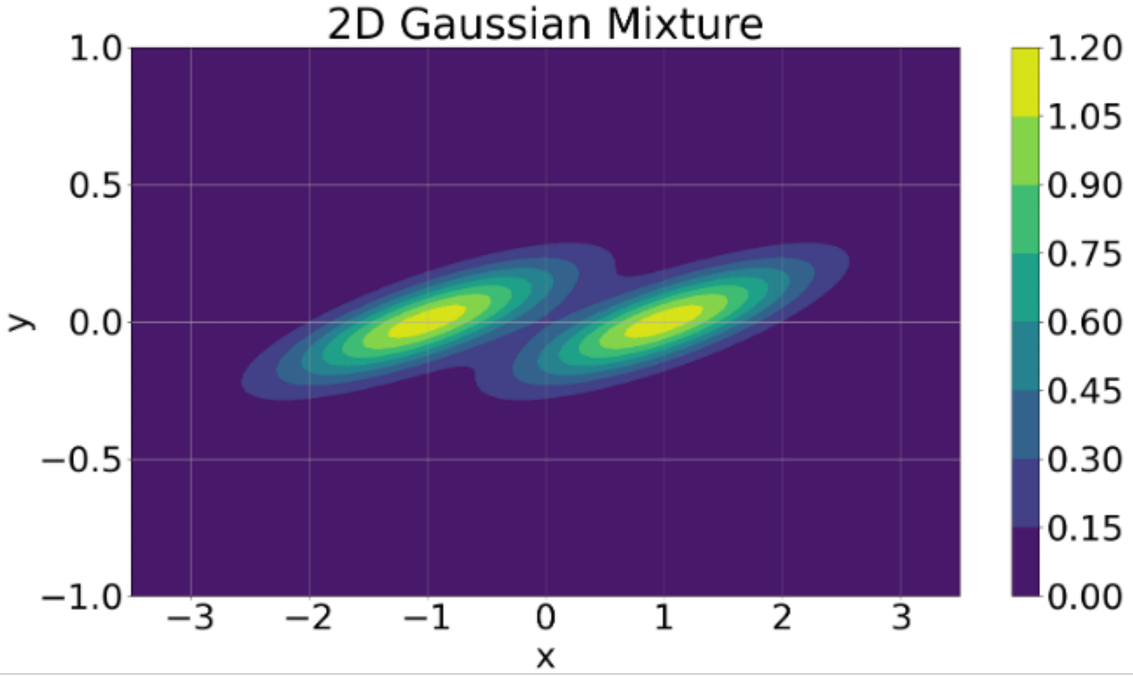
Metropolization removes bias due to numerical error but is also the principal drawback of HMC.

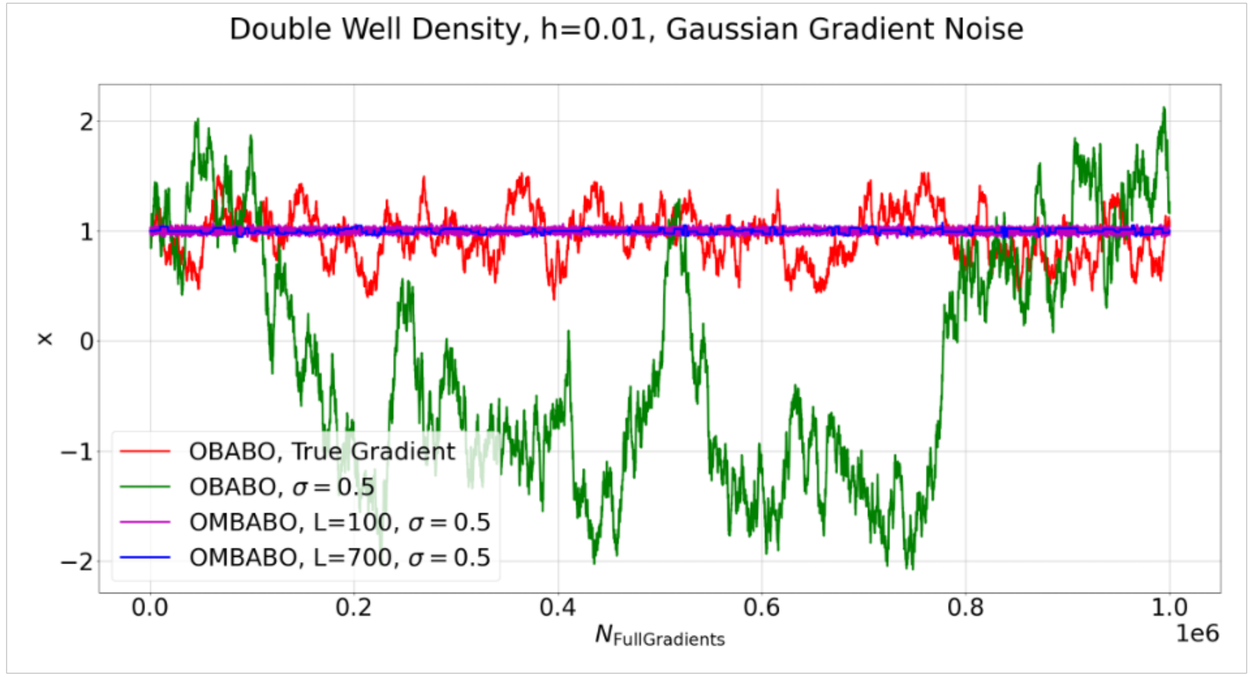
# Metropolization

→**First, it costs (wasted samples), sometimes a lot.**

→**Rejections can impede transitions, at least in thermally driven systems**

## Current Work with R. Lohmann (PhD student)



# Using HMC for rare event sampling [Two Ideas]

RHM*

**1. Randomize the time variable**

**N. Bou-Rabee** and **J.M. Sanz-Serna**, Randomized Hamiltonian Monte Carlo, 2017.

Randomizing the timestep size is a common technique in HMC. Alternatively one can consider the limiting case of exact Hamiltonian dynamics with an exponentially distributed duration parameter. Rigorously proved (for the unconstrained case).

**2. Just don't Metropolize**

Motivation comes from the ad-hoc "stochastic gradient HMC" used in the machine learning community. (While the subsampled gradient is cheap, the full energy needed for Metropolis correction is costly.) If our aim is not bias correction but exploration, can we get away with dropping it?

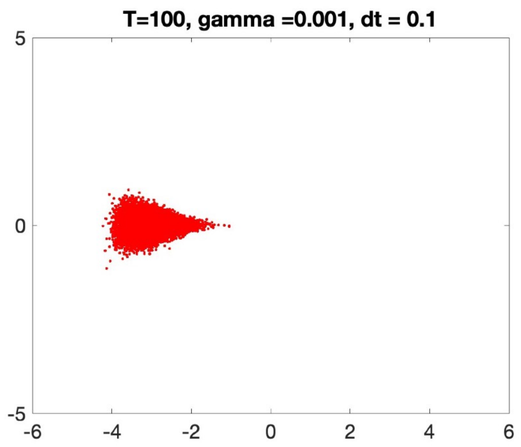
Origin: **Hybrid molecular dynamics** [**Andersen** 1980, **Duane** 1985]

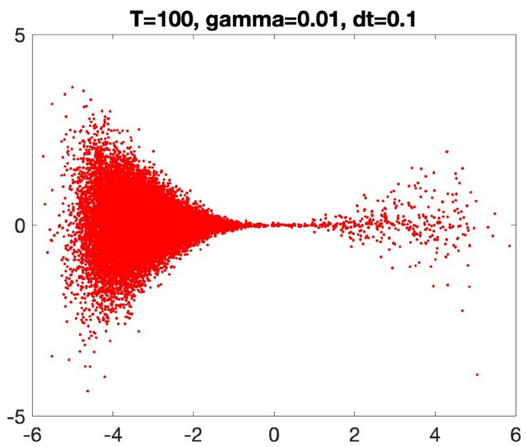
Some theory [**Manghoubi and Smith** 2017], but under the strong convexity conditions...

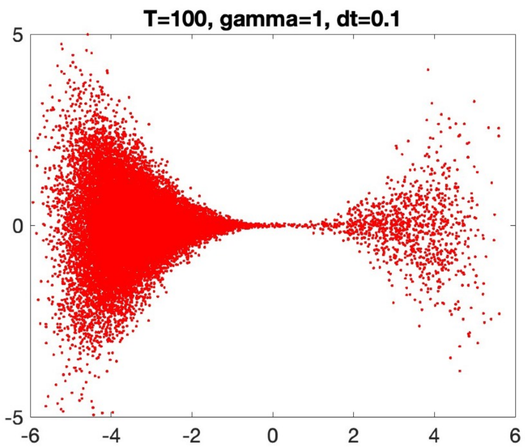
***There might be other better ways to unbias simulations!***

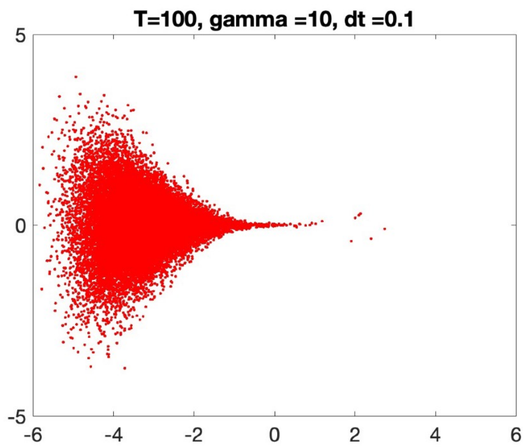
# parameter independence of RHM* (without correction and with randomized duration)

LD

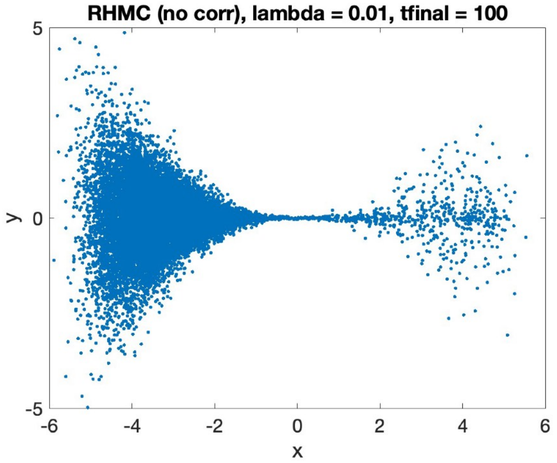


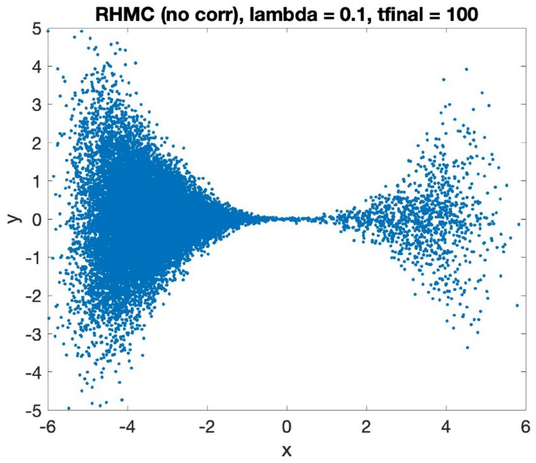



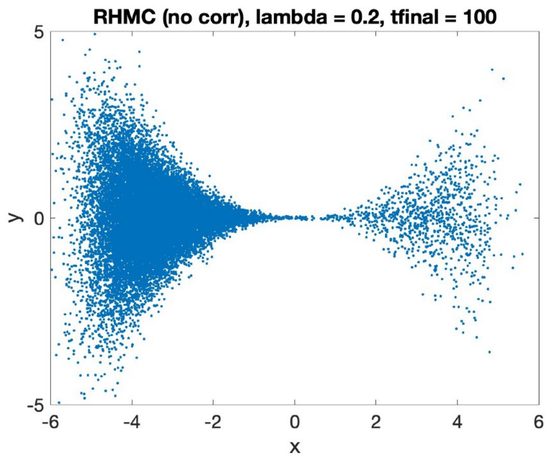



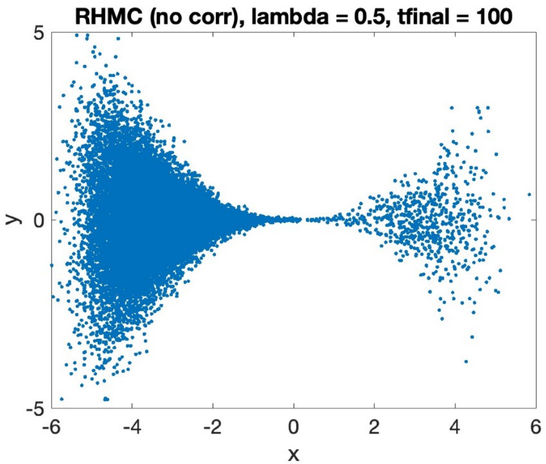




RHM*



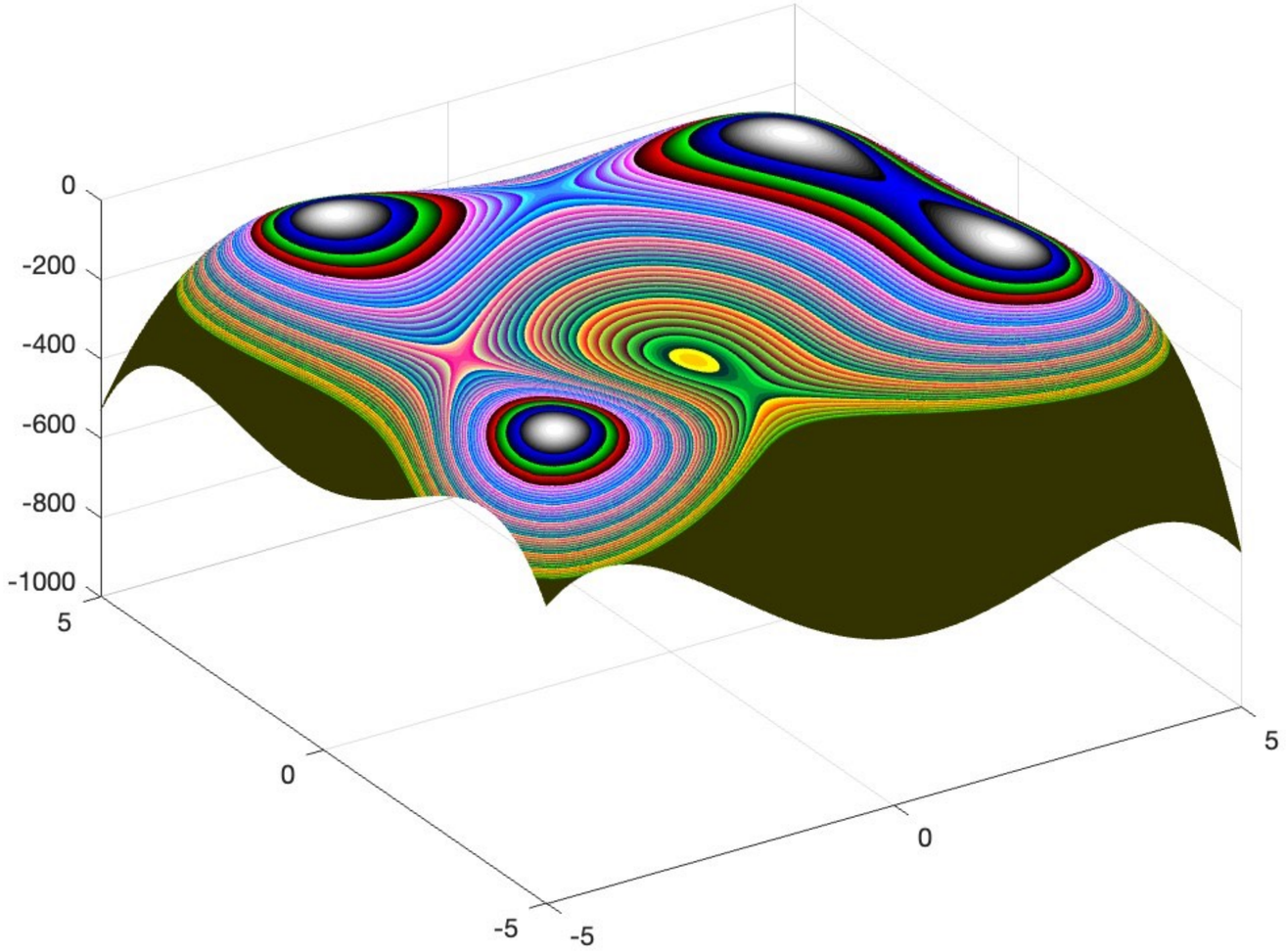



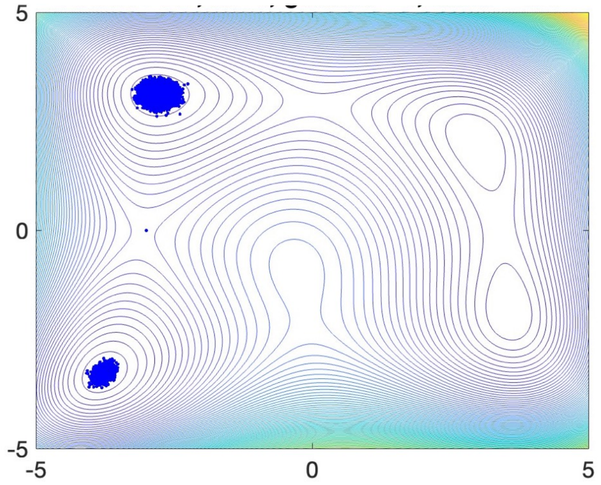

lambda =0.01 lambda =0.1 lambda =0.2 lambda =0.5

# Himmelblau

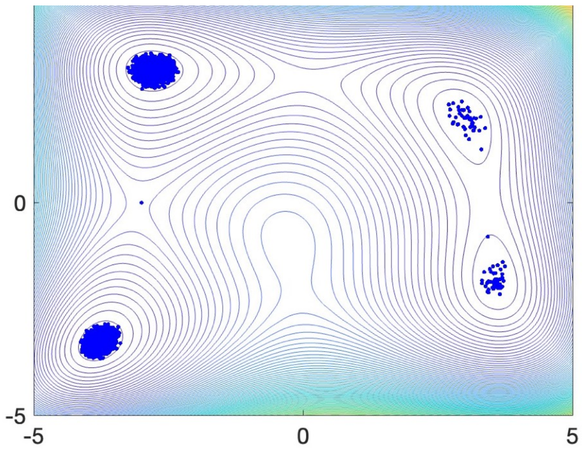


BAOAB is essentially unbiased but makes slow transitions to/from the isolated state

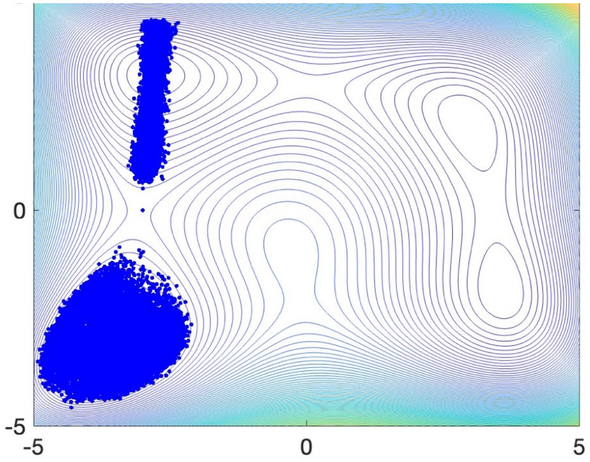
gamma = 1, tfinal = 400, dt=0.1



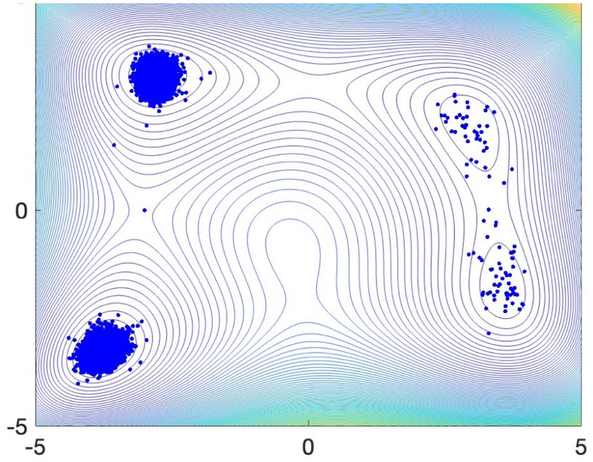

**gamma = 0.01**, **tfinal = 4000**, dt = 0.1

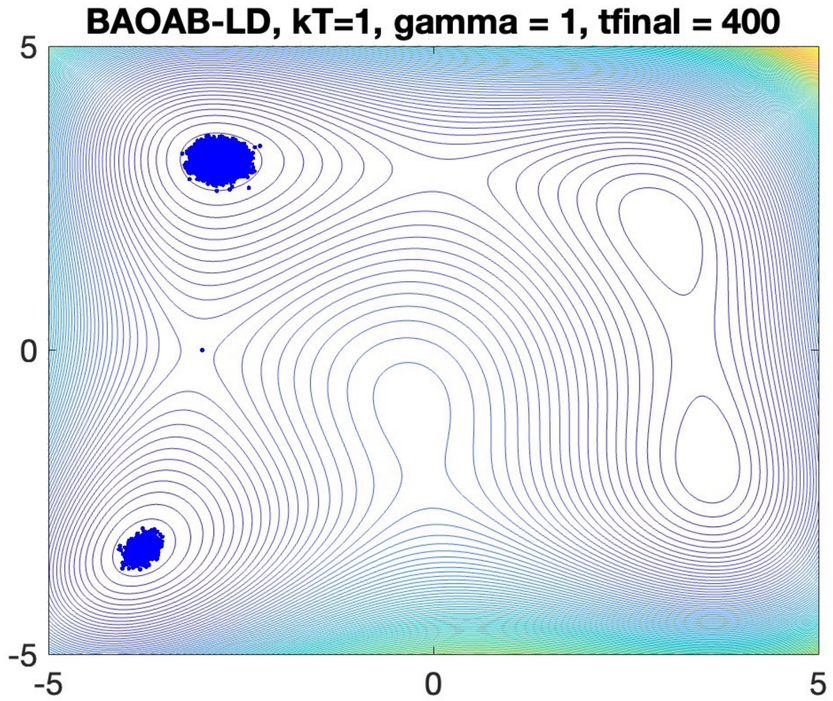


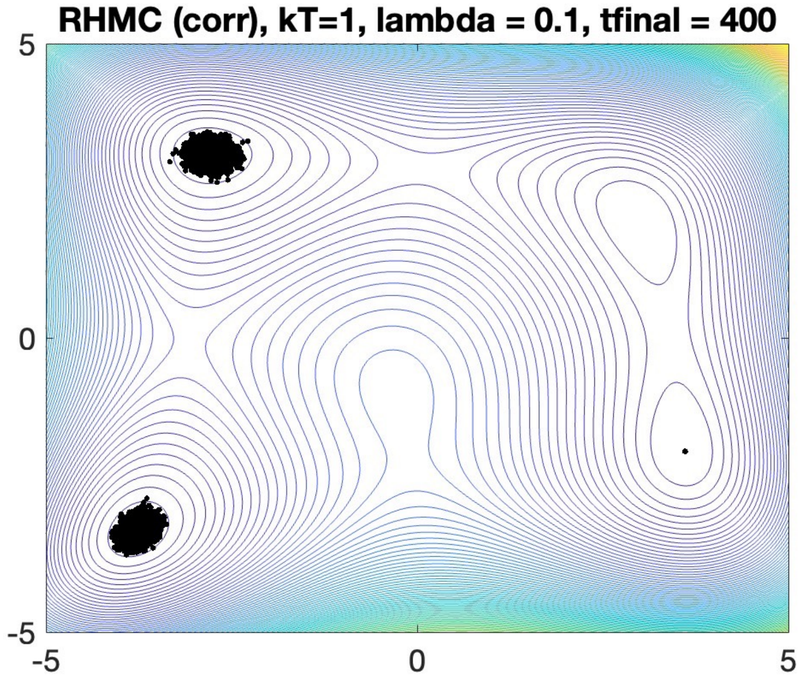

**gamma = .001**, tfinal = 400, **dt =0.05**

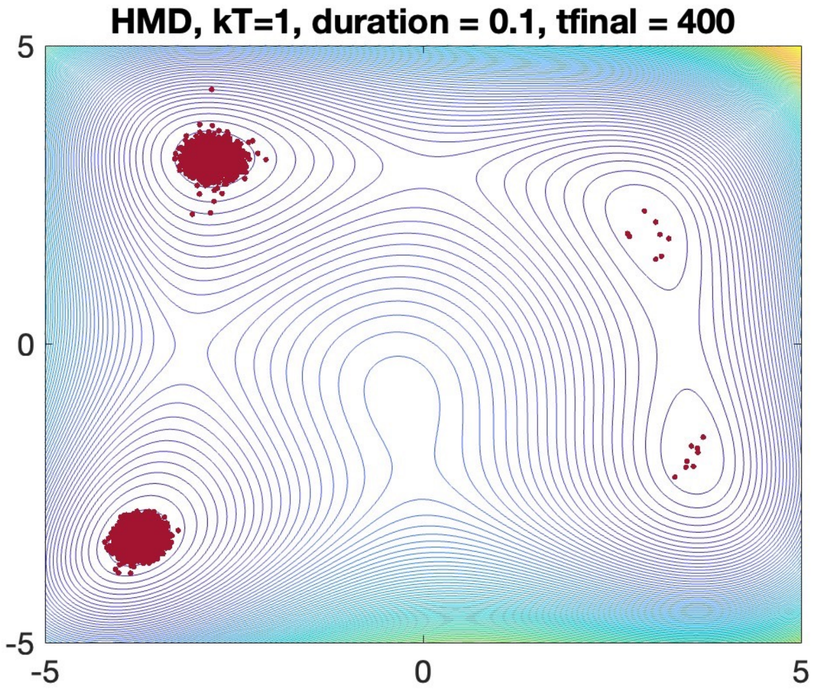


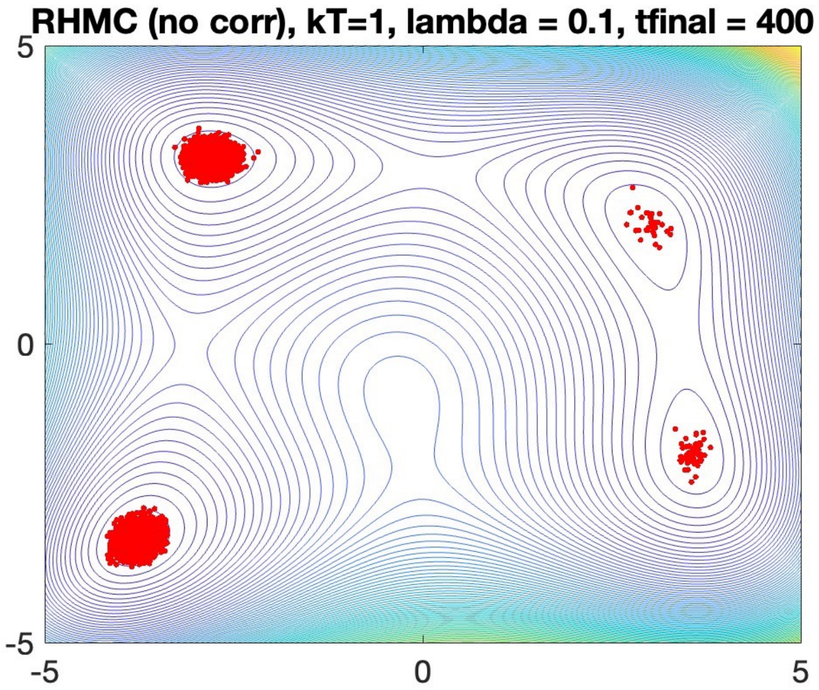

**gamma = 0.01**, tfinal = 400, dt = 0.1

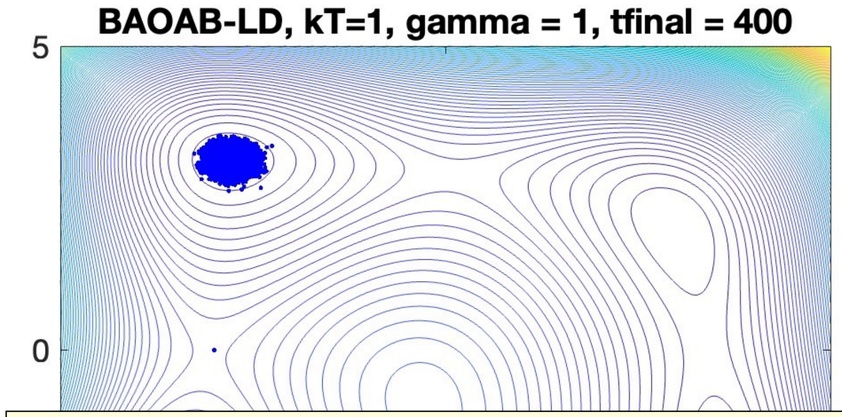


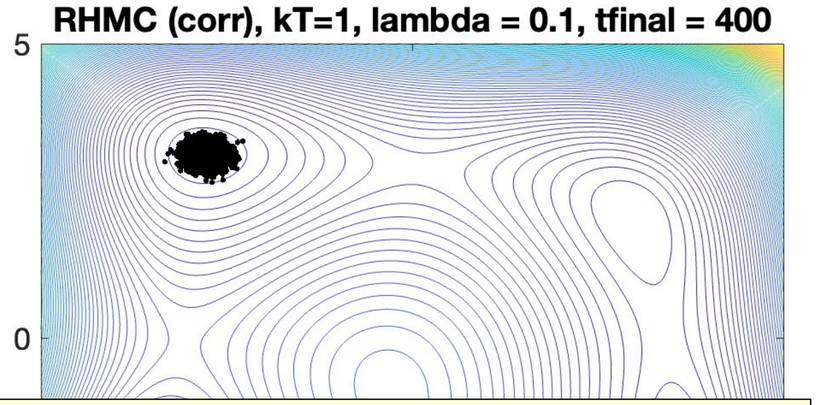

BAOAB-LD, kT=1, gamma = 1, tfinal = 400
5
0
-5
-5
0
5


RHMC (corr), kT=1, lambda = 0.1, tfinal = 400
5
0
-5
-5
0
5


HMD, kT=1, duration = 0.1, tfinal = 400
5
0
-5
-5
0
5

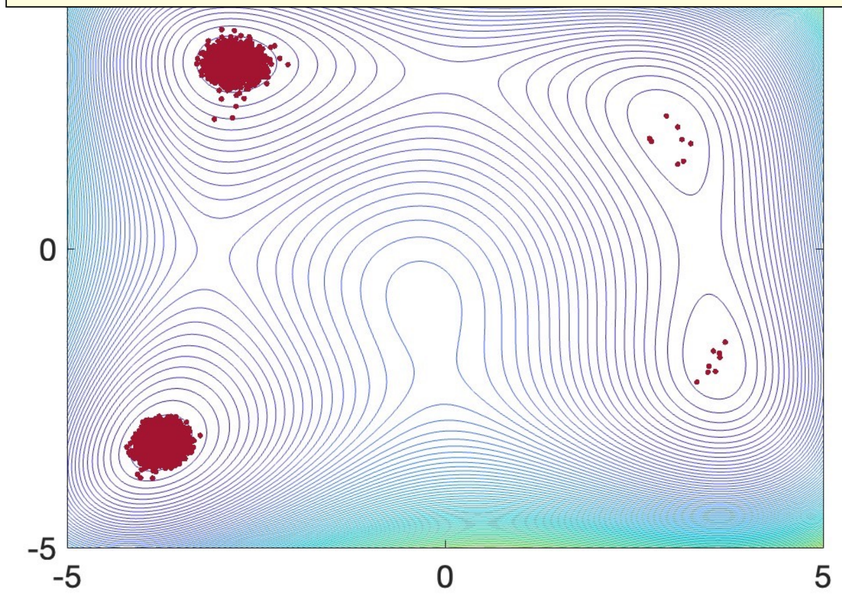

RHMC (no corr), kT=1, lambda = 0.1, tfinal = 400
5
0
-5
-5
0
5

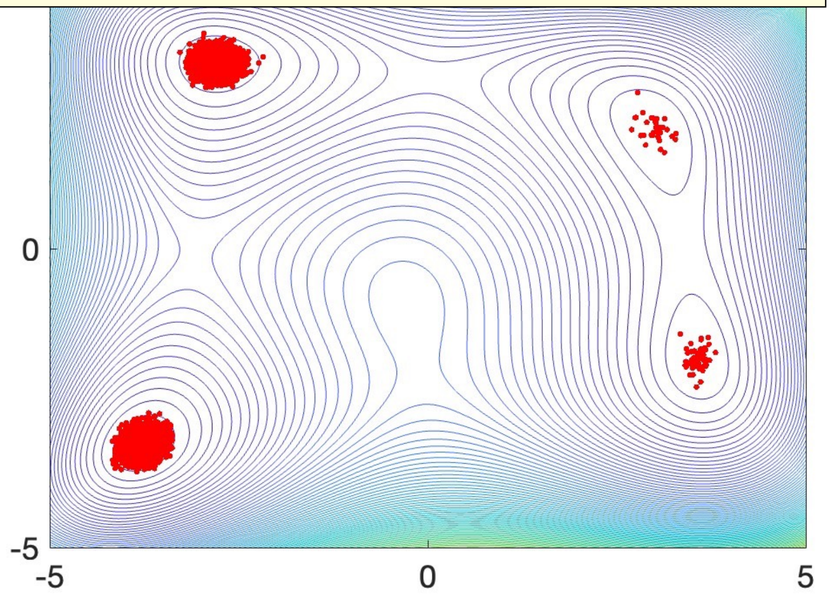



HMD (no correction, no time randomization) gains some exploration
RHMC without correction (i.e. RHMD) gains still more exploration

The uncorrected schemes lose accuracy in basins compared to Langevin but LD has poor exploration.  Correction inhibits convergence



# Randomized Time Riemannian Manifold Hybrid Monte Carlo

**Whalley, Paulin, Leimkuhler 2023**,Randomized Time Riemannian Manifold Hamiltonian Monte Carlo, ArXiv https://arxiv.org/abs/2206.04554

We work with the explicit constraint geometry, *not the Euclidean formulation. This makes the method efficient (semi-explicit).*

Uses deterministic paths which are (formally) geodesics of the manifold (but, in practice, are generated by RATTLE or else g-BAB)

Difficult to use ideas of hypocoercivity in the manifold setting. We assume compact manifold but don't show geometric ergodicity. Proof based on minorization (Doeblin condition).

Costs are **not worse than RATTLE** type constrained MD.

**Metropolization can be omitted, at least in the small stepsize limit**

# Invariance and Ergodicity

Assume: smooth compact manifold $M = \{x \mid g(x) = 0\}$

where $g : \mathbb{R}^d \to \mathbb{R}^m$ has full rank Jacobian

smooth potential $U(x) = -\log \pi_*(x)$

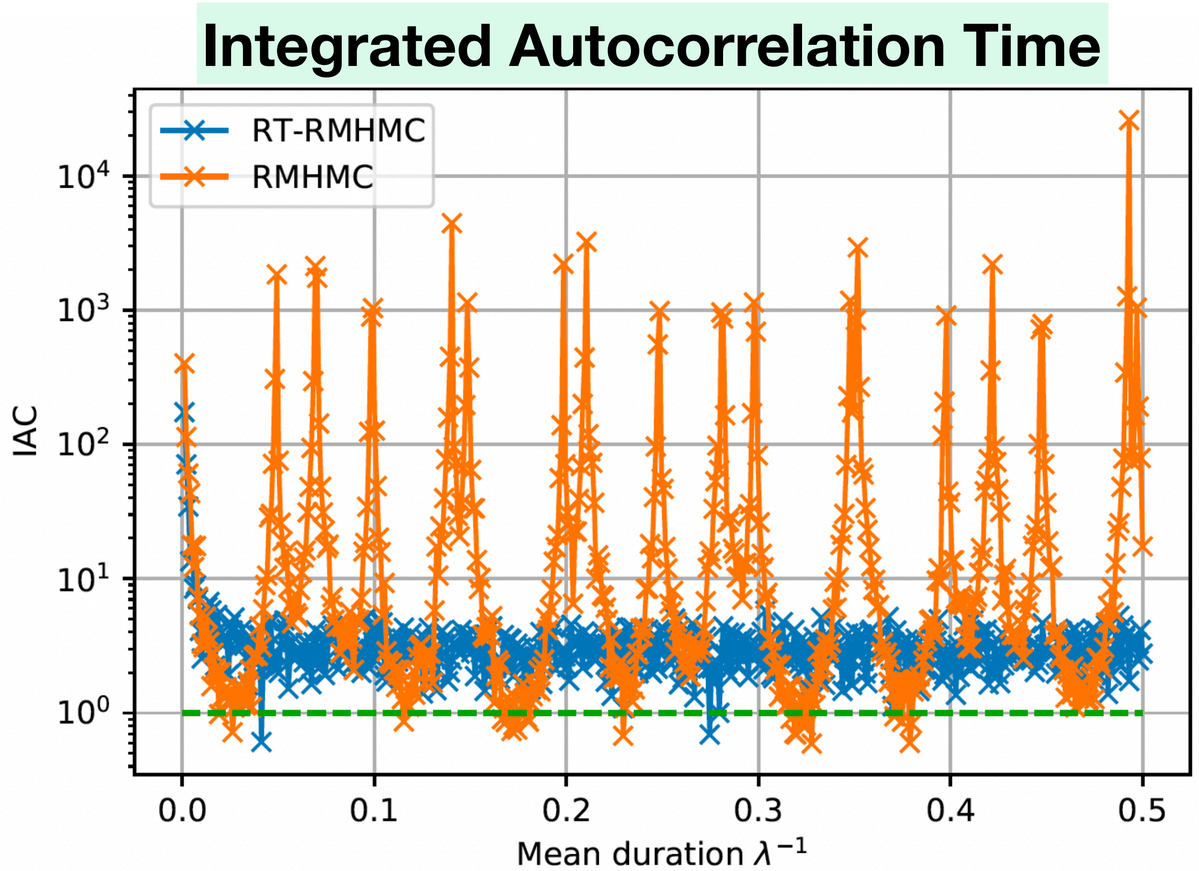
→ **RT-RMHMC** leaves invariant the Gibbs distribution restricted to $\mathrm{T}M$

→ **RT-RMHMC** is ergodic in the sense that transition density aligns to the target probability density over time

$$\lim_{t \to \infty} \int_M |K^t(x, y) - \pi_*(y)| \sigma_M(\mathrm{d}y) = 0$$

# Comparison - Bingham-von Mises-Fisher distribution in 3D

$$\pi_*(x) = \exp(c^t x + x^t A x), \qquad \|x\| = 1$$

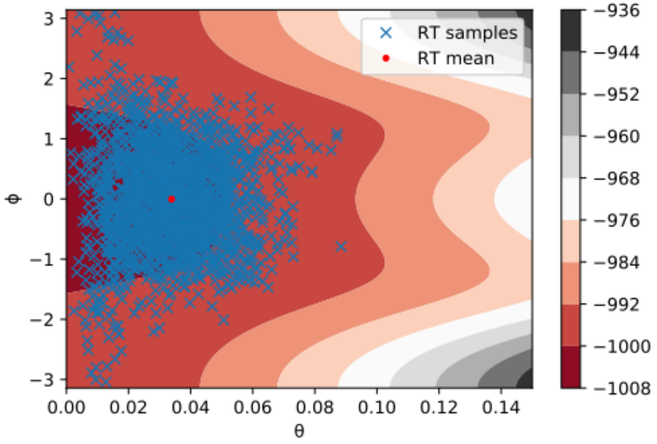



**The RT-RMHMC method stabilizes the convergence rate as a function of the mean duration compared to RMHMC (with fixed duration).**

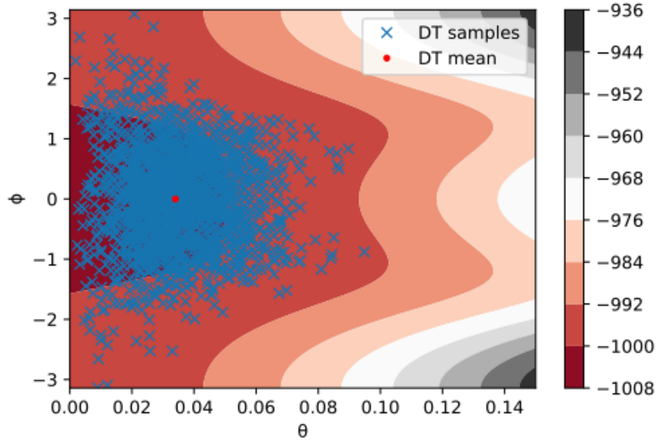
# BvMF distributions

**RT-RMHMC** **RMHMC**

$\lambda = 0.09$

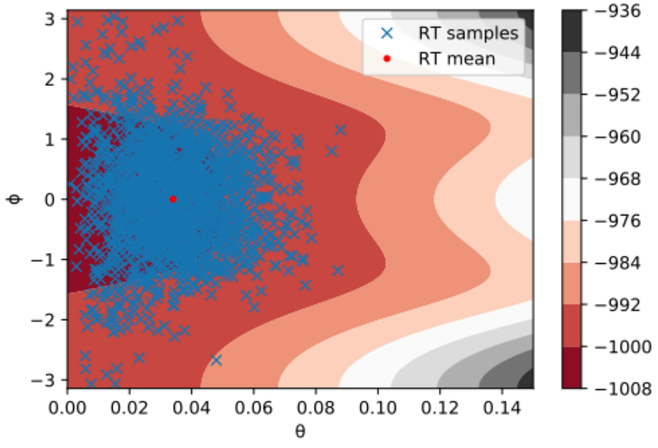



(a)

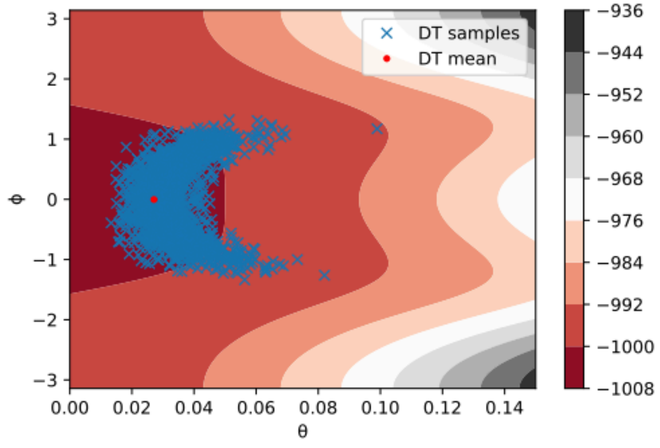



(b)

$\lambda = 0.1$




(c)




(d)

# Sparse Reconstruction for Astronomical Data

Estimate the covariance matrix when the number of data points is small relative to the dimension

The sample covariance is a poor estimator due to rank deficiency

To regularize the problem, we work in the space of matrices which are **"low rank + sparse"**

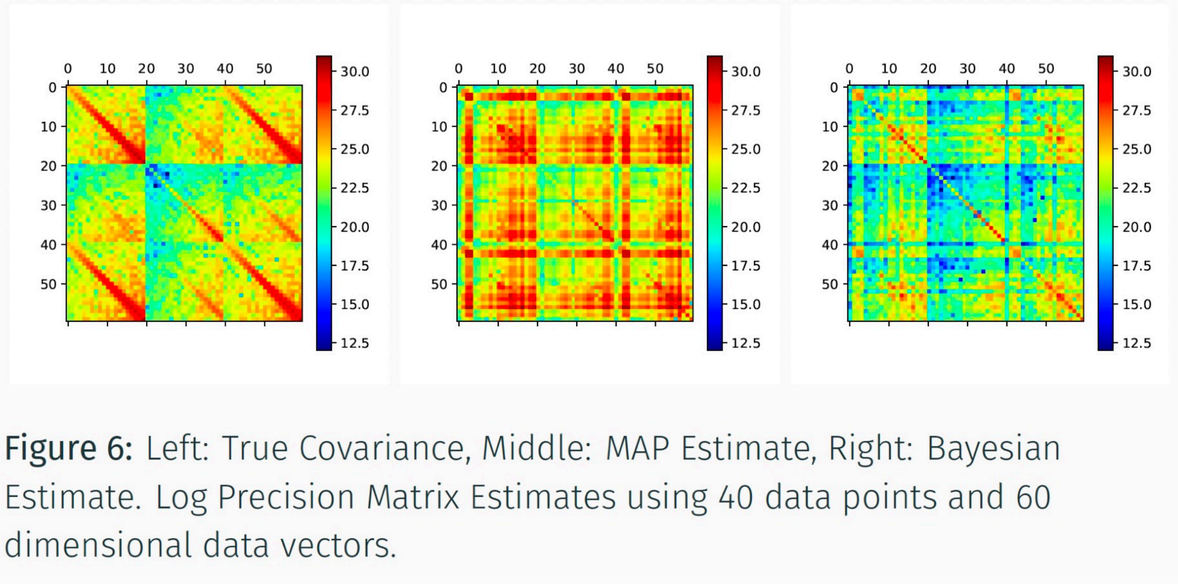
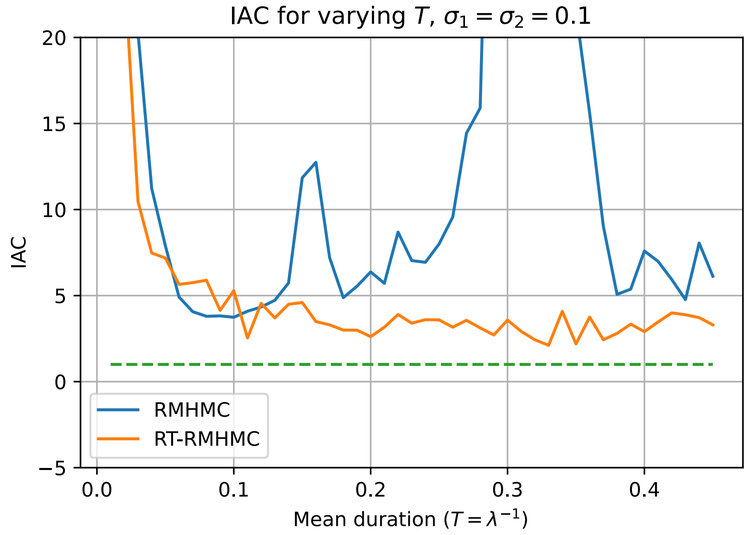


**Figure 6:** Left: True Covariance, Middle: MAP Estimate, Right: Bayesian Estimate. Log Precision Matrix Estimates using 40 data points and 60 dimensional data vectors.



# Summary

Langevin integrators constructed for a wide variety of SDEs provide efficient sampling algorithms for diverse applications

Accurate schemes are also possible for constrained systems, leading to particularly powerful methods in combination with other devices like multiple timestepping.

Numerical studies suggest that underdamped SDE discretization methods can have **low bias** in practice, particularly if the **friction is sufficiently high.**

However, the high friction limit can **slow convergence**, particularly for multimodal distributions.

Used without correction **RT-RMHMC** offers a powerful alternative that can improve convergence to equilibrium.