Repelling-Attracting Hamiltonian Monte Carlo

Siddharth Vishwanath
Joint work with Hyungsuk Tak

UC San Diego

Motivation

Sampling

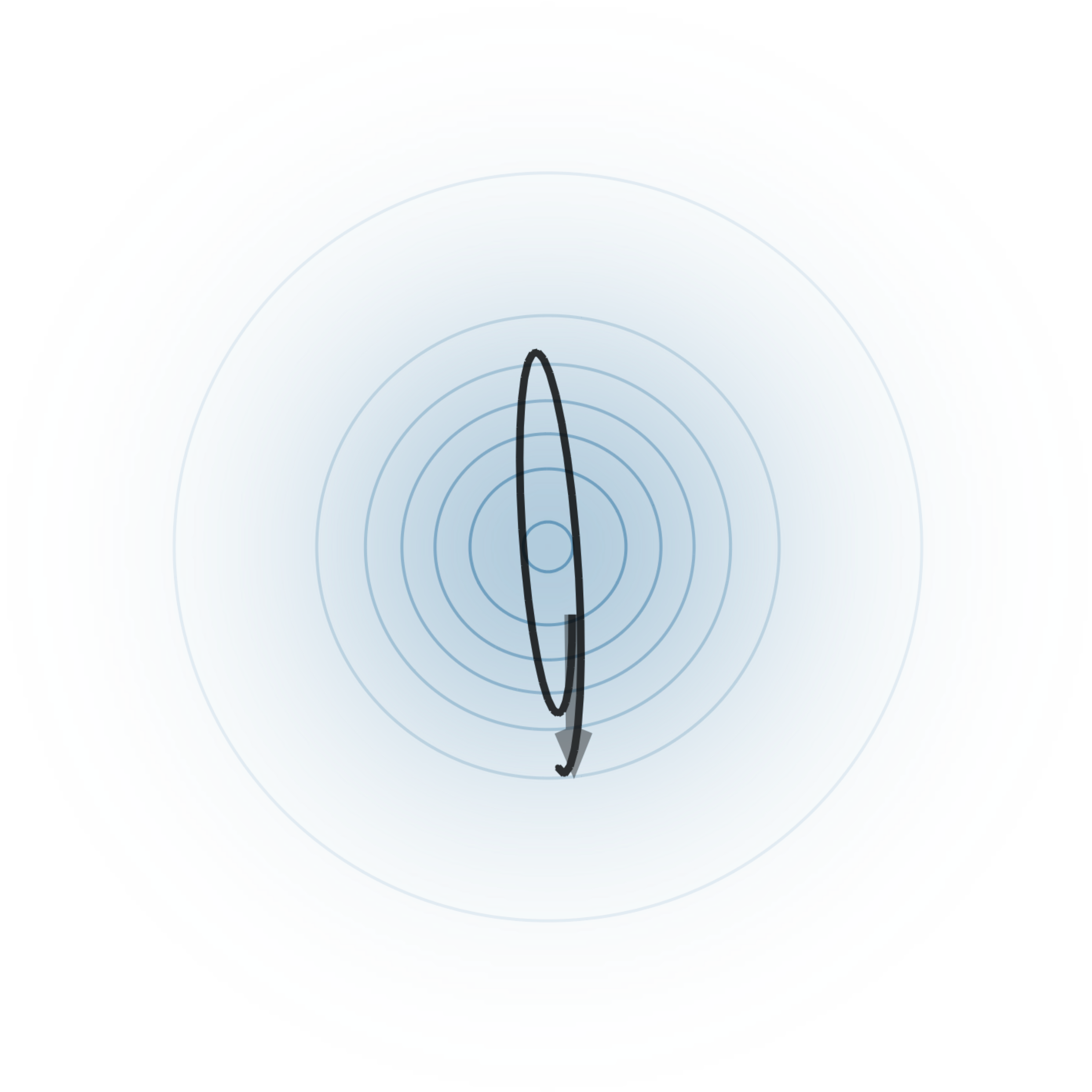
Generate samples $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n\} \sim \pi$ from a target distribution

$$\pi(\boldsymbol{q}) \propto e^{-U(\boldsymbol{q})}$$

Methods

1. Rejection sampling ❌
2. Random-walk Metropolis ✅
3. Metropolis adjusted Langevin algorithm 😮
4. Hamiltonian Monte Carlo 😲
5. Stein Variational gradient descent 🤯
6. Wasserstein gradient flows 🤯
7. Normalizing flows 🤯
8. ...

Repelling-Attracting Hamiltonian Monte Carlo



Simulation options	
Algorithm	RA-HMC
Target distribution	standard
Autoplay	✓
Autoplay delay	1200
Tweening delay	20
Step	
Reset	
Algorithm Options	
Leapfrog Steps	37
Leapfrog Δt	0.2
γ	0.1

Close Controls

Courtesy the template by Chi Feng

Objective

Can we design a sampler which:

1. Can efficiently sample from multimodal distributions
2. Preserves all the nice properties of existing mainstays

The Problem with Diffusions

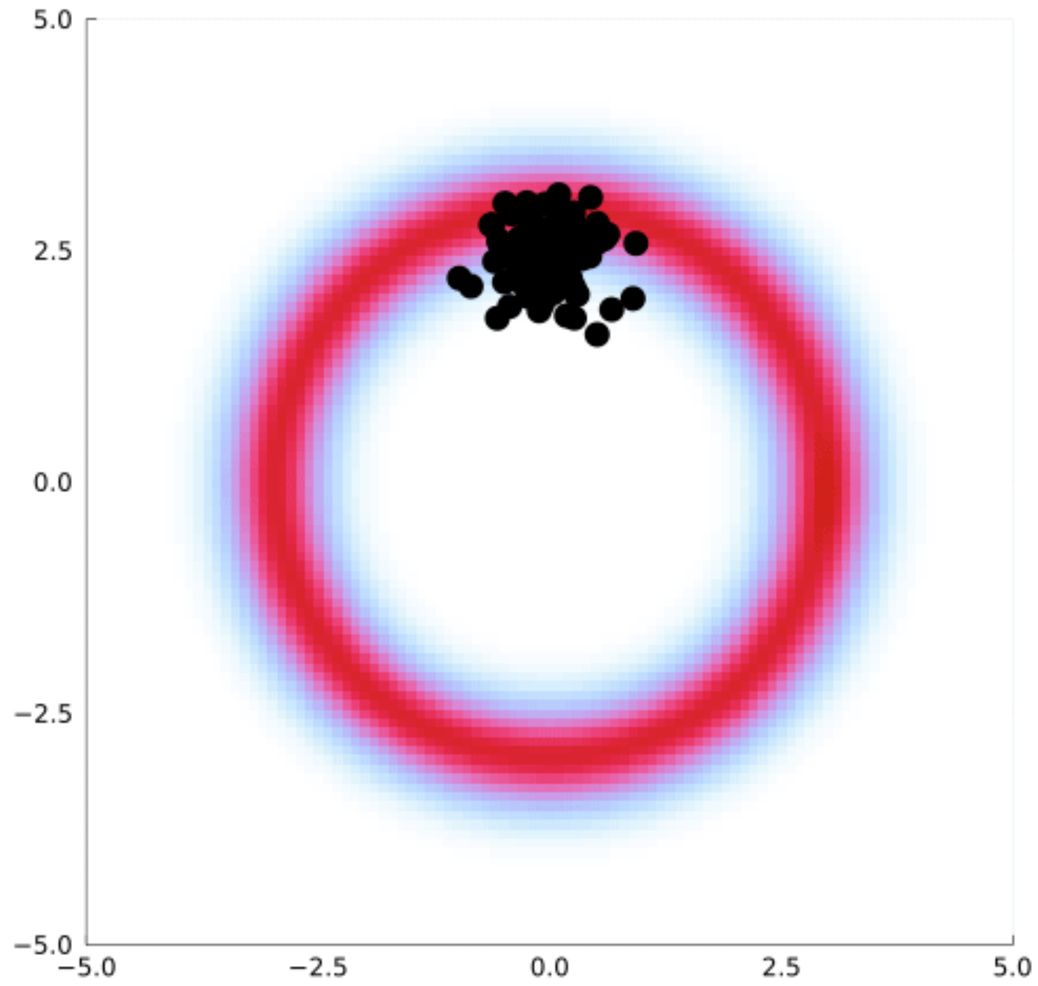
Random Walk Metropolis

1. Given the current state $\boldsymbol{q}_n$, propose a new state

$$\boldsymbol{q}_{n+1} \sim \kappa(\boldsymbol{q}|\boldsymbol{q}_n)$$

2. Accept / Reject with probability

$$\alpha(\boldsymbol{q}_{n+1}|\boldsymbol{q}_n) = 1 \wedge \frac{\kappa(\boldsymbol{q}_n|\boldsymbol{q}_{n+1})\pi(\boldsymbol{q}_{n+1})}{\kappa(\boldsymbol{q}_{n+1}|\boldsymbol{q}_n)\pi(\boldsymbol{q}_n)}$$



Wastes too much time in high dimensions.

More Intelligent Diffusions

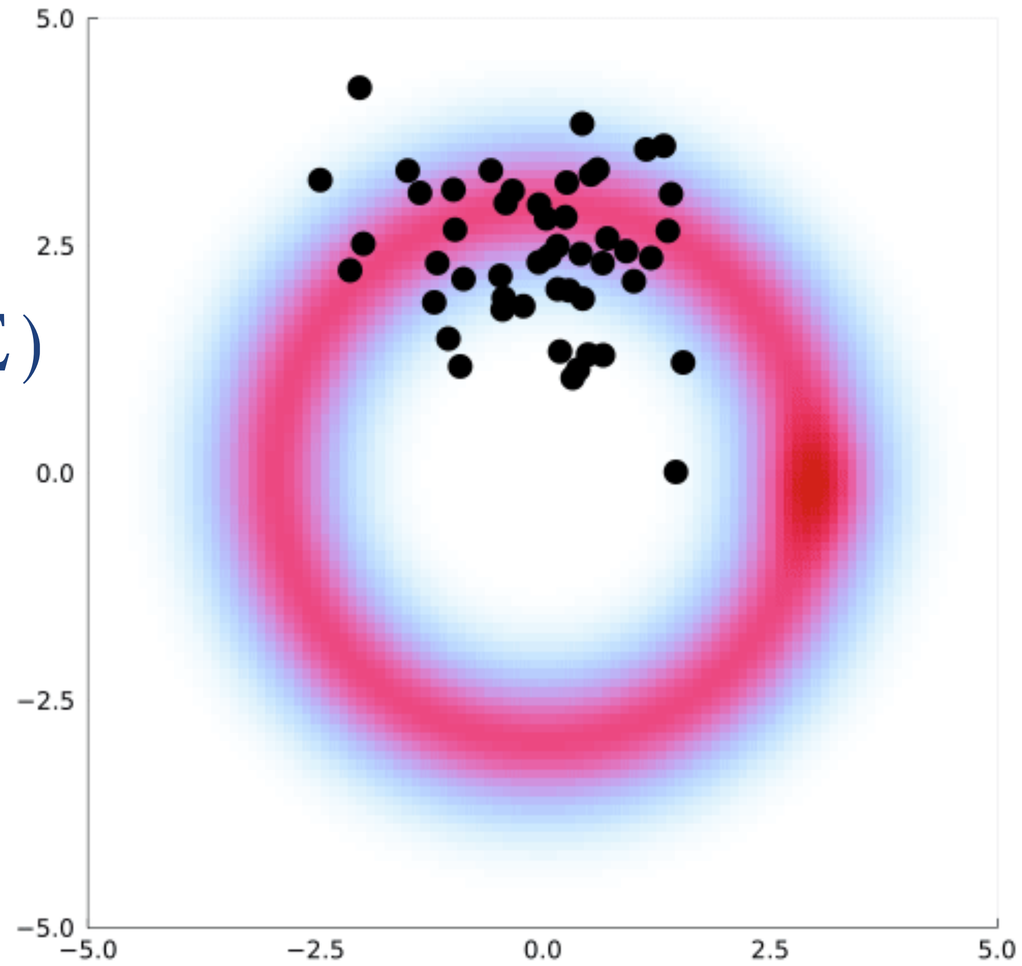
Langevin Monte Carlo

1. Given the current state $\boldsymbol{q}_n$, propose a new state

$$\boldsymbol{q}_{n+1} \sim \boldsymbol{q}_n + h \cdot \nabla \log \pi(\boldsymbol{q}_n) + 2h \cdot \mathrm{N}(\boldsymbol{0}, \Sigma)$$

2. Accept / Reject with probability

$$\alpha(\boldsymbol{q}_{n+1}|\boldsymbol{q}_0) = 1 \wedge \frac{\kappa(\boldsymbol{q}_n|\boldsymbol{q}_{n+1})\pi(\boldsymbol{q}_{n+1})}{\kappa(\boldsymbol{q}_{n+1}|\boldsymbol{q}_n)\pi(\boldsymbol{q}_n)}$$



Only works for small step sizes!

Hamiltonian Monte Carlo

1. Augment state space with auxiliary variables $\boldsymbol{p} \sim \mathrm{N}(\mathbf{0}, \Sigma)$
2. Joint distribution $(\boldsymbol{q}, \boldsymbol{p}) \sim \exp(-H(\boldsymbol{q}, \boldsymbol{p}))$ where $H(\boldsymbol{q}, \boldsymbol{p})$ is given by

$$H(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{p}^{\top}\Sigma^{-1}\boldsymbol{p}$$

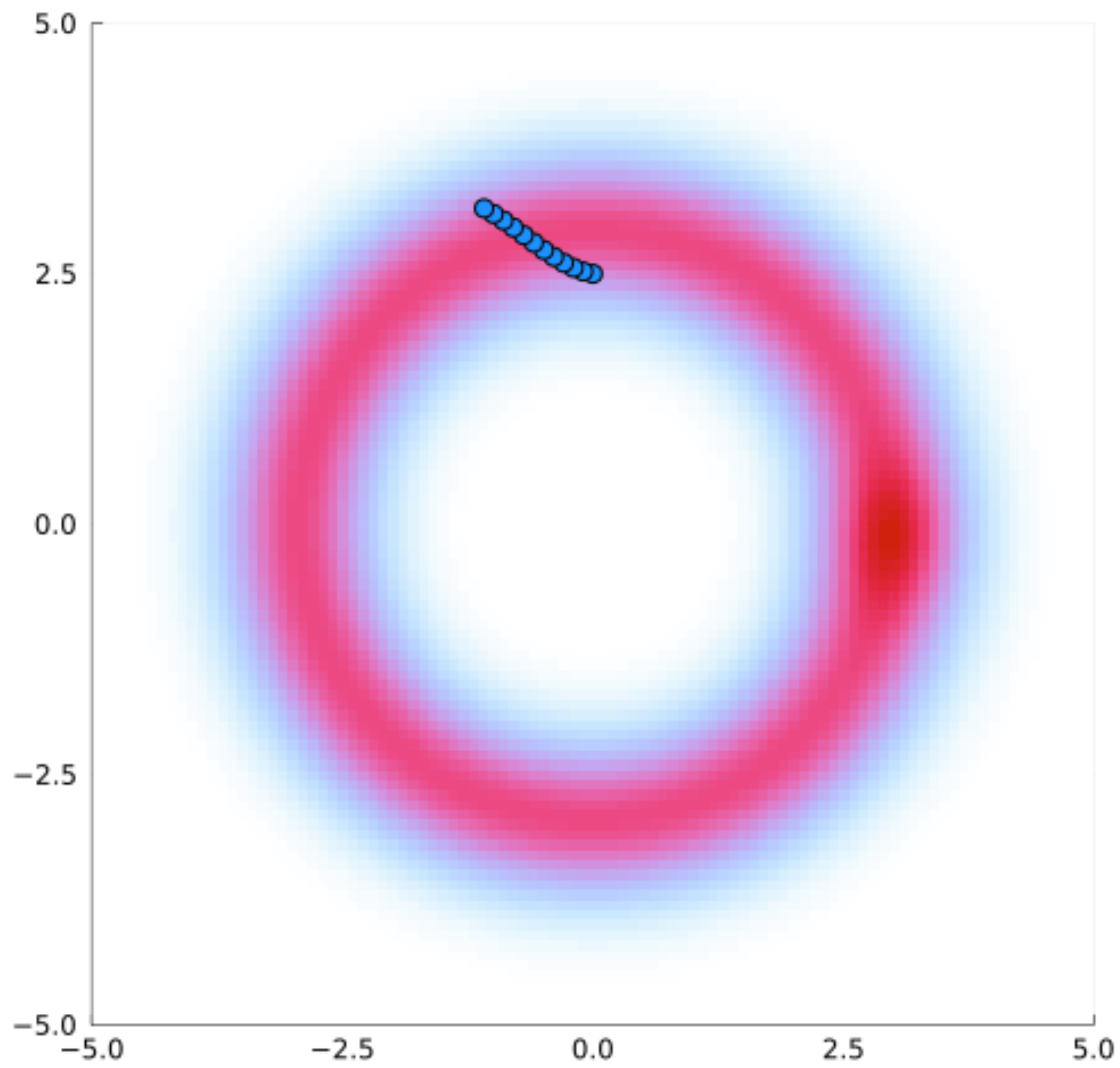
3. Treat $H(\boldsymbol{q}, \boldsymbol{p})$ as the Hamiltonian of a system and generate trajectories using Hamiltonian dynamics

$$\frac{d}{dt}\begin{bmatrix} \boldsymbol{q}_t \\ \boldsymbol{p}_t \end{bmatrix} = \begin{bmatrix} \mathbf{O} & \mathbb{I} \\ -\mathbb{I} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \nabla_{\boldsymbol{q}} H(\boldsymbol{q}_t, \boldsymbol{p}_t) \\ \nabla_{\boldsymbol{p}} H(\boldsymbol{q}_t, \boldsymbol{p}_t) \end{bmatrix} = \begin{bmatrix} \Sigma^{-1}\boldsymbol{p}_t \\ -\nabla U(\boldsymbol{q}_t) \end{bmatrix}.$$

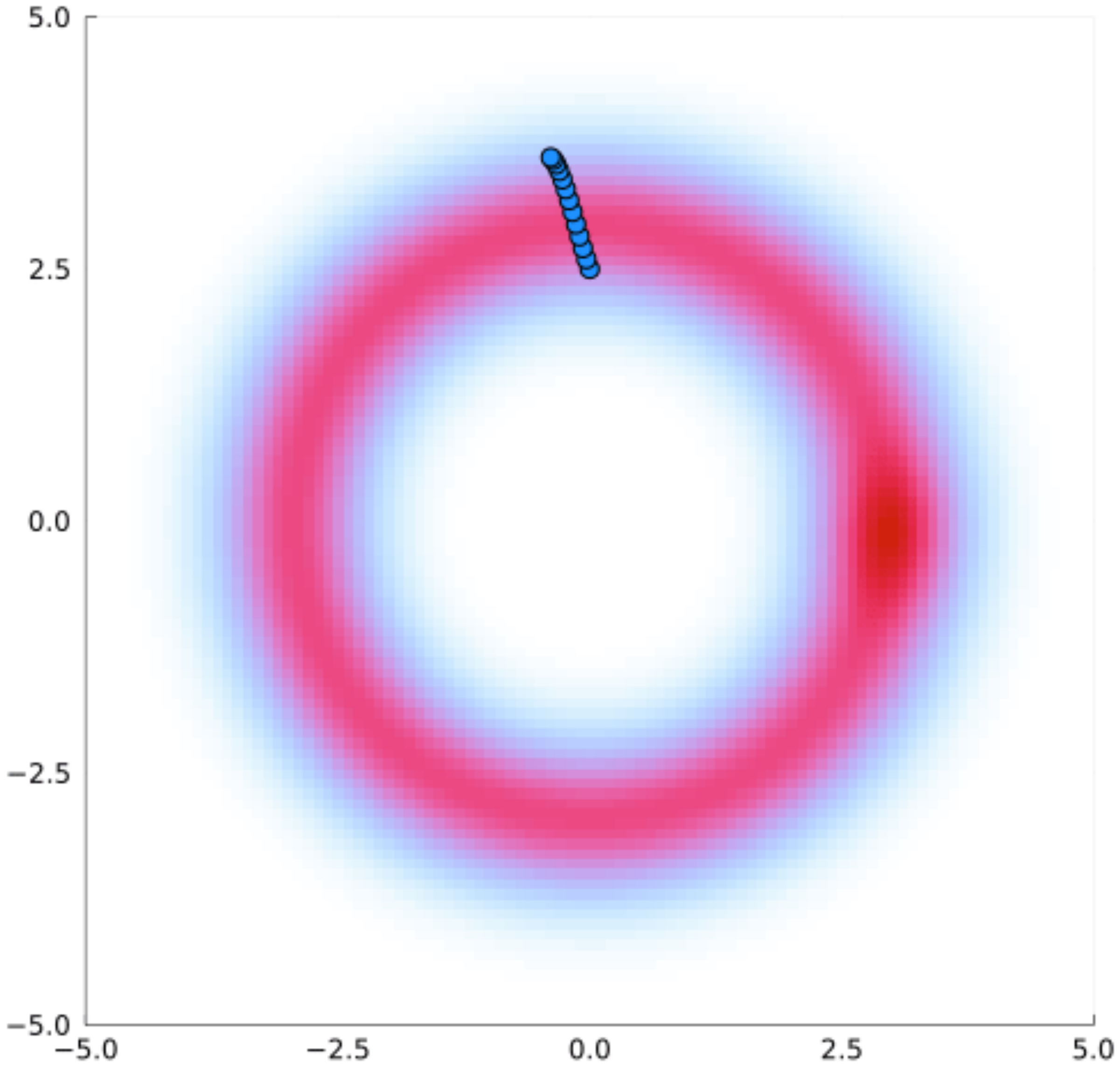
4. Accept/reject state $(\boldsymbol{q}_t, \boldsymbol{p}_t)$ with probability

$$\alpha(\boldsymbol{q}_t, \boldsymbol{p}_t | \boldsymbol{q}_0, \boldsymbol{p}_0) = \min\left\{1, \frac{\kappa(\boldsymbol{q}_0, \boldsymbol{p}_0 | \boldsymbol{q}_t, \boldsymbol{p}_t) \cdot e^{-H(\boldsymbol{q}_t, \boldsymbol{p}_t)}}{\kappa(\boldsymbol{q}_t, \boldsymbol{p}_t | \boldsymbol{q}_0, \boldsymbol{p}_0) \cdot e^{-H(\boldsymbol{q}_0, \boldsymbol{p}_0)}} \middle| \cdot \frac{\partial(\boldsymbol{q}_t, \boldsymbol{p}_t)}{\partial(\boldsymbol{q}_0, \boldsymbol{p}_0)} \middle| \right\}.$$

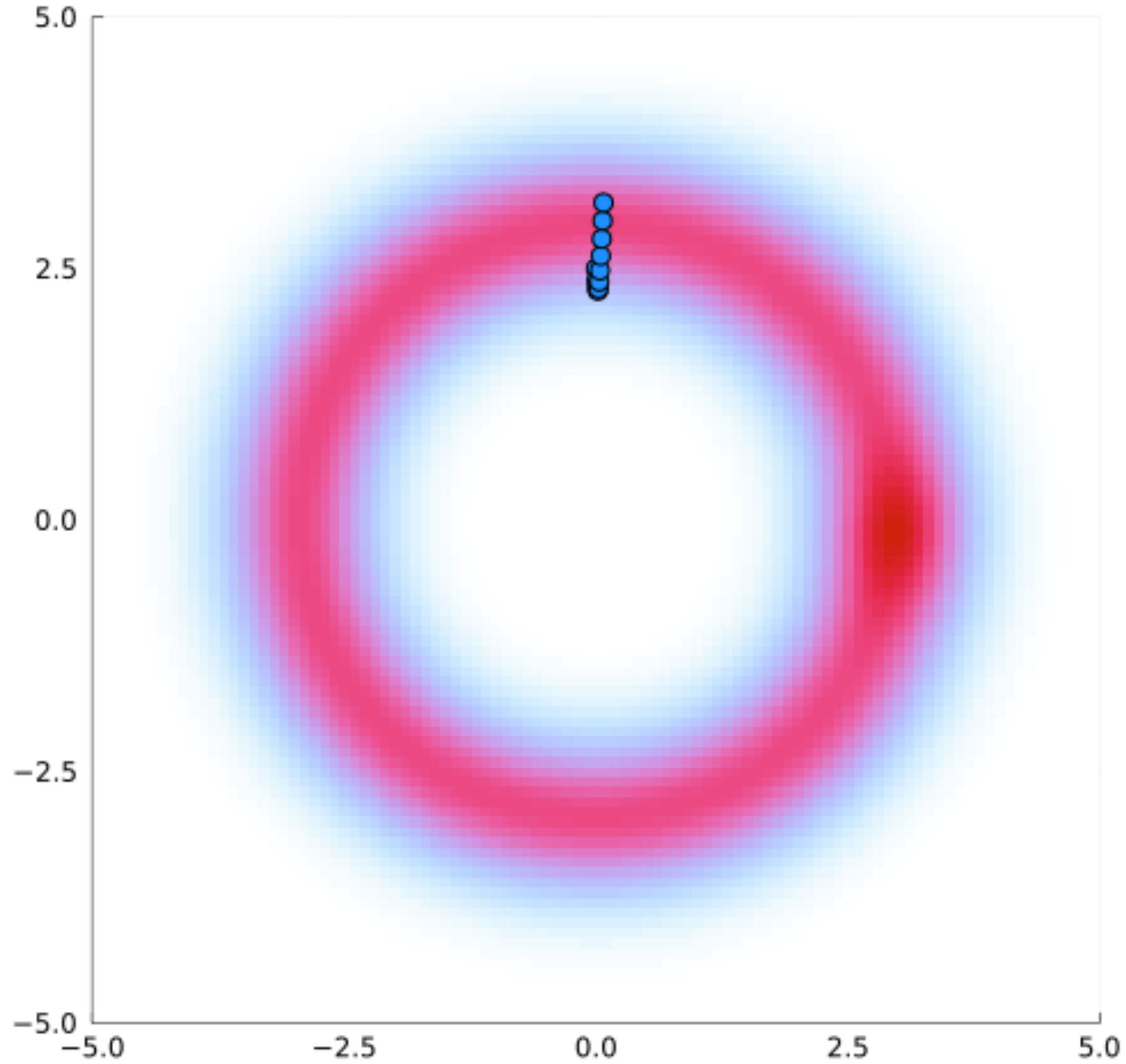
Hamiltonian Monte Carlo



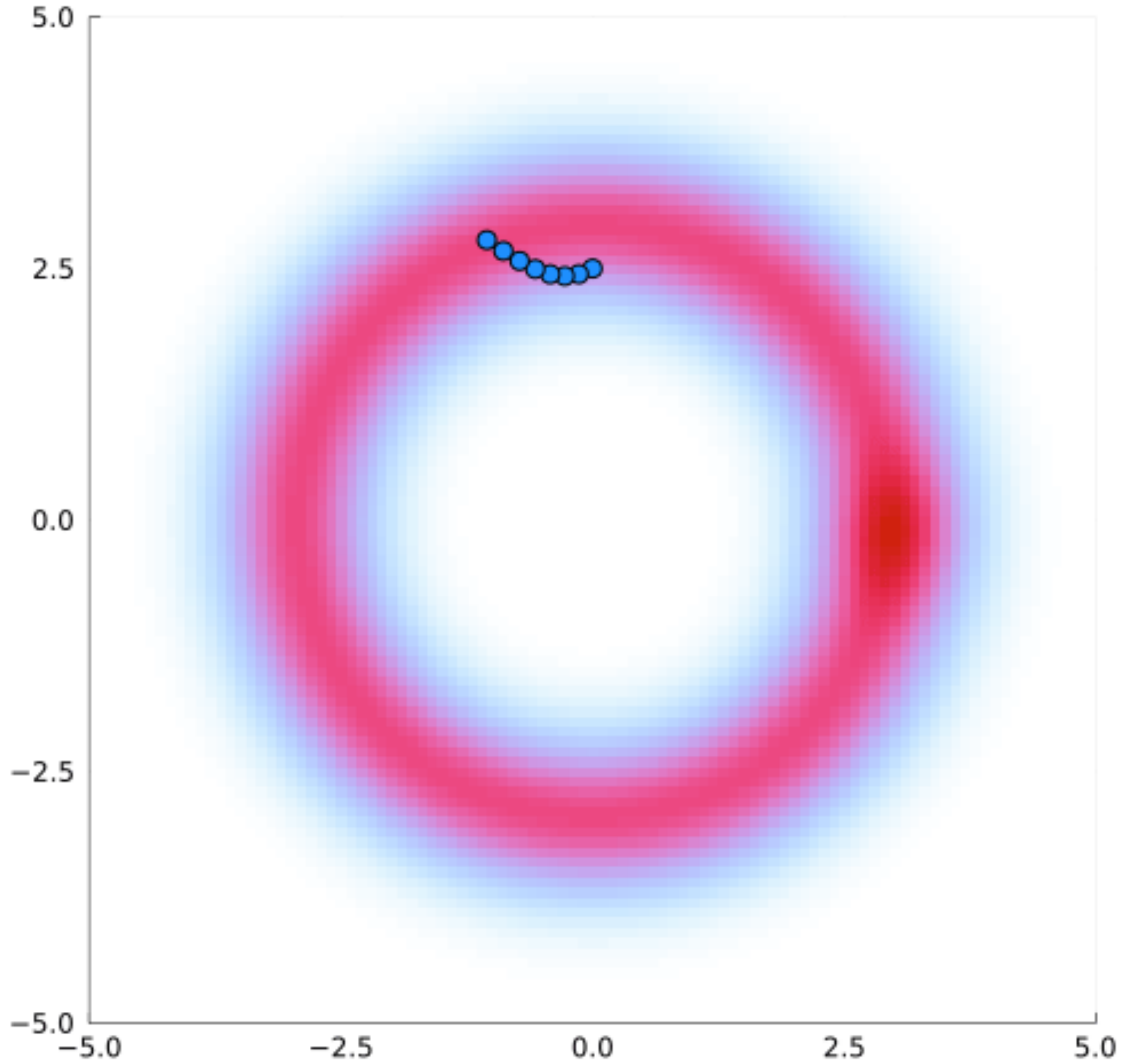
Hamiltonian Monte Carlo



Hamiltonian Monte Carlo



Hamiltonian Monte Carlo



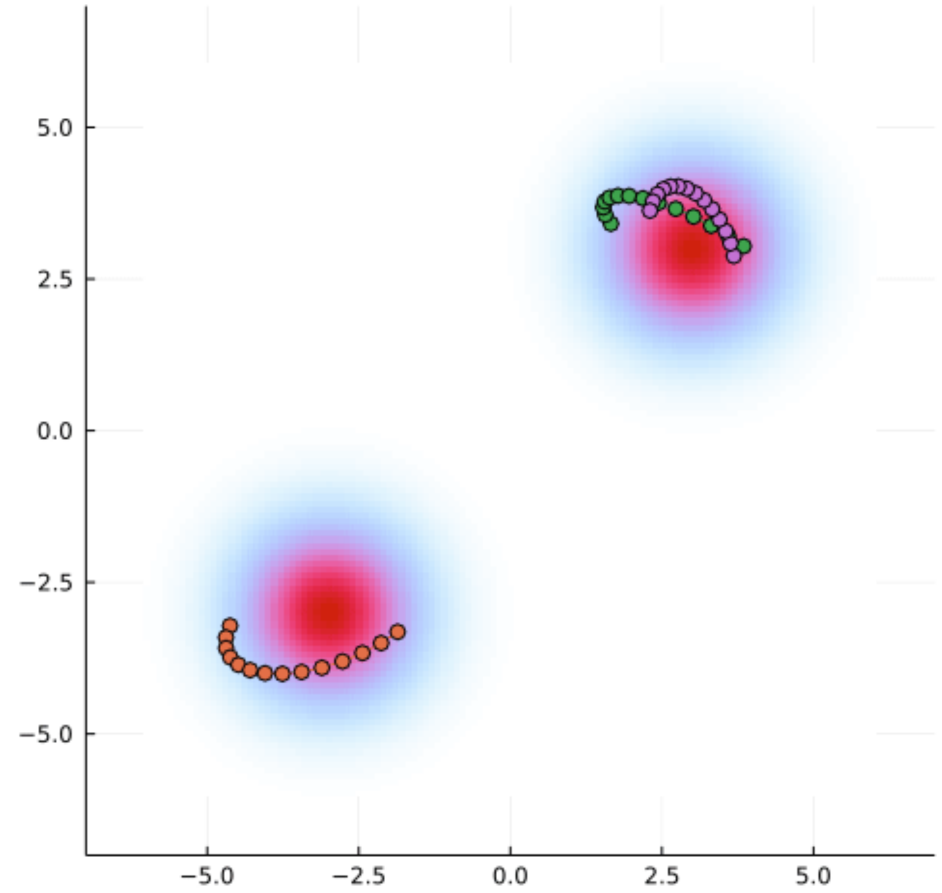
The four pillars of Hamiltonian Monte Carlo

Energy Conservation | Reversibility | Volume Preservation | Symplecticity

Theorem (Arnold (2013))

For every trajectory $\{(\boldsymbol{q}_t, \boldsymbol{p}_t) : t > 0\}$ satisfying the Hamiltonian dynamics

$$\frac{d}{dt} H(\boldsymbol{q}_t, \boldsymbol{p}_t) = 0.$$



$$\alpha(\boldsymbol{q}_t, \boldsymbol{p}_t | \boldsymbol{q}_0, \boldsymbol{p}_0) = \min\left\{1, \frac{\kappa(\boldsymbol{q}_0, \boldsymbol{p}_0 | \boldsymbol{q}_t, \boldsymbol{p}_t) \cdot \cancel{e^{-H(\boldsymbol{q}_t, \boldsymbol{p}_t)}}}{\kappa(\boldsymbol{q}_t, \boldsymbol{p}_t | \boldsymbol{q}_0, \boldsymbol{p}_0) \cdot \cancel{e^{-H(\boldsymbol{q}_0, \boldsymbol{p}_0)}}} \cdot \left|\frac{\partial(\boldsymbol{q}_0, \boldsymbol{p}_0)}{\partial(\boldsymbol{q}_t, \boldsymbol{p}_t)}\right|\right\}$$

Symplectic Integration

- The solution $\mathbf{\Phi}_t : (\boldsymbol{q}_0, \boldsymbol{p}_0) \mapsto (\boldsymbol{q}_t, \boldsymbol{p}_t)$ to the system of **nonautonomous differential equations**

$$\tfrac{d}{dt}\boldsymbol{q}_t = \Sigma^{-1}\boldsymbol{p}_t, \qquad \tfrac{d}{dt}\boldsymbol{p}_t = -\nabla U(\boldsymbol{q}_t)$$

is rarely available in practice. It can be numerically solved using the leapfrog scheme

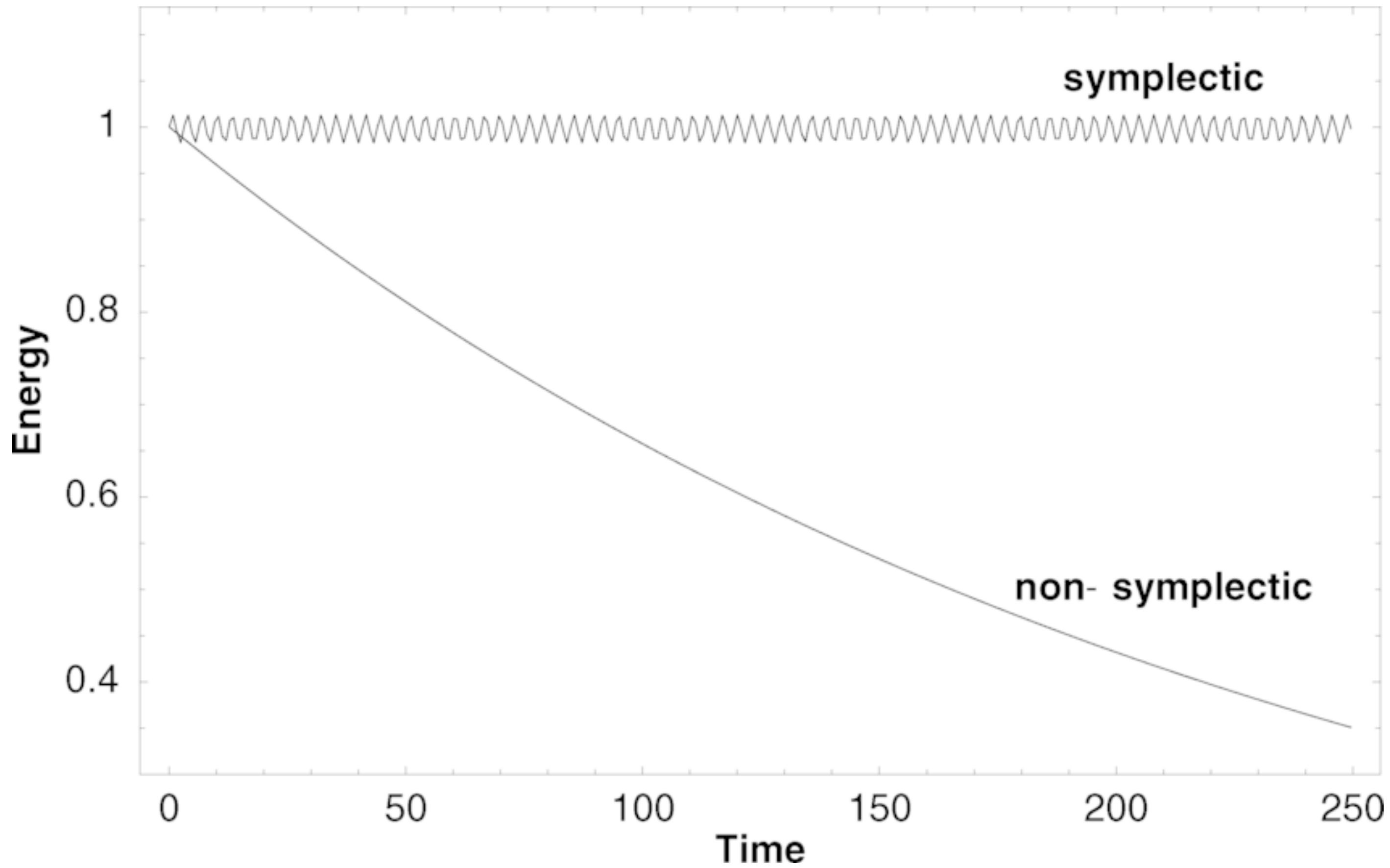
- Let $\epsilon \approx dt$ be a small **step-size**, then $\mathbf{\Phi}_{dt} \approx \mathbf{\Phi}_\epsilon : (\boldsymbol{q}_t, \boldsymbol{p}_t) \mapsto (\boldsymbol{q}_{t+\epsilon}, \boldsymbol{p}_{t+\epsilon})$ is given by

[Math Processing Error] and for time T and $L = \lfloor T/\epsilon \rfloor$, we have

$$\mathbf{\Phi}_T \approx \mathbf{\Phi}_{\epsilon,L} = (\mathbf{\Phi}_\epsilon)^{\otimes L}.$$

Why Symplectic Integration?

Theorem (Hairer, Lubich, and Wanner (2006)) *[Math Processing Error]* Therefore, $\alpha(\boldsymbol{q}_T, \boldsymbol{p}_T | \boldsymbol{q}_0, \boldsymbol{p}_0) \gtrsim e^{-\epsilon^3}$.



Is HMC the panacea?

Proposition (SV and Tak (2024)) Consider

- The bimodal target: $\pi \propto \mathsf{N}(-b\mathbf{1}, \sigma^2 \mathbb{I}_d) + \mathsf{N}(+b\mathbf{1}, \sigma^2 \mathbb{I}_d)$
- The initial state: $\|\boldsymbol{q}_0 + b\mathbf{1}\| \leq (1 - \delta) \cdot b\sqrt{d}.$

Let $\mathrm{E}(b, d)$ be event given by

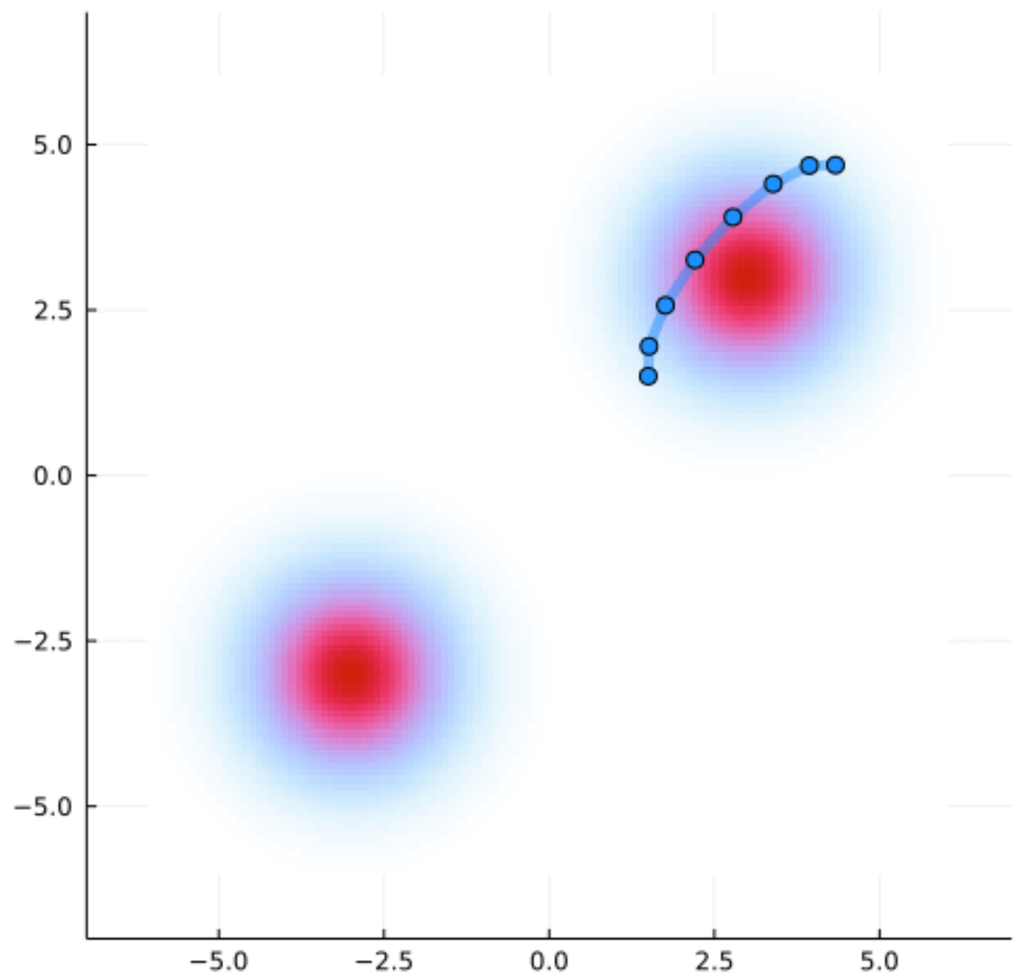
$$\mathrm{E}(b, d) := \{\|\boldsymbol{\Phi}_T(\boldsymbol{q}_0) - b\mathbf{1}\| \leq \|\boldsymbol{\Phi}_T(\boldsymbol{q}_0) + b\mathbf{1}\|\}$$

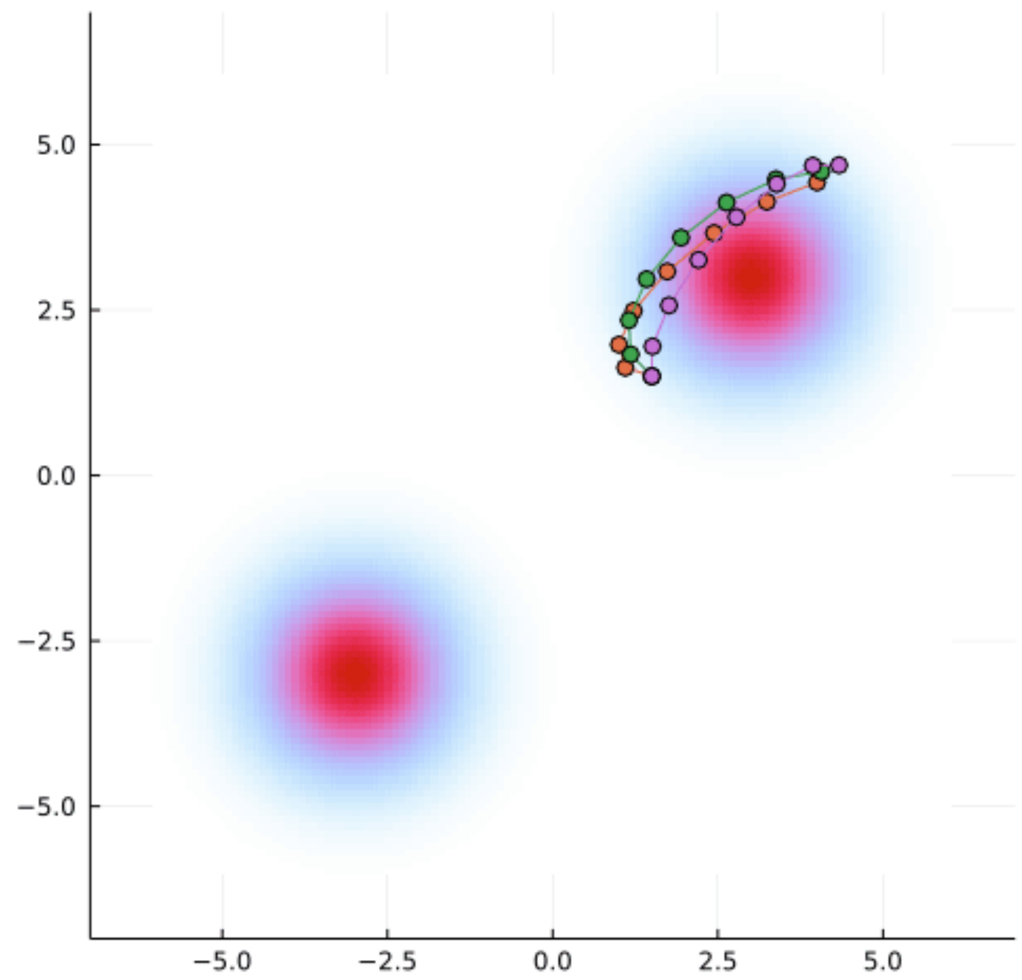
Then for sufficiently large $\delta \in (0, 1)$

$$\mathbb{P}\Big(\mathrm{E}(b, d)\Big) \leq \exp\left(- d \cdot \frac{\delta b^2}{2\sigma^2} \cdot \left(1 - \frac{2\sigma^2}{\delta b^2}\right)^2\right)$$

Mode transitions are incredibly rare in high dimensions

Is HMC the panacea?





Repelling-Attracting Hamiltonian Monte Carlo

Observation #1: Friction ↓ energy

Consider the trajectory of a particle $(\boldsymbol{q}_t, \boldsymbol{p}_t)$ on a rough surface with friciton $\gamma > 0$

$$\frac{d}{dt}\boldsymbol{q}_t = \Sigma^{-1}\boldsymbol{p}_t, \qquad \frac{d}{dt}\boldsymbol{p}_t = -\nabla U(\boldsymbol{q}_t) - \gamma \boldsymbol{p}_t$$

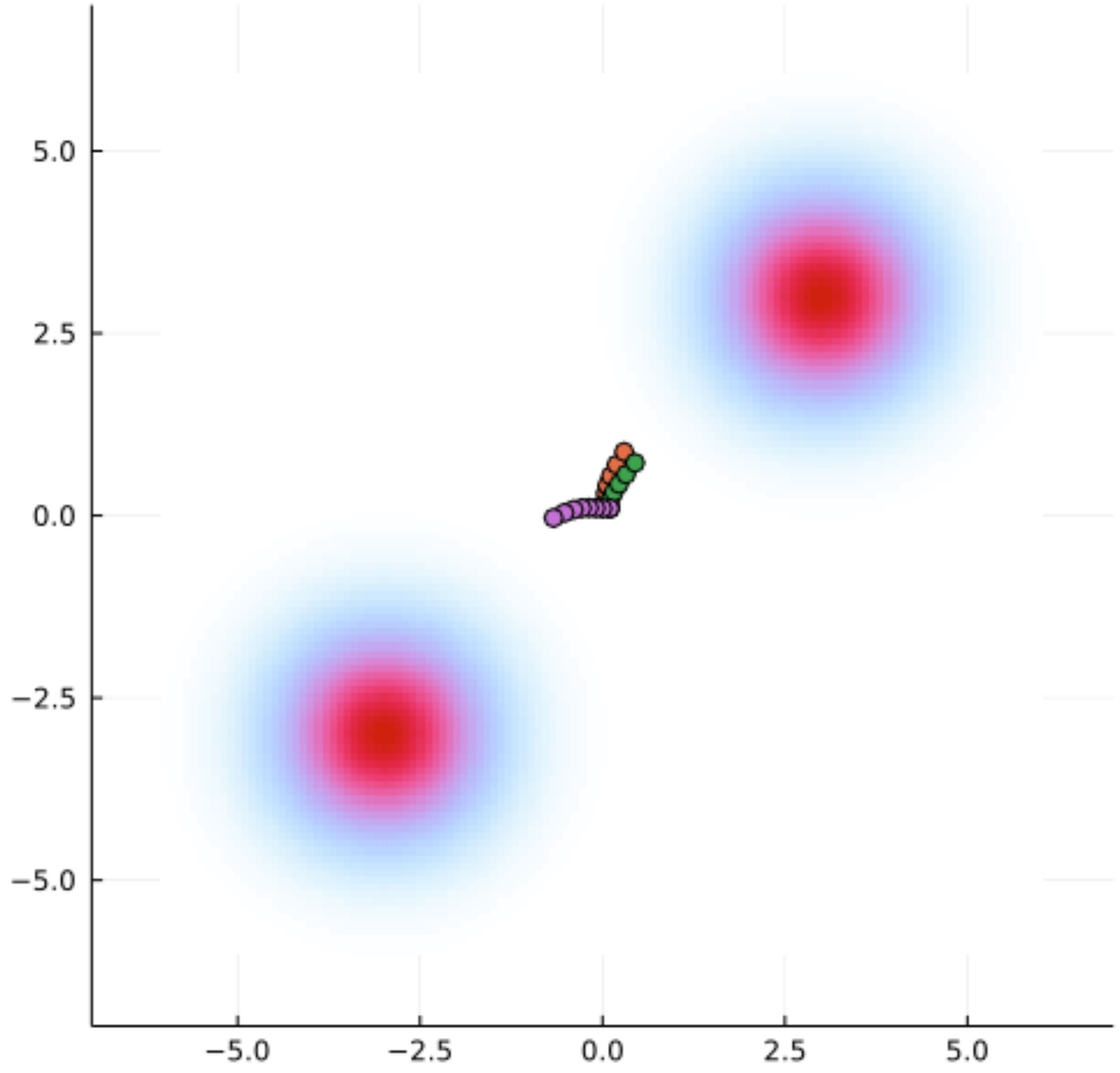
This can be rewritten as

$$\frac{d}{dt}\begin{bmatrix} \boldsymbol{q}_t \\ \boldsymbol{p}_t \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{O} & \mathbb{I} \\ -\mathbb{I} & \mathbf{O} \end{bmatrix}}_{=\boldsymbol{\Omega}} \underbrace{\begin{bmatrix} \nabla_{\boldsymbol{q}} H(\boldsymbol{q}_t, \boldsymbol{p}_t) \\ \nabla_{\boldsymbol{p}} H(\boldsymbol{q}_t, \boldsymbol{p}_t) \end{bmatrix}}_{=\nabla H(\boldsymbol{q}_t, \boldsymbol{p}_t)} - \underbrace{\begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \gamma\mathbb{I} \end{bmatrix}}_{=\boldsymbol{\Gamma}} \begin{bmatrix} \boldsymbol{q}_t \\ \boldsymbol{p}_t \end{bmatrix}.$$

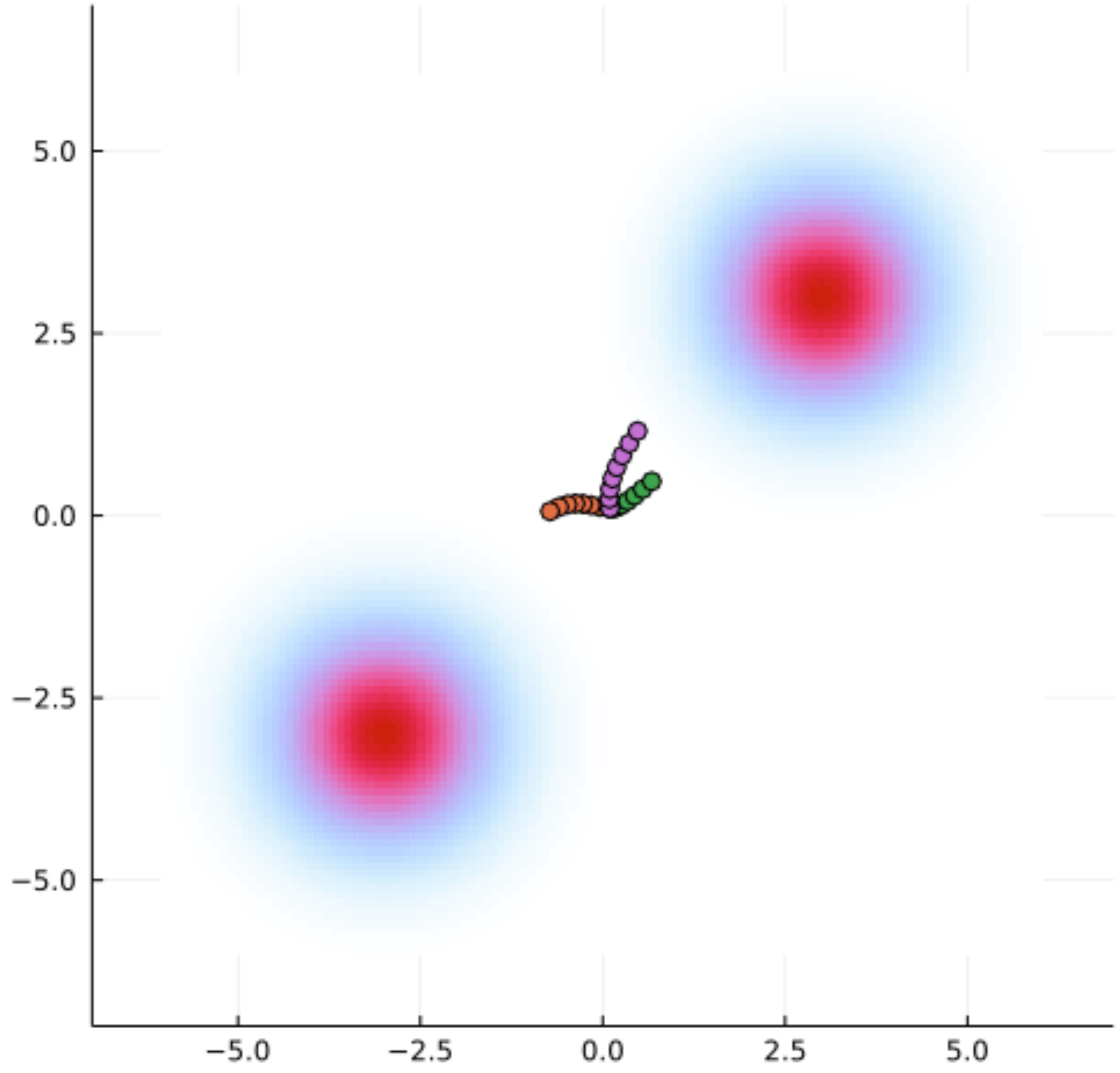
Therefore, we get the conformal Hamiltonian system (A):

$$\frac{d}{dt}(\boldsymbol{q}_t, \boldsymbol{p}_t) = \boldsymbol{\Omega}\nabla H(\boldsymbol{q}_t, \boldsymbol{p}_t) - \boldsymbol{\Gamma}(\boldsymbol{q}_t, \boldsymbol{p}_t). \qquad \text{(A)}$$

Observation #1: Friction ↓ energy



Without friction



With friction

Observation #2: Negative friction ↑ energy

If we just flip the sign of the friction parameter

$$+\gamma \to -\gamma$$

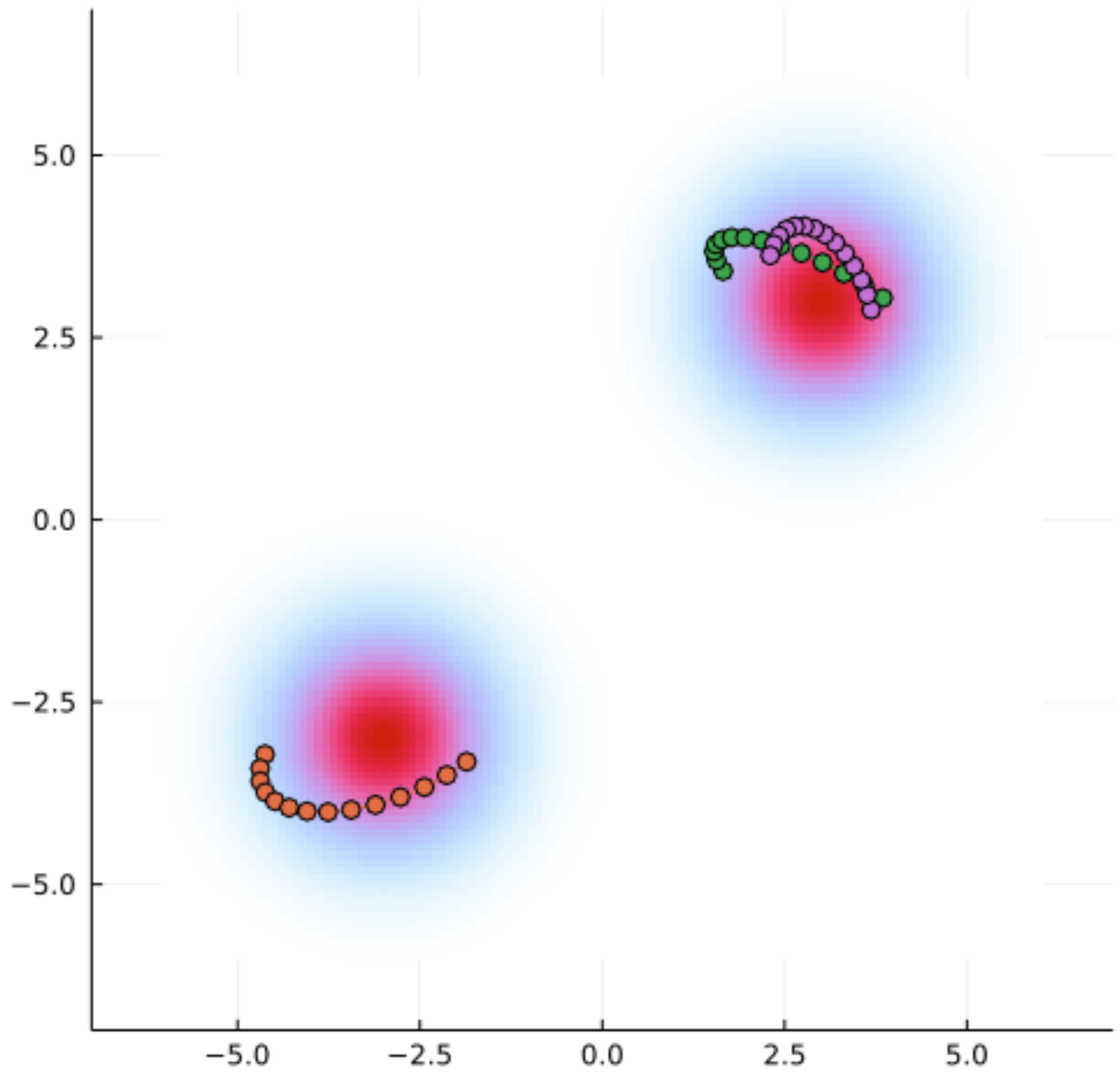
we get

$$\frac{d}{dt}\begin{bmatrix} \boldsymbol{q}_t \\ \boldsymbol{p}_t \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{O} & \mathbb{I} \\ -\mathbb{I} & \mathbf{O} \end{bmatrix}}_{=\boldsymbol{\Omega}} \underbrace{\begin{bmatrix} \nabla_{\boldsymbol{q}} H(\boldsymbol{q}_t, \boldsymbol{p}_t) \\ \nabla_{\boldsymbol{p}} H(\boldsymbol{q}_t, \boldsymbol{p}_t) \end{bmatrix}}_{=\nabla H(\boldsymbol{q}_t, \boldsymbol{p}_t)} + \underbrace{\begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \gamma \mathbb{I} \end{bmatrix}}_{=\boldsymbol{\Gamma}} \begin{bmatrix} \boldsymbol{q}_t \\ \boldsymbol{p}_t \end{bmatrix}.$$

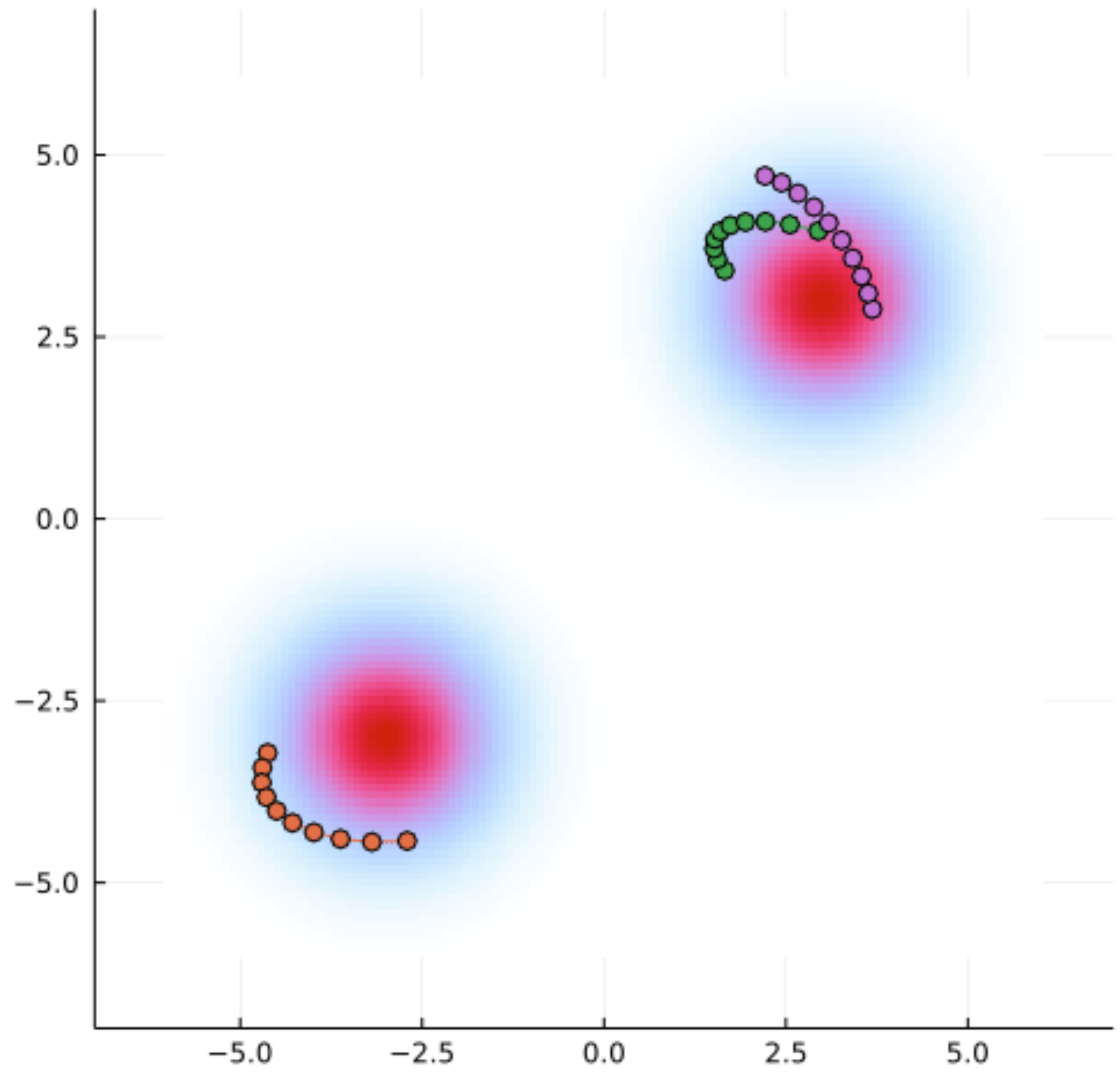
i.e., another conformal Hamiltonian system (B):

$$\frac{d}{dt}(\boldsymbol{q}_t, \boldsymbol{p}_t) = \boldsymbol{\Omega} \nabla H(\boldsymbol{q}_t, \boldsymbol{p}_t) + \boldsymbol{\Gamma}(\boldsymbol{q}_t, \boldsymbol{p}_t). \qquad \text{(B)}$$

Observation #2: Negative friction ↑ energy



Without friction



With negative friction

Repelling-Attracting HMC

1. Choose a hypothetical friction parameter $\gamma \in (0, \infty)$, and integration time T
2. For time $t \in [0, T/2]$ generate conformal Hamiltonian dynamics $\mathbf{\Phi}_t^+$ using negative friction

$$\frac{d}{dt}(\boldsymbol{q}_t, \boldsymbol{p}_t) = \mathbf{\Omega}\nabla H(\boldsymbol{q}_t, \boldsymbol{p}_t) + \mathbf{\Gamma}(\boldsymbol{q}_t, \boldsymbol{p}_t)$$

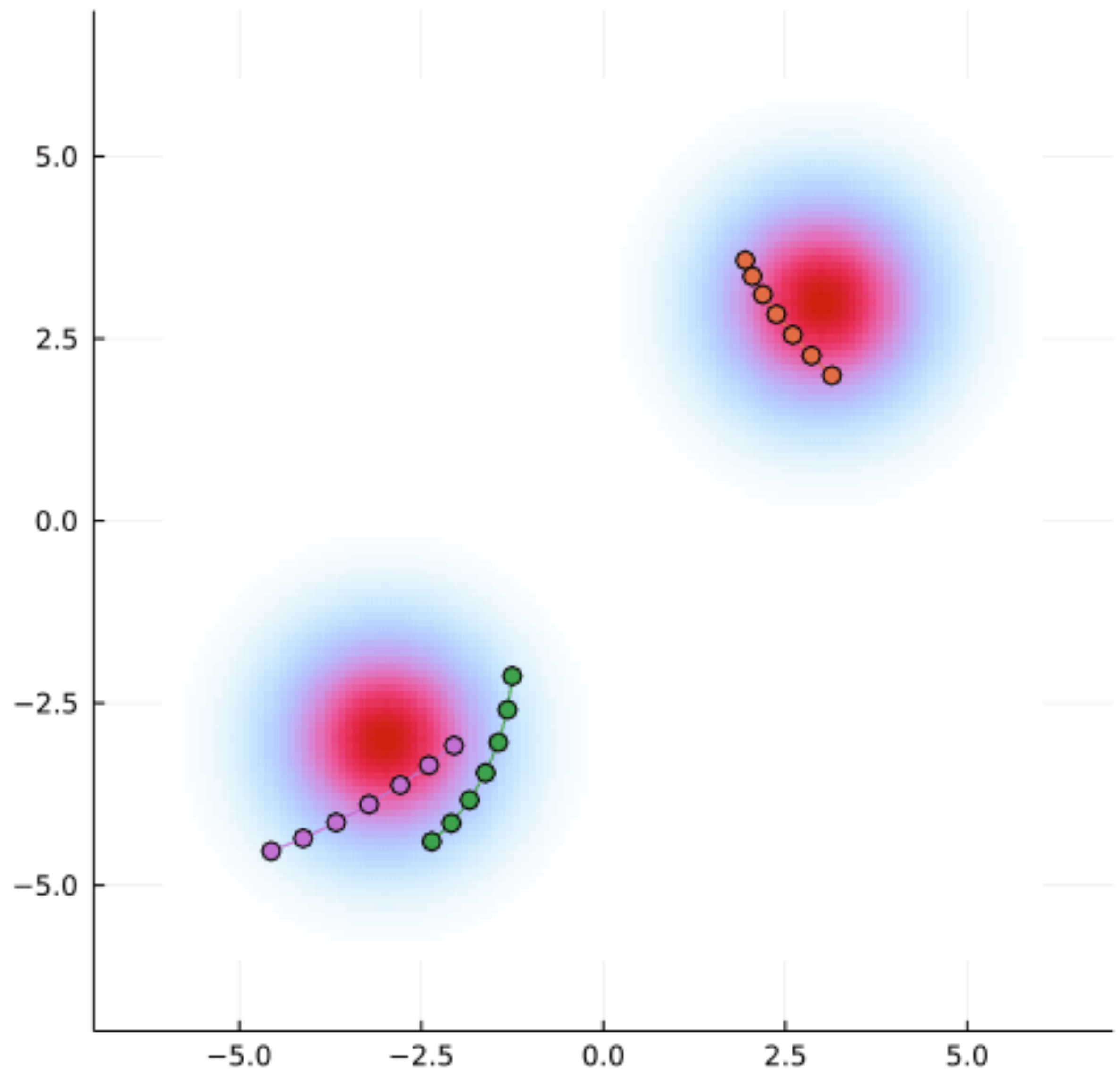
3. For time $t \in [T/2, T]$ generate conformal Hamiltonian dynamics $\mathbf{\Phi}_t^-$ using positive friction

$$\frac{d}{dt}(\boldsymbol{q}_t, \boldsymbol{p}_t) = \mathbf{\Omega}\nabla H(\boldsymbol{q}_t, \boldsymbol{p}_t) - \mathbf{\Gamma}(\boldsymbol{q}_t, \boldsymbol{p}_t)$$

4. Accept/reject state $(\boldsymbol{q}_T, \boldsymbol{p}_T) = \Psi_T(\boldsymbol{q}_0, \boldsymbol{p}_0)$ with MH probability

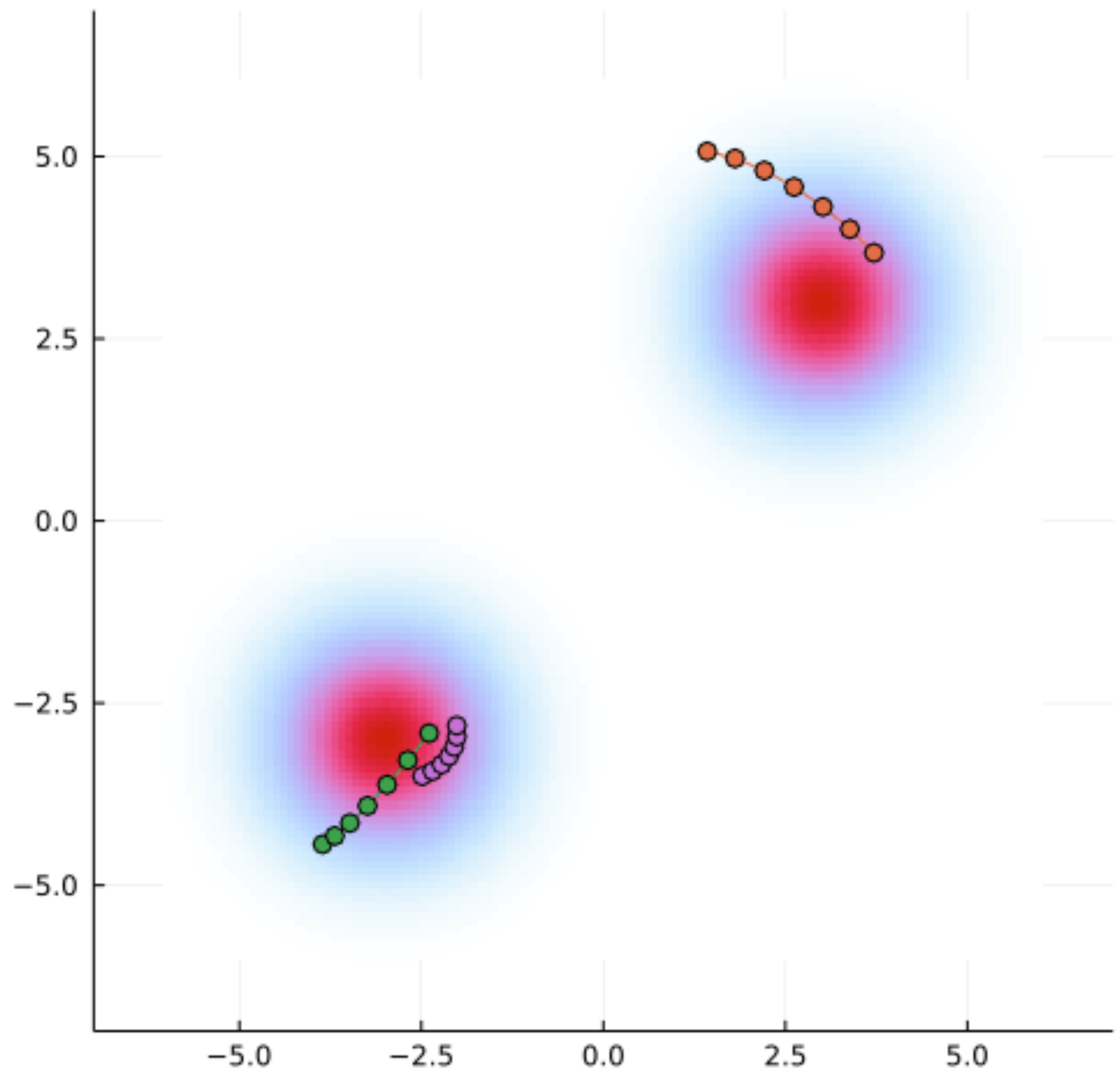
$$\alpha(\boldsymbol{q}_T, \boldsymbol{p}_T | \boldsymbol{q}_0, \boldsymbol{p}_0) = \min\left\{1, \frac{\kappa(\boldsymbol{q}_0, \boldsymbol{p}_0 | \boldsymbol{q}_T, \boldsymbol{p}_T) \cdot e^{-H(\boldsymbol{q}_T, \boldsymbol{p}_T)}}{\kappa(\boldsymbol{q}_T, \boldsymbol{p}_T | \boldsymbol{q}_0, \boldsymbol{p}_0) \cdot e^{-H(\boldsymbol{q}_0, \boldsymbol{p}_0)}} \cdot \left| \frac{\partial(\boldsymbol{q}_t, \boldsymbol{p}_t)}{\partial(\boldsymbol{q}_0, \boldsymbol{p}_0)} \right| \right\}$$

Repelling-Attracting HMC



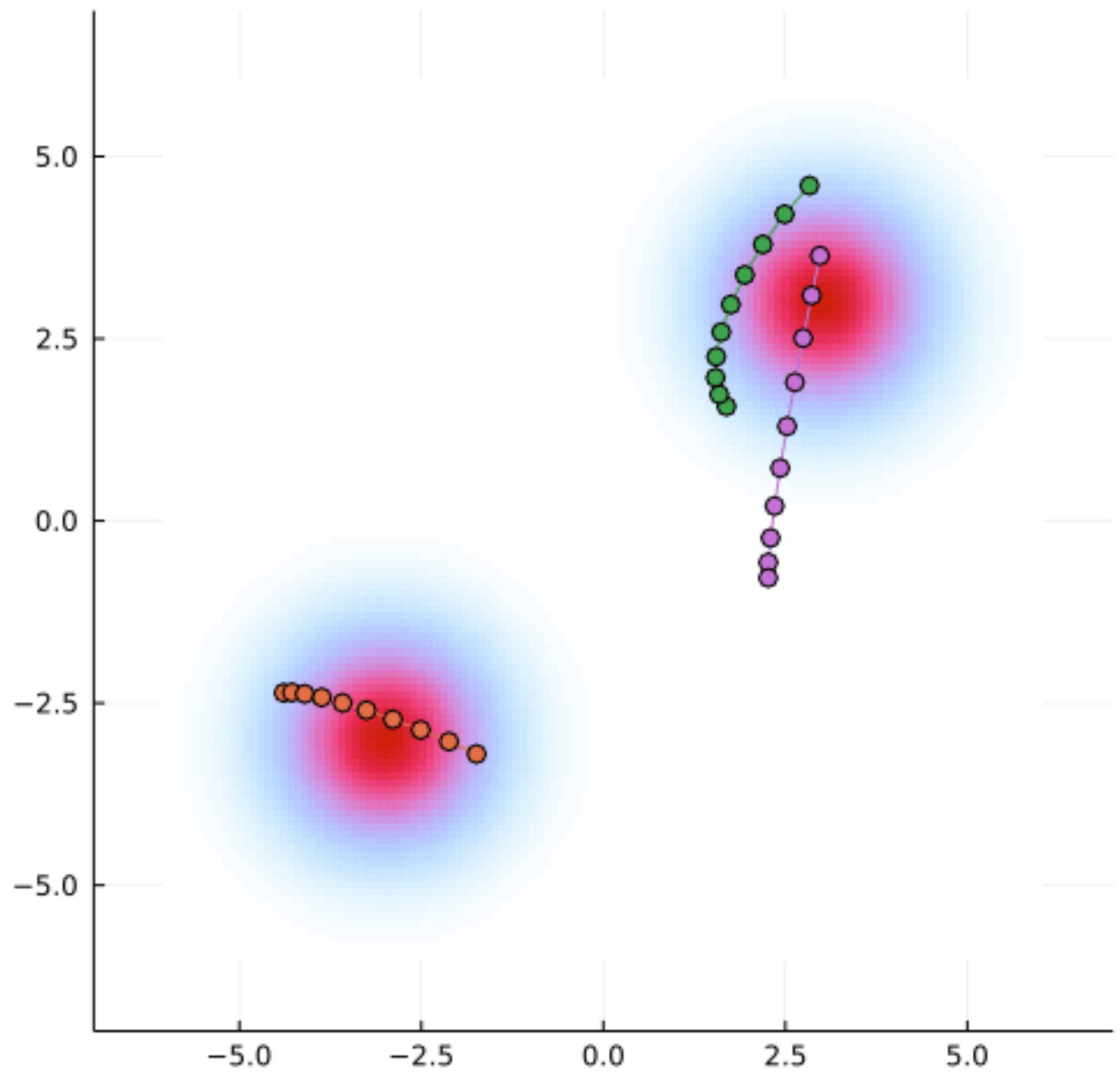
$$\Psi_T = \Phi^{-}_{T/2} \circ \Phi^{+}_{T/2}$$

Repelling-Attracting HMC



$$\Psi_T = \Phi^{-}_{T/2} \circ \Phi^{+}_{T/2}$$

Repelling-Attracting HMC



$$\Psi_T = \Phi^{-}_{T/2} \circ \Phi^{+}_{T/2}$$

Properties of RA-HMC

Reversibility | Volume | Symplecticity | Energy | Numerical Integration

Proposition (SV and Tak (2024))

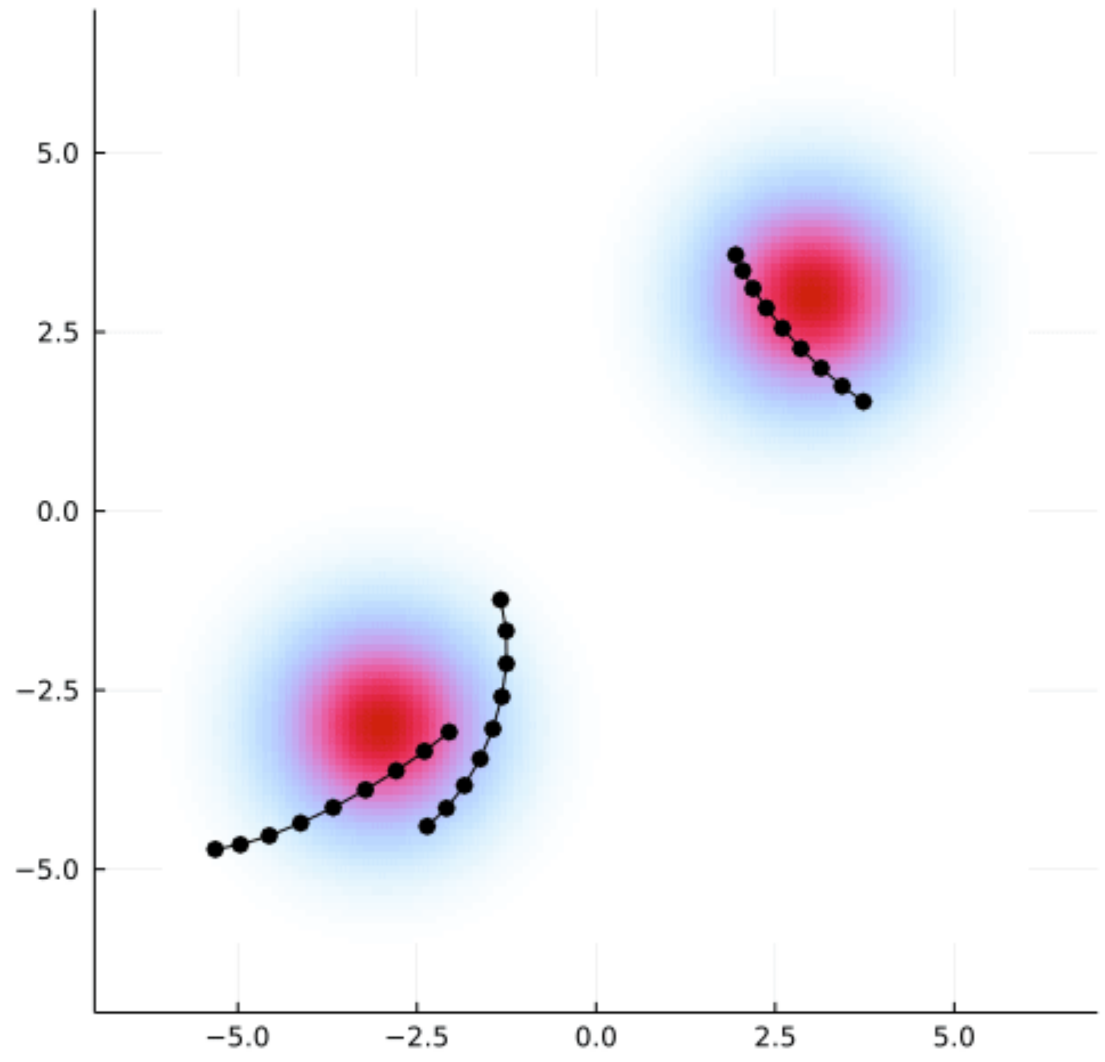
- $\mathbf{F} : (\boldsymbol{q}, \boldsymbol{p}) \to (\boldsymbol{q}, -\boldsymbol{p})$
- $\boldsymbol{\Phi}_t^+$ and $\boldsymbol{\Phi}_t^-$ for the repelling-attracting dynamics

Then

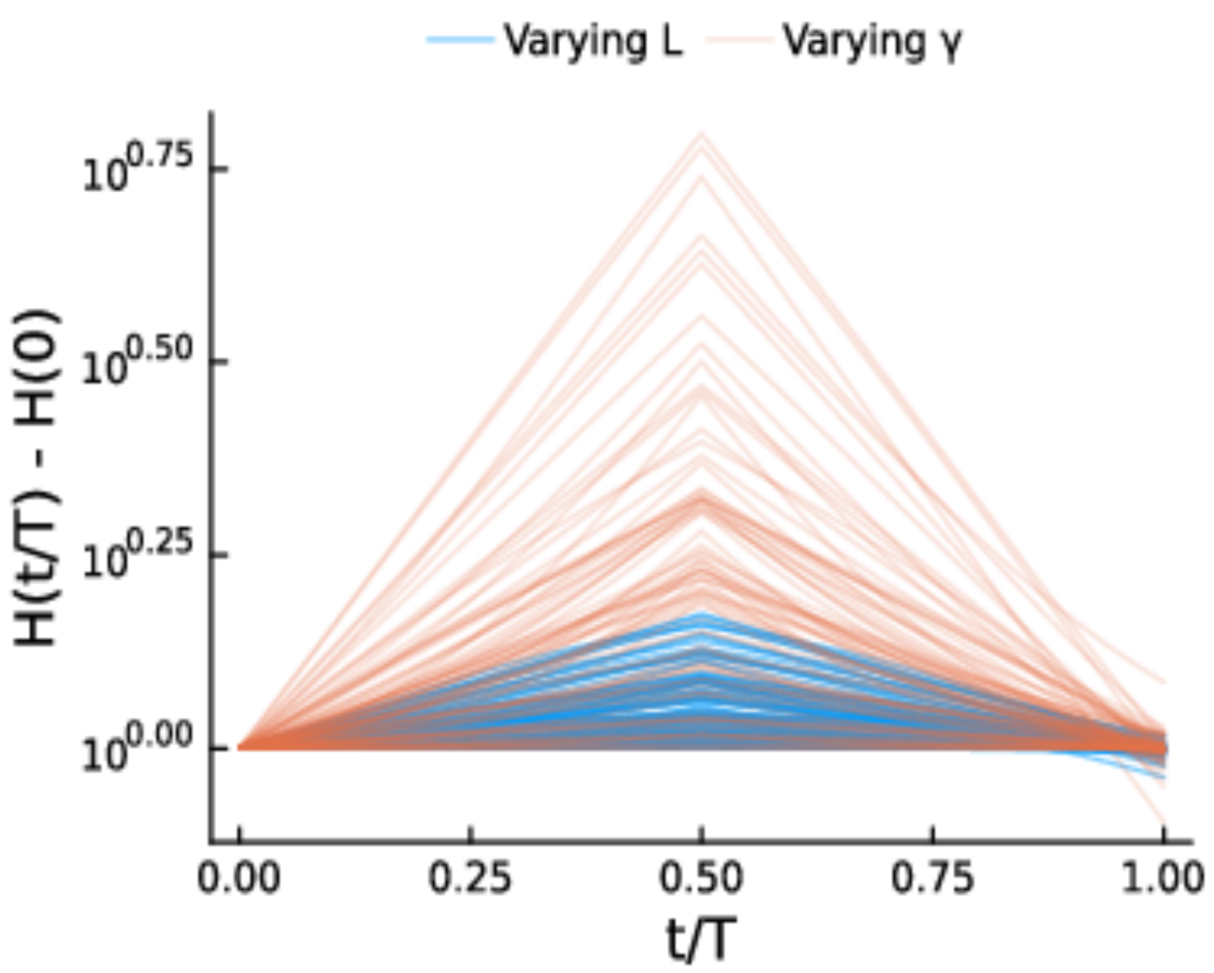
$$\mathbf{F} \circ (\boldsymbol{\Phi}_t^+)^{-1} = \boldsymbol{\Phi}_t^- \circ \mathbf{F}$$

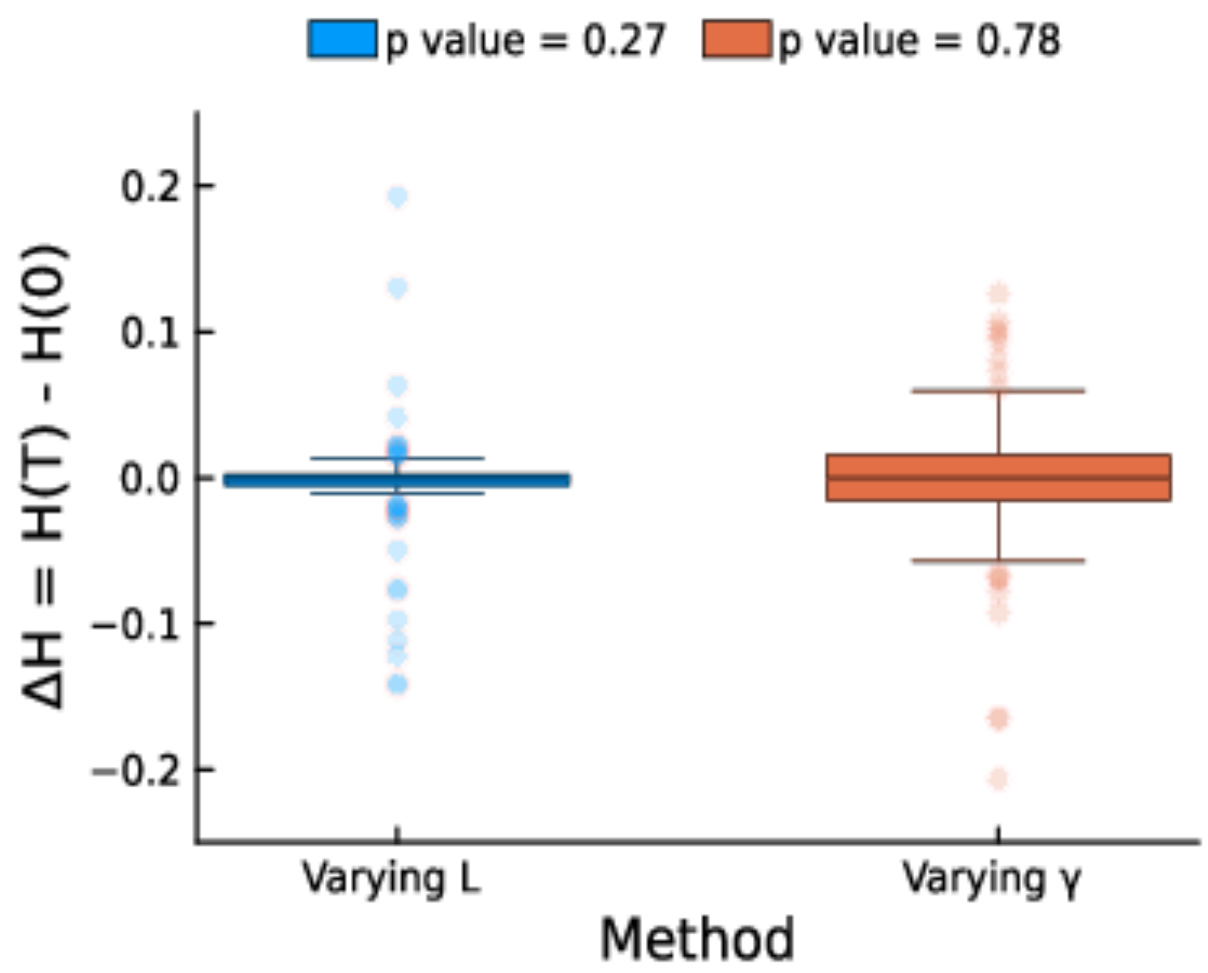
In particular, for $\boldsymbol{\Psi}_{2t} = \boldsymbol{\Phi}_t^- \circ \boldsymbol{\Phi}_t^+$

$$\left(\mathbf{F} \circ \boldsymbol{\Psi}_{2t}\right) \circ \left(\mathbf{F} \circ \boldsymbol{\Psi}_{2t}\right) = \mathrm{id}.$$



Energy drift

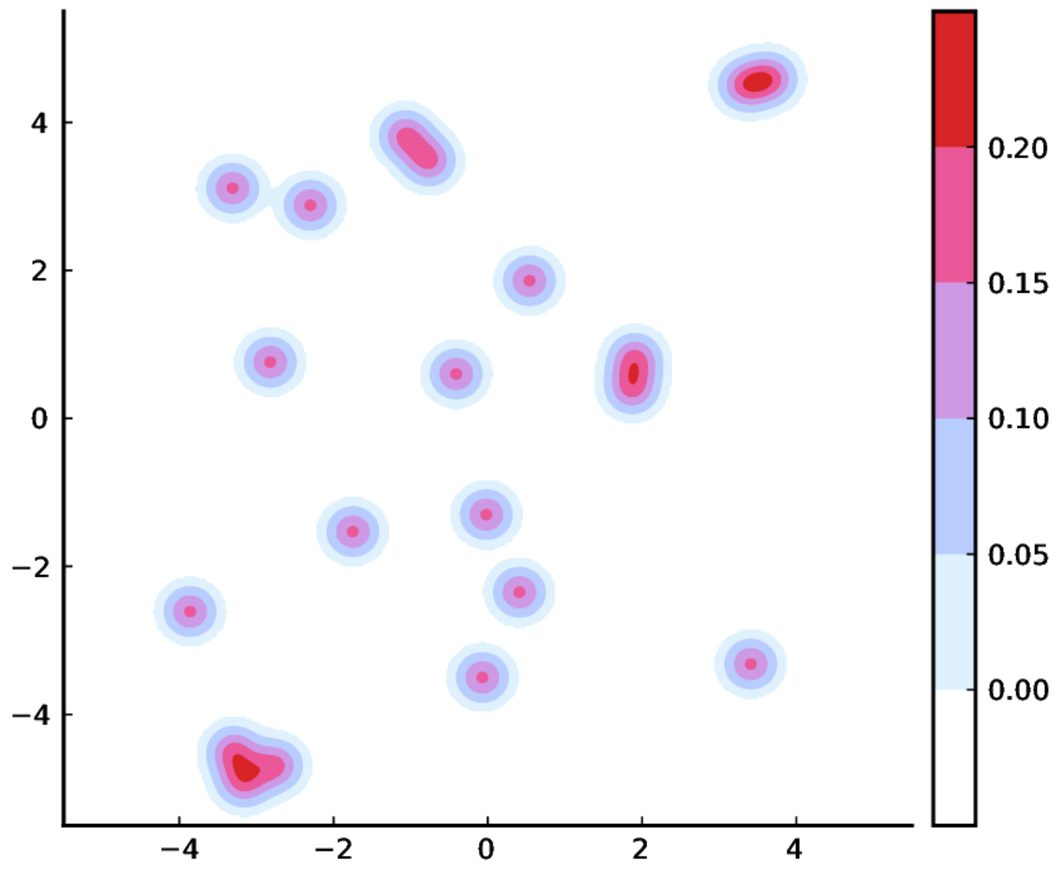




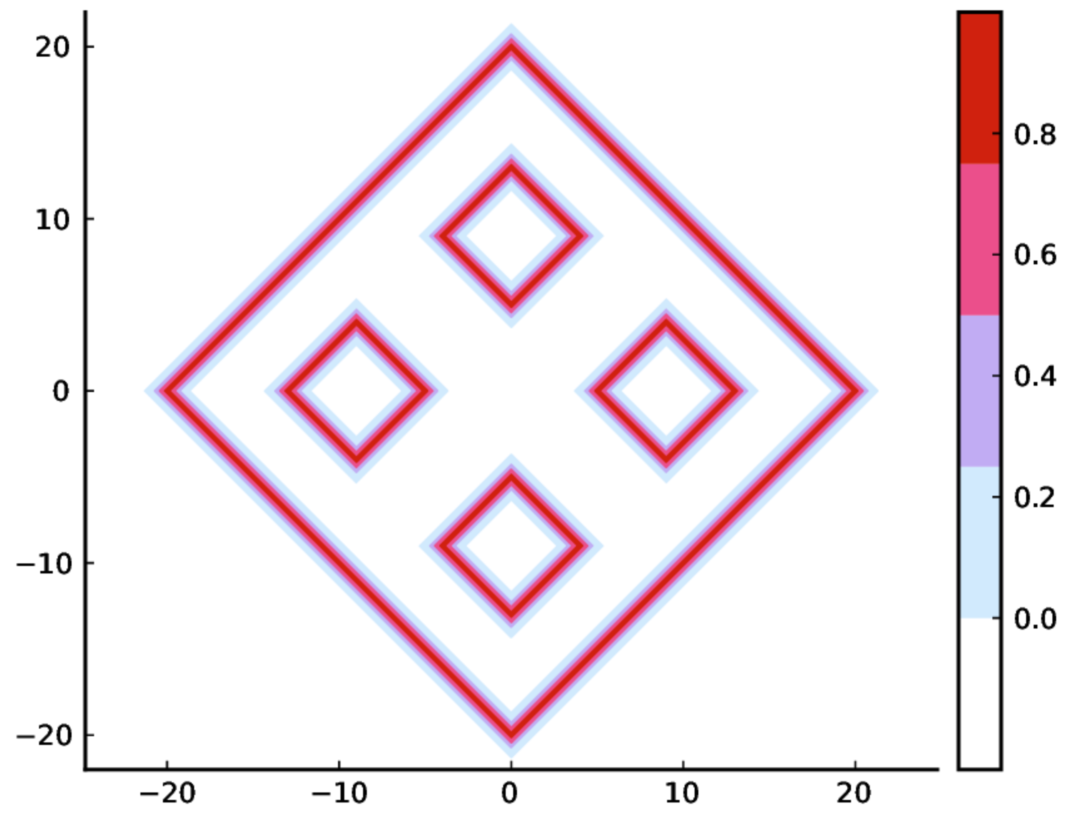
Experiments

Benchmark data & Nested ℓ_1 target

Setup Results Scatter(1) Scatter(2) Trace



Kou, Zhou, and Wong (2006)



Nested ℓ_1

Comparisons:

- RAM (Tak, Meng, and Dyk (2018))
- PEHMC (Nemeth et al. (2019))
- WHMC (Lan, Streets, and Shahbaba (2014))

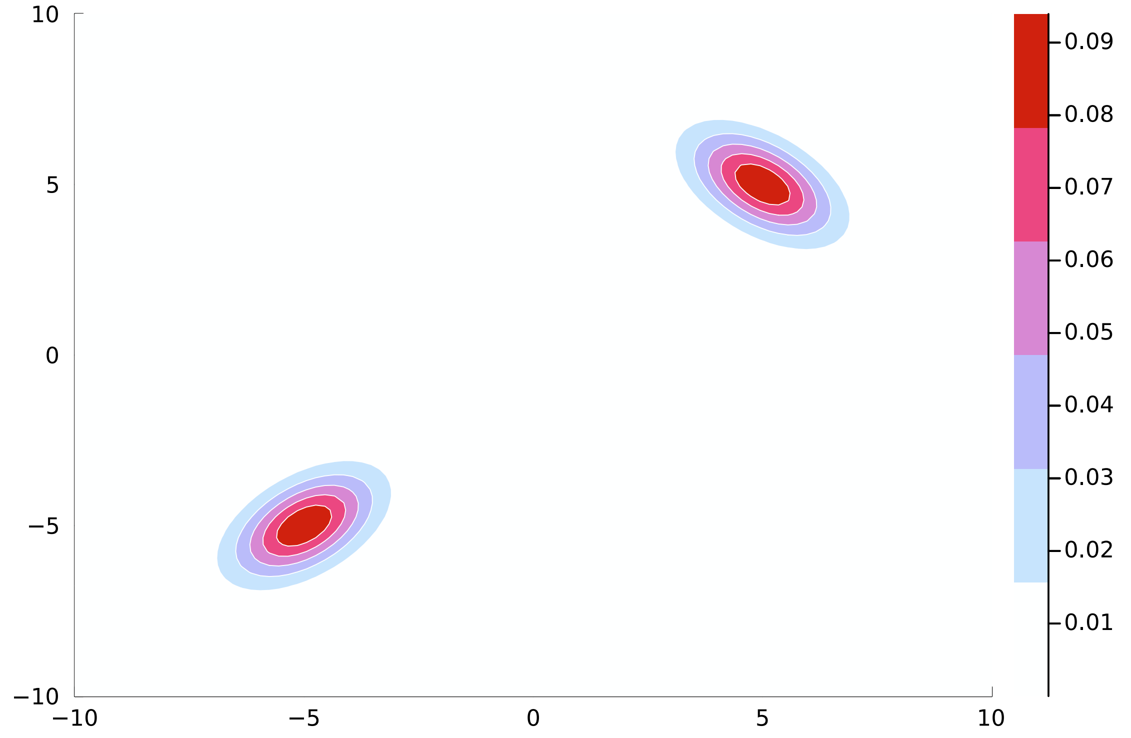
Bimodal Anisotropic Gaussian

Setup | Results | Summary | Mixing | Trace (d=2) | Trace (d=20) | Trace (d=100)

$$\pi \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma_1) + \mathrm{N}(-\boldsymbol{\mu}, \Sigma_2)$$

where

- $\boldsymbol{\mu} = 5\mathbf{1}_d \in \mathbb{R}^d$
- $d \in \{2, 10, 50\}$
- $\Sigma_1(i, j) = 0.5^{|i-j|}$
- $\Sigma_2 = Q\Sigma_1 Q^\top$ for $Q \in SO(d)$



Enhanced Mixing Unimodal Gaussian

Setup | W_2 metric | Mixing

$$\boxed{\pi \propto \mathrm{N}(0_d, \mathbb{I}_d)}$$

With:

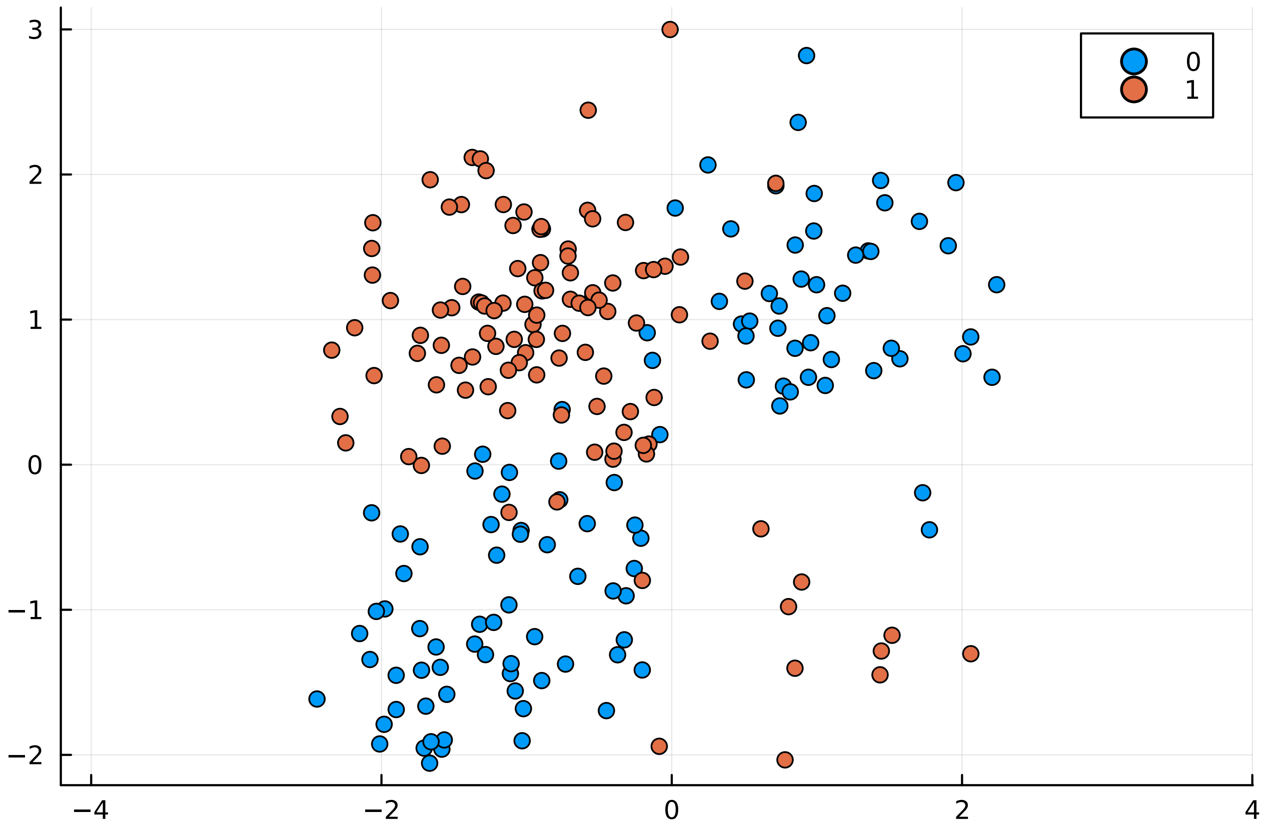
- $d \in \{3, 10, 50, 100\}$
- $n = 5,000$ with $n_{\text{warm-up}} = 5,000$

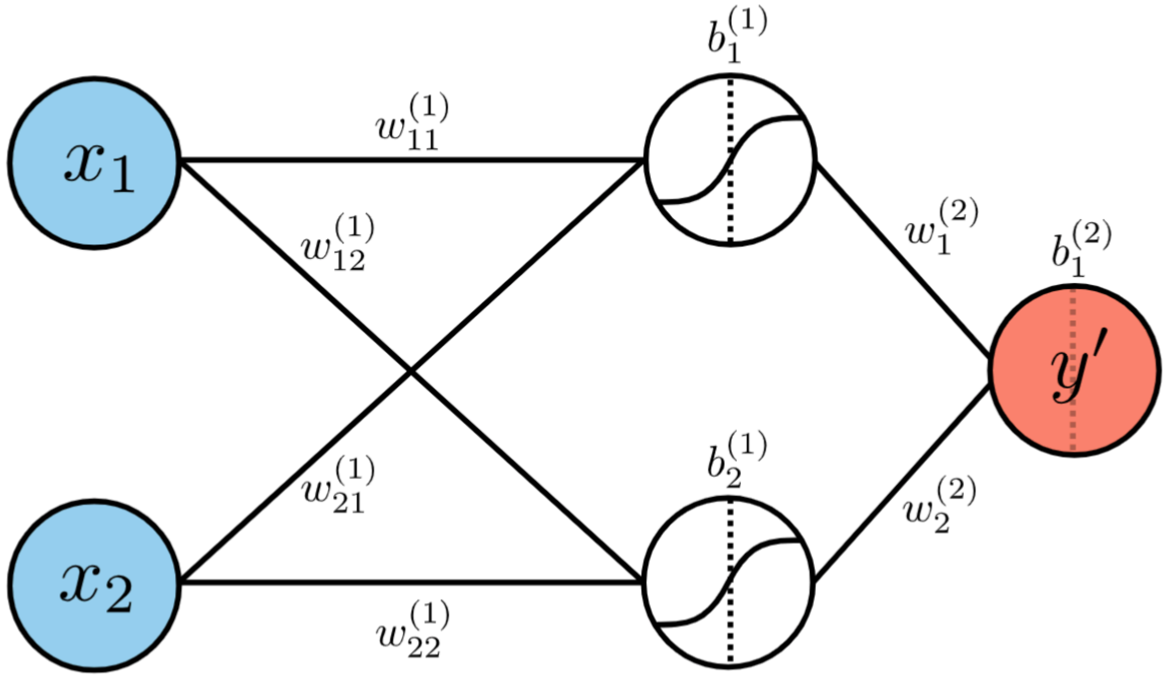
Methods:

- **HMC** with $\epsilon = 0.5$ and $L = 20$
- **RAHMC** with same ϵ, L and $\gamma = 0.05$

Bayesian Neural Network

Setup Prediction Surface Posterior Samples





Bayesian NN posterior for **unbalanced** XOR data exhibits two main modalities (Yallup et al. (2022))

Summary

- RAHMC:

$$\frac{d}{dt}(\boldsymbol{q}_t, \boldsymbol{p}_t) = \begin{cases} \boldsymbol{\Omega}\nabla H(\boldsymbol{q}_t, \boldsymbol{p}_t) + \boldsymbol{\Gamma}(\boldsymbol{q}_t, \boldsymbol{p}_t) & 0 \le t \le T/2 \\ \boldsymbol{\Omega}\nabla H(\boldsymbol{q}_t, \boldsymbol{p}_t) - \boldsymbol{\Gamma}(\boldsymbol{q}_t, \boldsymbol{p}_t) & 0 \le t \le T/2 \end{cases}$$

- Framework can be extended to other variants of HMC:
  - e.g., Magnetic HMC, Non-canonical HMC, Relativistic Monte carlo
- Moving away from the physical analogy of HMC:
  - provides more reliable sampling from multimodal target distributions
  - can lead to better mixing even when target doesn't have any modalities

Open questions

- Can No-U-Turn framework be incorporated to enhance efficiency?
- Ergodicity? Convergence rate?
- Beyond Euclidean spaces

Code

```
using main
using Distributions, Dyna

x = randn(1000)
y = x .+ randn(1000) .* e

@model function lr(x, y)
    σ2 ~ Truncated(Normal
    b0 ~ Cauchy(2.0)
    b1 ~ Normal(0.0, 1.0)
    y ~ MvNormal(b0 .+ b1
end

samples, accepts = mcmc(
  DualAverage(λ=10, δ=0.8
  HaRAM();
  lr(x, y),
  n=1e4, n_burn=1e4
);
```

References

Arnold, Vladimir Igorevich. 2013. *Mathematical Methods of Classical Mechanics*. Vol. 60. Springer Science & Business Media.

França, Guilherme, Jeremias Sulam, Daniel P Robinson, and René Vidal. 2020. "Conformal Symplectic and Relativistic Optimization." *Journal of Statistical Mechanics: Theory and Experiment* 2020 (12): 124008.

Hairer, Ernst, Christian Lubich, and Gerhard Wanner. 2006. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Vol. 31. Springer Science & Business Media.

Kou, S. C., Qing Zhou, and Wing Hung Wong. 2006. "Equi-energy sampler with applications in statistical inference and statistical mechanics." *The Annals of Statistics* 34 (4): 1581–1619.

Lan, Shiwei, Jeffrey Streets, and Babak Shahbaba. 2014. "Wormhole Hamiltonian Monte Carlo." *Proceedings of the AAAI Conference on Artificial Intelligence*, 1953–59.

Nemeth, Christopher, Fredrik Lindsten, Maurizio Filippone, and James Hensman. 2019. "Pseudo-Extended Markov Chain Monte Carlo." *Advances in Neural Information Processing Systems* 32.

SV, and Hyungsuk Tak. 2024. "Repelling-Attracting Hamiltonian Monte Carlo." *arXiv Preprint arXiv:2403.04607*.

Tak, Hyungsuk, Xiao-Li Meng, and David A. van Dyk. 2018. "A Repelling-Attracting Metropolis Algorithm for Multimodality." *Journal of Computational and Graphical Statistics* 27 (3): 479–90.

Yallup, David, Will Handley, Mike Hobson, Anthony Lasenby, and Pablo Lemos. 2022. "Split Personalities in Bayesian Neural Networks: The Case for Full Marginalisation." *arXiv Preprint arXiv:2205.11151*.

Questions?