Enhancing Sampling with Learning: Adaptive Monte Carlo with Normalizing Flows

Inhomogeneous Random Systems:
Statistical Mechanics and Data Science

January 25, 2022

Marylou Gabrié
(CMAP, École Polytechnique)

with:
Grant Rotskoff (Stanford), **Éric Vanden-Eijnden** (Courant Institute, NYU)
James Brofos & Roy Lederman (Yale University), Marcus Brubacker (York university)
Pilar Cossio (Flatiron, CCM), Olga Lopez Acevedo & Ana Molina Taborda (Universidad de Antioquia)
Kaze Wong & Dan Foreman-Mackey (Flatiron, CCA)

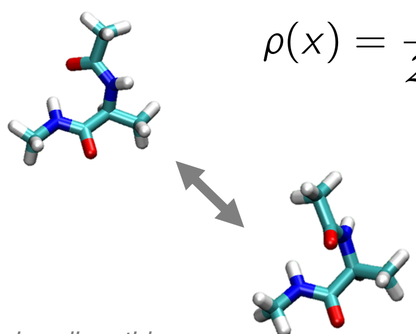
High-dimensional probabilistic models

▷ Ubiquitous in statistical mechasnics / scientific computing in general

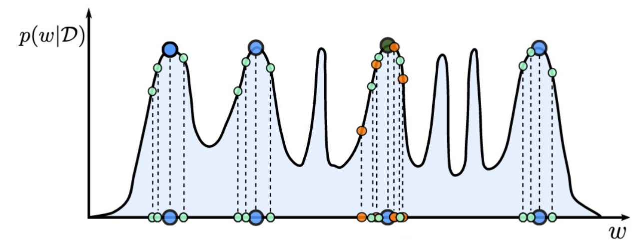
ex: Bayesian models $\rho(\theta|D) = \frac{1}{\mathcal{Z}_D} L(D;\theta)\rho(\theta)$

ex: molecular configurations

$$\rho(x) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta U(x)}$$



Alanine-dipeptide
Jiang et al J. Phys. Chem. B 2019



Deep neural net parameters posterior
Wilson et al. NeurIPS 2020

ex: Training of energy-based models in ML

▷ Random variable $x \in \Omega \subset \mathbb{R}^D$, and density $\rho(x) = \frac{1}{\mathcal{Z}} e^{-U(x)}$ with unknown $\mathcal{Z}$

▷ Task: Compute expectations $\mathbb{E}_\rho[f(x)] = \int_\Omega f(x)\rho(x)\mathrm{d}x$

▷ Method: Monte Carlo approximations, generate $x_1, \dots x_N, \dots$

such that $\mathbb{E}_\rho[f(x)] = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} f(x_i)$

How to obtain samples? Markov Chain Monte Carlo

- ▷ Idea: design transition kernel $\pi(x_{t+1}|x_t)$ such that chain $x_0, x_1, \ldots, x_t$ produces samples from ρ_* for t large enough

[e.g. Liu. *Monte Carlo Strategies in Scientific Computing*, 2004]

- ▷ Important example:

Metropolis-Hastings sampler

Initialize: x_0

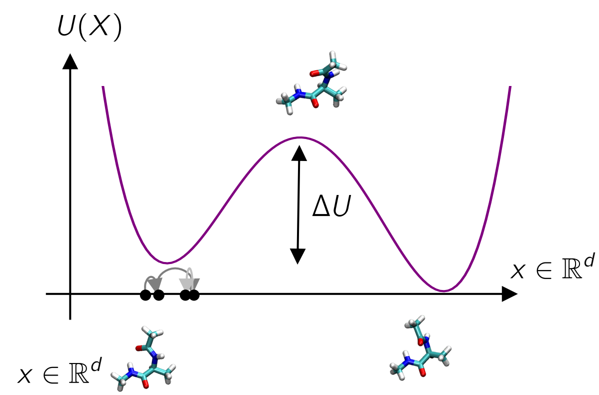
Iterate:

- Propose $\rho_p(x_{t+1}|x_t)$
- Accept reject

$$\mathrm{acc}(x_{t+1}|x_t) = \min\left[1, \frac{\rho_*(x_{t+1})\rho_p(x_t|x_{t+1})}{\rho_*(x_t)\rho_p(x_{t+1}|x_t)}\right]$$

- Update if accept otherwise stay

$$\rho_*(x) = e^{-U(x)}/Z$$



Alanine-dipeptide
Jiang et al J. Phys. Chem. B 2019

- ▷ Issue: decorrelation time
  - Trade-off acceptance/non-local moves *ex: Hamiltonian MC*
  - May not converge/equilibrate in acceptable time if multimodality
    ex: Mode jumping Monte Carlo, "Darting" Monte Carlo

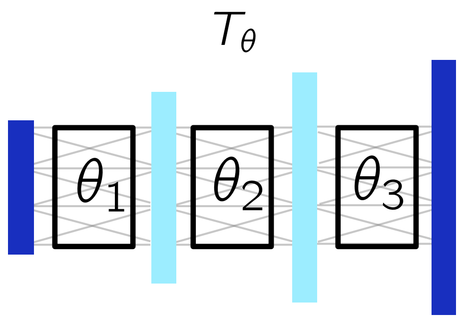
[Duane et al. *PRB* 1987, Tjelmeland & Hegstad Scandinavian *J. of. Stat.* 2001, Andricioaei *J. Chem. Phys*,. 2001, Sminchisescu & Welling *AISTAT* 2017, Pompe et al *Annals of Statistics.* 2020 etc ...]

Deep generative models

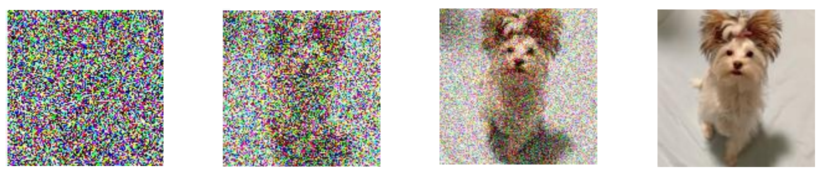
▷ Use transformation T_θ (deep neural network) from simple base distribution ρ_B :

["GANs" Goodfellow et al. *NeurIPS* 2014,
"VAEs" Kingma & Welling *ICLR* 2014,
"Normalizing flows" Papamakarios et al. *JMLR* 2021]

$z \sim \rho_{\mathrm{B}}(z)$



$x = T_\theta(z) \quad \sim \rho_\theta(x)$ "push-forward" distribution



Song et al. *ICLR* 2021

Create independent samples of complicated distributions!

▷ But:

- Needs learn T_θ (do we need data?)
- Even with data from $\rho_*(x) = e^{-U_*(x)}/Z$, unlikely that T_θ creates perfect samples

Can **generative modelling** and **MCMC** be combined into a better solution?

Outline

- ▷ Adaptive MCMC with Normalizing Flows
- ▷ Convergence properties
- ▷ First applications

Initial idea:
Accept/Reject to correct generative model samples

Target density: $\rho_*(x) = e^{-U_*(x)}/Z$

Generative model parametrized density: $\rho_\theta(x)$

▷ Algorithm: Metropolis-Hastings with generative model proposal

Initialize: x_0

Loop:

- Draw from generative model $x_{t+1} \sim \rho_\theta(x)$
- Accept-reject $\text{acc}(x_{t+1}|x_t) = \min\left[1, \frac{\rho_*(x_{t+1})\rho_\theta(x_t)}{\rho_*(x_t)\rho_\theta(x_{t+1})}\right]$

▷ Practical algorithm?

- Can we evaluate and sample from $\rho_\theta(x)$?
- Do we have fast decorrelation? Can we get $\rho_\theta(x) \approx \rho_*(x)$?

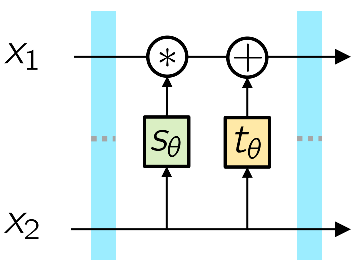
Use Normalizing Flows (NF): Invertible networks (with easy Jacobian)

▷ Parametrized invertible map $T_\theta : \Omega \mapsto \Omega \qquad \Omega \subset \mathbb{R}^d$

- Base distribution $z \sim \rho_B(z)$
- Push-forward distribution $x = T_\theta(z) \quad \sim \rho_\theta(x) = \rho_{\mathrm{B}}(T_\theta^{-1}(x)) \det \left|\nabla_x T_\theta^{-1}\right|$

▷ "Coupling layers": easy-to-compute inverse and Jacobian

Affine coupling layer $T_\theta(x)$

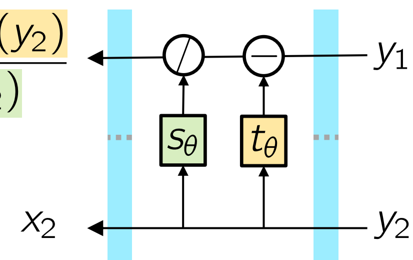


$y_1 = s_\theta(x_2) * x_1 + t_\theta(x_2)$

$y_2 = x_2$

Inverse layer $T_\theta^{-1}(y)$

$x_1 = \dfrac{y_1 - t_\theta(y_2)}{s_\theta(y_2)}$

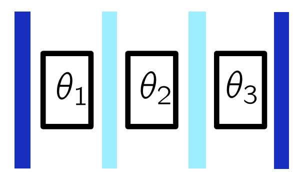


$y_2 = x_2$

Block diagonal Jacobian: $\nabla_x T_\theta(x) = \begin{bmatrix} s_\theta(x_2) I_{d/2} & 0 \\ 0 & I_{d/2} \end{bmatrix}$

$T_\theta = T_{\theta_3} \circ T_{\theta_2} \circ T_{\theta_1}$

▷ Composition to encode for sophisticated transformations



Easy to **sample** and easy to **evaluate density**

[Tabak & V.-E. *Commun. Math. Sci.* 2010, Dinh, L. et al *ICLR* 2017, Papamakarios, G et al *JMLR* 2021]

Training to get $\rho_\theta(x) \approx \rho_*(x)$

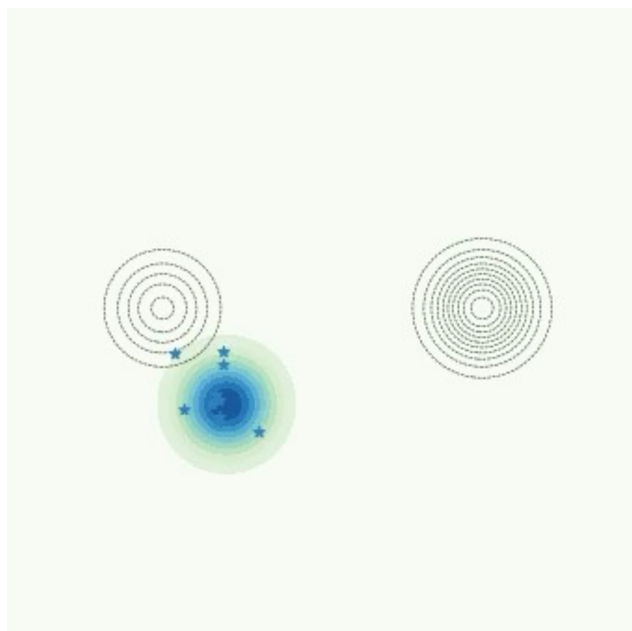
▷ No data a priori, first idea:

minimize “Backward” Kullback-Leibler – “Self-learning” – Variational Inference

easy to obtain!

$$D_{\mathrm{KL}}(\rho_\theta \| \rho_*) = \int \log \frac{\rho_\theta(x)}{\rho_*(x)} \rho_\theta(x) \mathrm{d}x \quad \Longrightarrow \quad L[\rho_\theta] = -\sum_{i=1}^{N} \log \frac{\rho_\theta(x_i)}{\rho_*(x_i)} \quad x_i \sim \rho_\theta(x)$$

example:
learn mixture of 2 Gaussians (2d)



prone to **mode collapse !**

Training to get $\rho_\theta(x) \approx \rho_*(x)$

▷ No data a priori:

minimize "Backward" Kullback-Leibler – "Self-learning" – Variational Inference

easy to obtain!

$$D_{\mathrm{KL}}(\rho_\theta \| \rho_*) = \int \log \frac{\rho_\theta(x)}{\rho_*(x)} \rho_\theta(x) \mathrm{d}x \quad \Longrightarrow \quad L[\rho_\theta] = -\sum_{i=1}^{N} \log \frac{\rho_\theta(x_i)}{\rho_*(x_i)} \qquad x_i \sim \rho_\theta(x)$$

▷ With samples:

minimize "Forward" KL – maximize log-likelihood

hard to obtain!

$$D_{\mathrm{KL}}(\rho_* \| \rho_\theta) = \int \log \frac{\rho_*(x)}{\rho_\theta(x)} \rho_*(x) \mathrm{d}x \quad \Longrightarrow \quad L[\rho_\theta] = -\sum_{i=1}^{N} \log \rho_\theta(x_i) \qquad x_i \sim \rho_*(x)$$

Idea: concurrent sampling-training scheme = **adaptive MCMC**

Adaptive MCMC with Normalizing Flows

Inputs: target energy U_*
normalizing flow T_θ, ρ_B
initial chains $\{x_i(0)\}$ N
training time step η
local kernel $\pi_{\text{local}}(\cdot|\cdot)$,

for $t = 1 : t_{\text{max}}$ **do**
  for i=1,. . . , N **do**

Non-local re-sampling

$x'_{\text{B,i}} \sim \rho_{\text{B}}$, $x'_i = T_\theta(x'_{\text{B},i})$

$x_i(t) \leftarrow x'_i$ with probability $\text{acc}(x_i(t), x'_i)$

Local sampling

$x_i(t+1) \sim \pi_{\text{local}}(x(t+1)|x_i(t))$

$\theta \leftarrow \theta + \eta \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \log \rho_\theta(x_i(t+1))$

NF training step

return: $\{x_i(k)\}_{t=0,i=1}^{t_{\text{max}},N}$

▷ Related to

- Adaptive / "non-linear" Monte Carlo

  [Haario at al *Bernoulli* 2001,
  Jasra et al *Statistics and Computing*, 2007,
  Andrieu et al *Bernoulli* 2011,
  Sejdinovic et al *ICML* 2014,
  Naesseth et al. Neurips 2020]

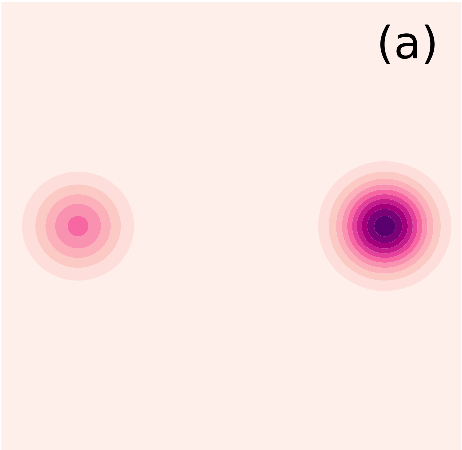
- Local + Mode jumping methods

  [Sminchisescu & Welling *AISTAT* 2017,
  Pompe et al. Ann. Stat 2020,
  Sbailò et al. J. Chem. Phys. 2021]

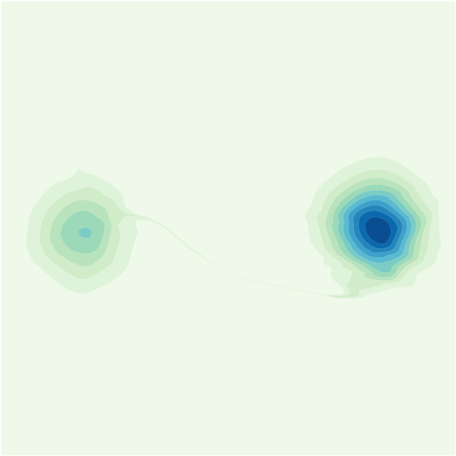
[Gabrié, Rotskoff & Vanden-Eijnden *arxiv:2105.12603 – to appear in PNAS*]

A first small dimensional example: Mixture of two Gaussians in 2d

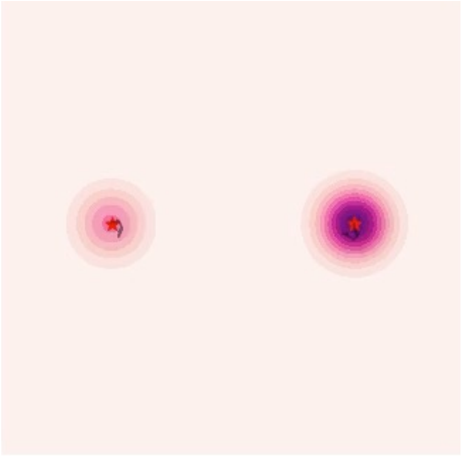
Target density:



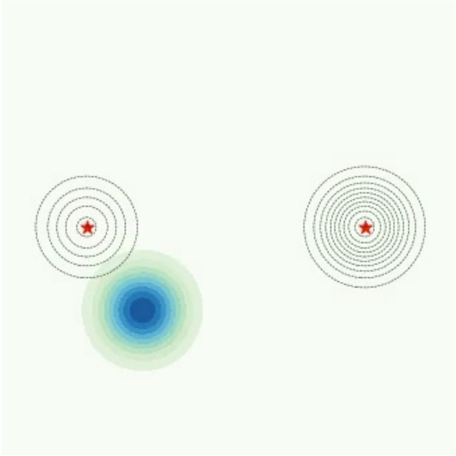
Final learned density:



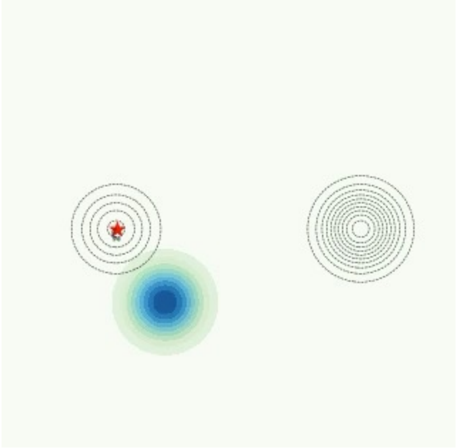
Local method only:



Concurrent:
careful intialization



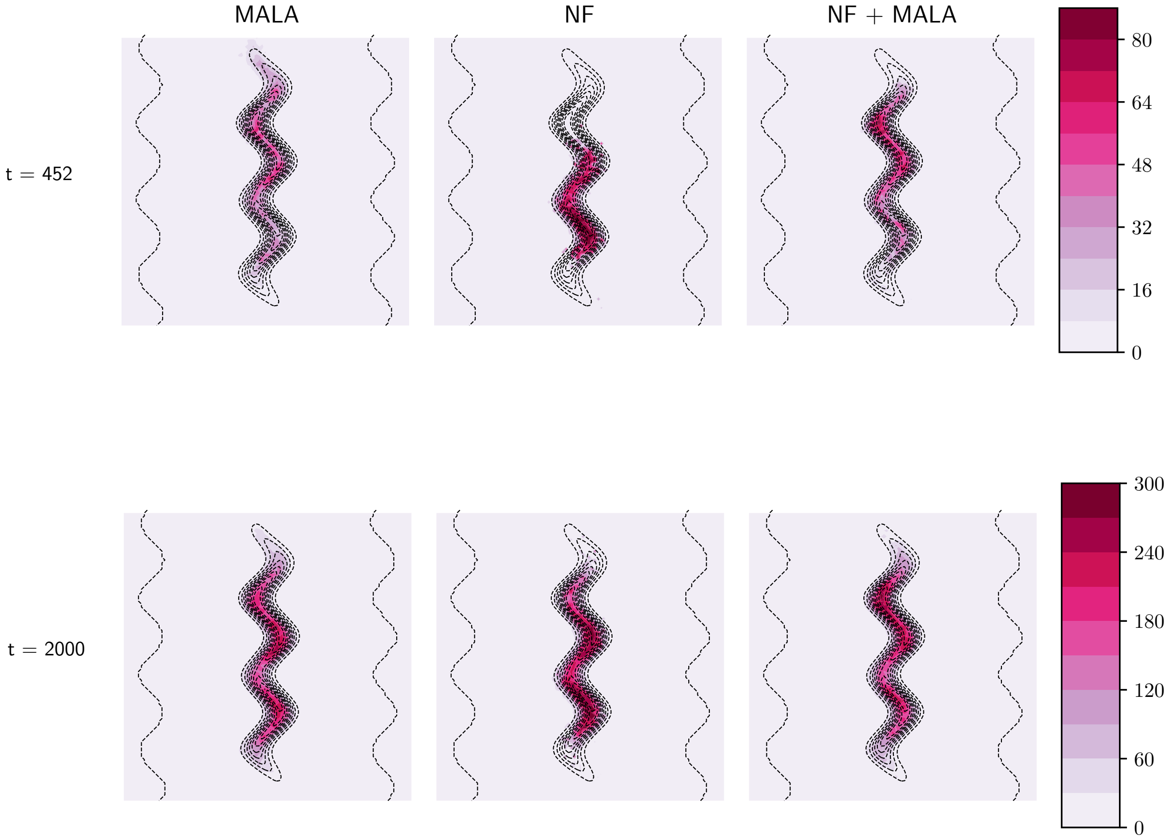
Concurrent:
starting with one walker

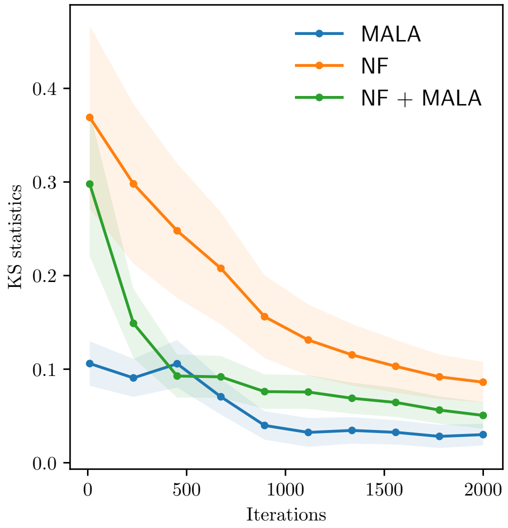


No mode discovery!

Why keep a local update kernel?

Unimodal 2d wiggly distribution:





faster exploration of modes driving learning
+ compensating for imperfect maps.

Outline

- ▷ Adaptive MCMC with Normalizing Flows
- ▷ Convergence properties
- ▷ First applications

Continuous time analysis

▷ Fokker-Planck equation for Langevin dynamics (local sampler)

$$\mathrm{d}x = -\nabla U_*(x)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}W_t \qquad \partial_t \rho_t = \nabla \cdot [\rho_t \nabla U_* + \nabla \rho_t]$$

▷ Resampling as a birth-death process

- Transition kernel when proposing with the NF density

$$\pi(y|x) = \mathrm{acc}(y|x)\rho_\theta(y) + \left(1 - \int_\Omega \mathrm{d}y' \mathrm{acc}(y'|x)\rho_\theta(y')\right)\delta(y-x)$$

- Evolution of density with “birth-death”

$$\partial_t \rho_t(x) = -\alpha \rho_t(x) \int_\Omega \mathrm{acc}(y|x)\rho_\theta(y)\mathrm{d}y + \alpha \rho_\theta(x) \int_\Omega \mathrm{acc}(x|y)\rho_t(y)\mathrm{d}y$$

“particules killed” *“particles resampled”*

▷ Combined dynamics

$$\partial_t \rho_t(x) = \nabla \cdot [\rho_t \nabla U_* + \nabla \rho_t] - \alpha \rho_t(x) \int_\Omega \mathrm{acc}(y|x)\rho_\theta(y)\mathrm{d}y + \alpha \rho_\theta(x) \int_\Omega \mathrm{acc}(x|y)\rho_t(y)\mathrm{d}y$$

Langevin — global resampling

Convergence under perfect training

$$\partial_t \rho_t(x) = \nabla \cdot [\rho_t \nabla U_* + \nabla \rho_t] - \alpha \rho_t(x) \int_\Omega \mathrm{acc}(y|x)\rho_\theta(y)\mathrm{d}y + \alpha \rho_\theta(x) \int_\Omega \mathrm{acc}(x|y)\rho_t(y)\mathrm{d}y$$

Langevin

global resampling

▷ Assume $\forall t$, $\rho_\theta = \rho_t$ (perfect training at all times)

$$D_t \leq \frac{D_0}{1 + 2\alpha D_0 E_0^{-1} t}$$

with Pearson's χ^2 divergences $D_t = \int_\Omega \frac{\rho_t^2}{\rho_*} dx - 1$ and $E_t = \int_\Omega \frac{\rho_*^2}{\rho_t} dx - 1$

Importance of initialization captured by: $E_0 < \infty$, $D_0 < \infty$

Discrete time ergodicity theory

▷ Theory for independent Metopolis-Hastings sampler:

- Independent proposal: $\pi_{\text{prop}}(x^{n+1}|x^n) = \rho_\theta(x^n)$
- Metropolis-Hastings Markov kernel:

$$\pi_{\theta^n}(y|x) = \text{acc}(y|x)\rho_{\theta^n}(y) + \left(1 - \int_\Omega \text{d}y'\text{acc}(y'|x)\rho_{\theta^n}(y')\right)\delta(y-x)$$

▷ The seqence of Markov kernels exhibits **diminishing adaptation** if

$$\lim_{n\to+\infty} \|\pi_{\theta^n}(\cdot) - \pi_{\theta^{n+1}}(\cdot)\|_{\text{TV}} = 0 \text{ in probability.}$$

- e.g.: probability to adapt goes to 0, or converging sequence of

▷ The sequence of Markov kernels exhibits **containement** if:

For any δ , there exists $M(\delta) > 0$ such that

$$\Pr\left(\frac{\rho_*}{\rho_{\theta^n}} \leq M(\delta)\,,\ \forall x \in \mathcal{X}\right) \geq 1-\delta \quad \forall n \in \mathbb{N}$$

▷ Theorems: (Andrieu & Moulines 2006, Roberts & Rosenthal 2007):

If the sequence of Markov kernels exhibits diminishing **adaptation** and **containment**, the chain is ergodic for the distribution ρ^*.

[Brofos, Gabrié, Lederman & Brubacker *arxiv:2110.13216 – AISTAT 2022*]

Outline

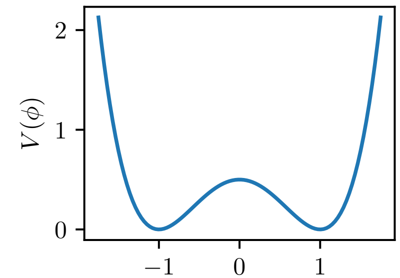
- ▷ Adaptive MCMC with Normalizing Flows
- ▷ Convergence properties
- ▷ First applications

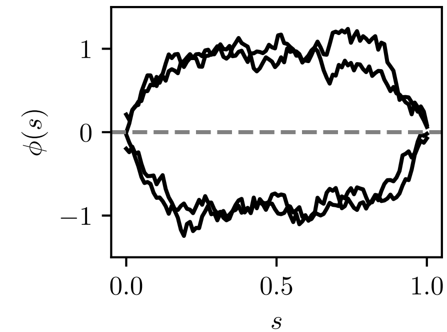
High-dimensional models field system

- Examples: Φ^4 model
  - Random field $\phi\colon [0,1] \mapsto \mathbb{R} \in C([0,1];\mathbb{R})$
  - Energy functional $U_*(\phi) = \int_{[0,1]} \left(\frac{a}{2} |\nabla_s \phi|^2 + V(\phi) \right) \mathrm{d}s$
  - Local potential $V(\phi) = \frac{1}{2}(\phi^2 - 1)^2$
  - Dirichlet boundary conditions $\phi(0) = 0, \phi(1) = 0$
  - Target distribution $\rho(\phi) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta U(\phi)}$

local potential

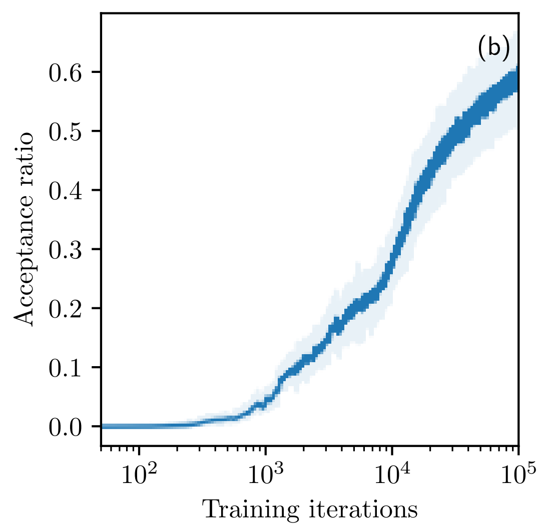
coupling term



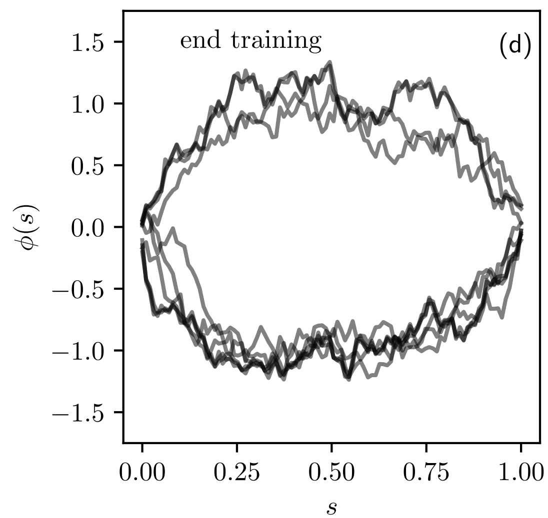


- Discretized: N=100

Acceptance ~ 60%



Fast mixing



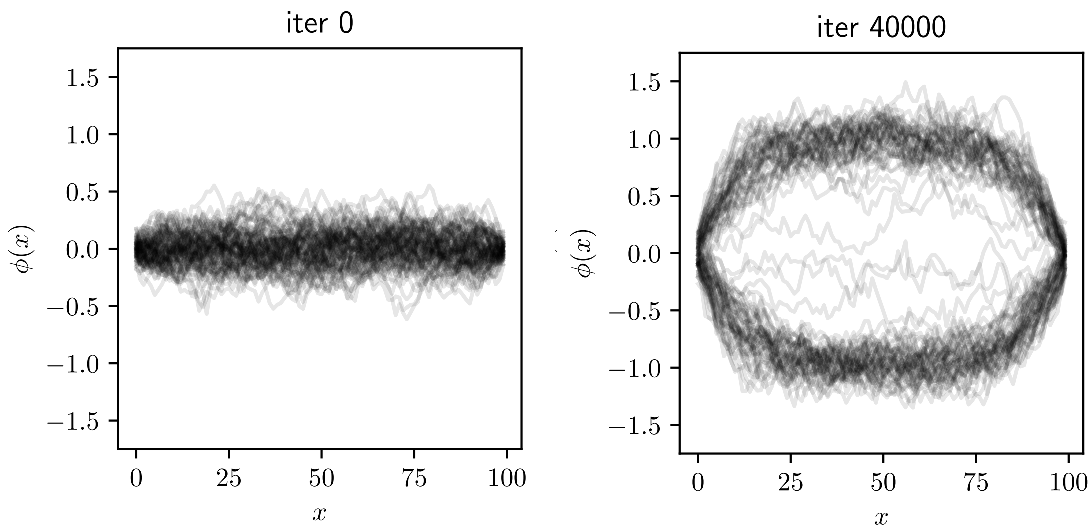
[Gabrié, Rotskoff & Vanden-Eijnden *arxiv:2105.12603 – to appear in PNAS*]

Uncoupled vs coupled base distributions

Gaussian informed (coupled)

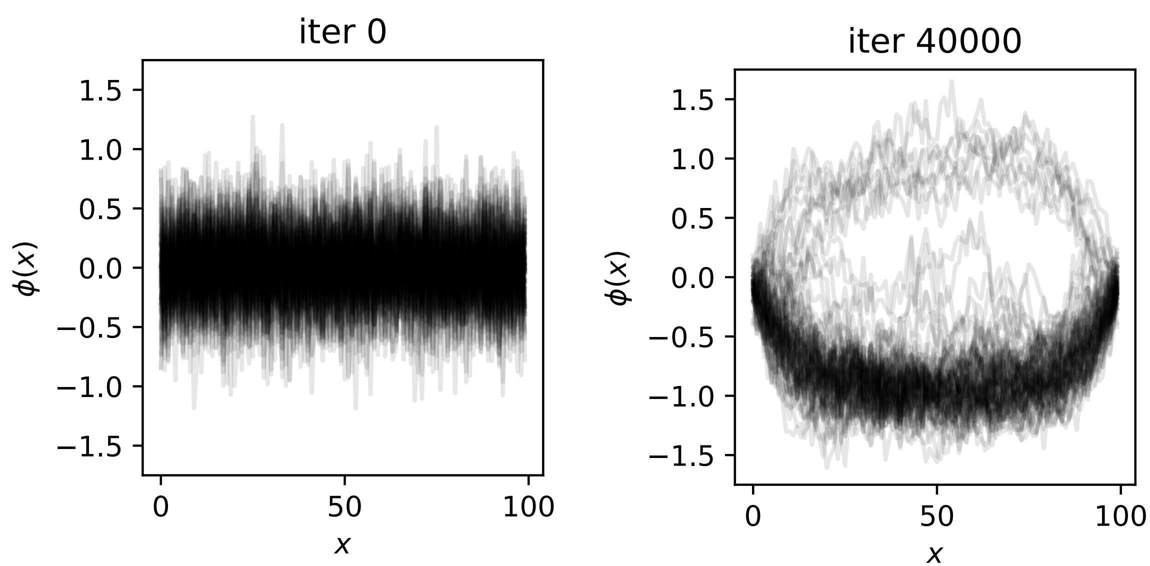
$$U_{\mathrm{B}}(\phi) = \int \left(\frac{a}{2} |\nabla_x \phi|^2 + \frac{1}{2\sigma^2} \phi^2 \right) \mathrm{d}x$$

$$\phi(0) = 0, \phi(1) = 0$$

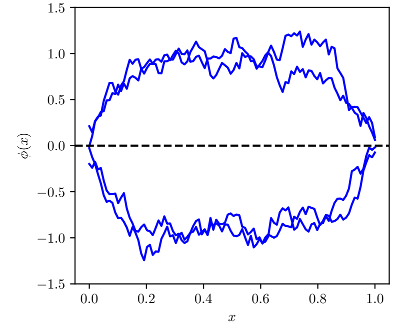


Gaussian uninformed (uncoupled)

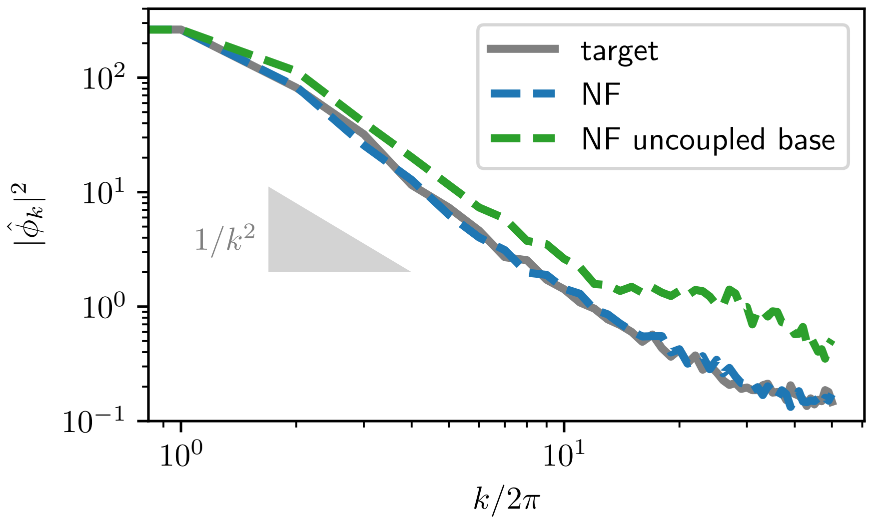
$$U_{\mathrm{B}}(\phi) = \int \frac{1}{2\sigma^2} \phi^2 \mathrm{d}x$$



More numerical checks



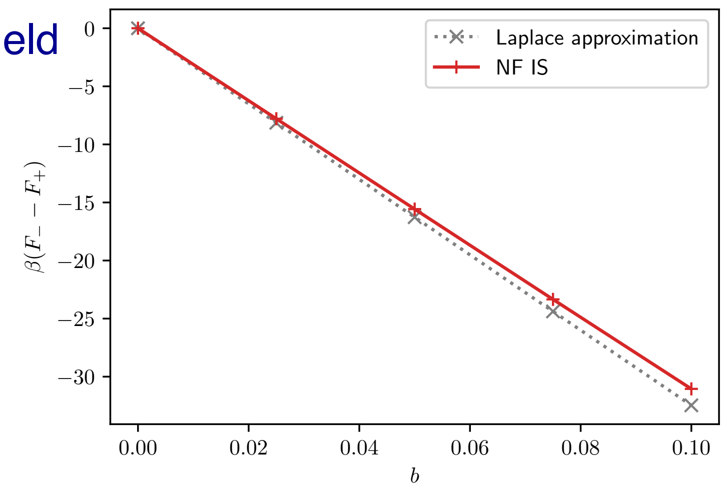
- ▷ Learned Fourier spectrum matches target up to fine scales



- ▷ Tilt distribution towards -1 configuration with local field

$$U_{*,\boldsymbol{b}}(\phi) = \int \left(\frac{a}{2}|\nabla_x \phi|^2 + V(\phi) + \boldsymbol{b}\,\phi\right) \mathrm{d}x$$

free energy difference

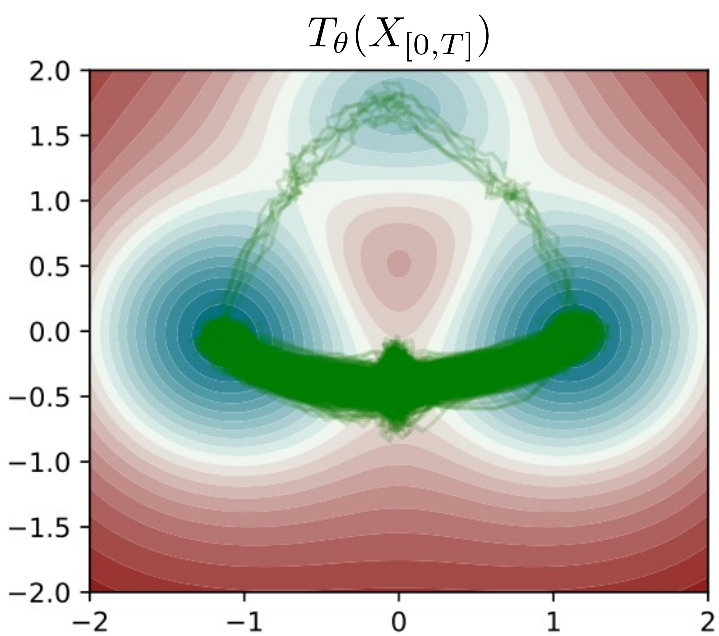


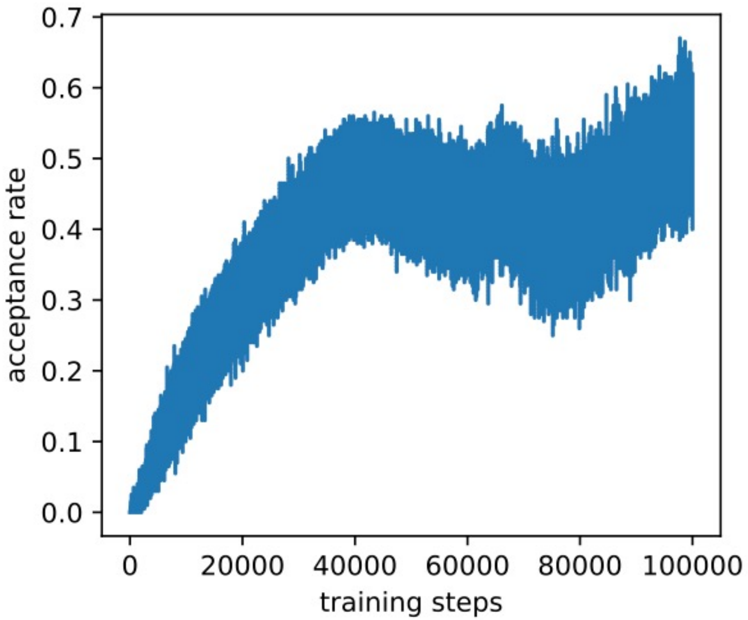
Another field-like example: Transition paths sampling

- ▷ Diffusion with potential drifts (possibly non-conservative forces)

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}W_t \qquad \mathbb{P}(X_{[0,T]}) \propto \exp\left[-\frac{\beta}{2}\int_0^T |\dot{x}_t + b(x_t)|^2 \mathrm{d}t\right]$$

- ▷ Path metastability
  - ○ Base distribution $\mathbb{P}(X_{[0,T]}) \propto \exp\left[-\frac{\beta}{2}\int_0^T |\dot{x}_t|^2 \mathrm{d}t\right] \quad (x_0 = x_A, x_T = x_B)$



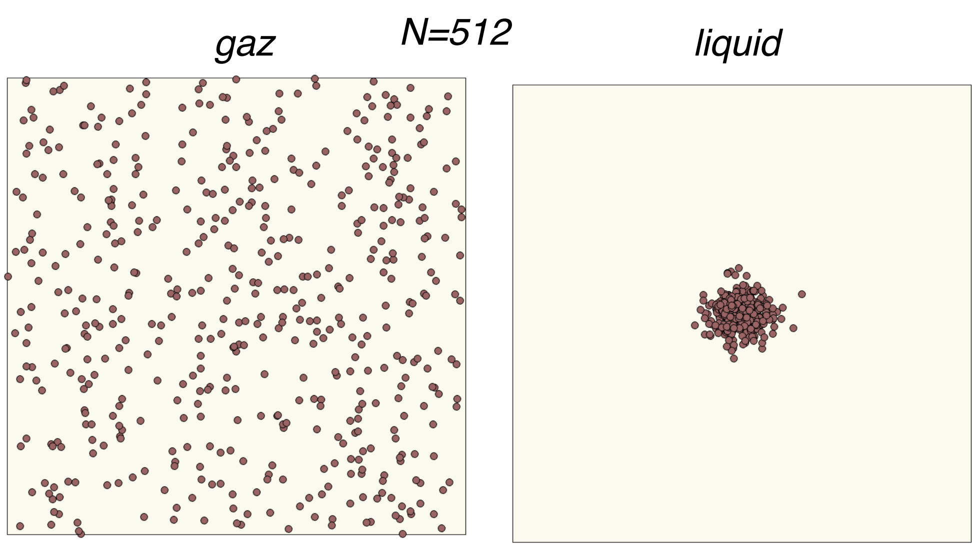


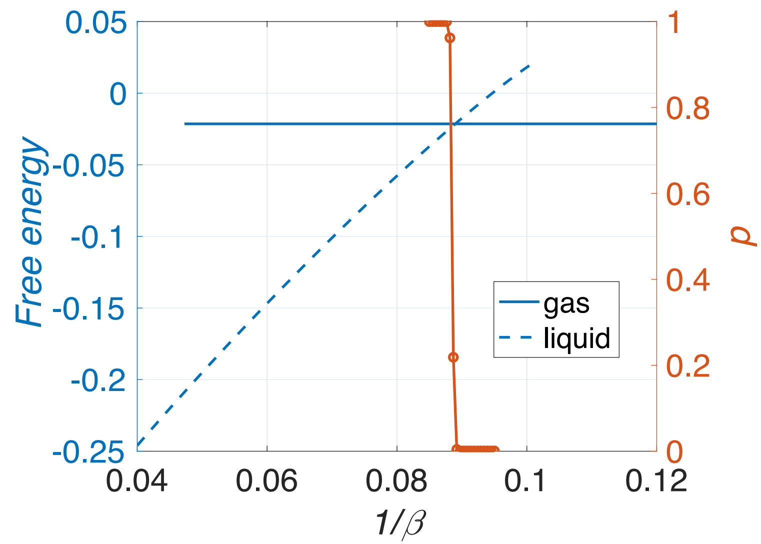
Sampling simple particle systems with phase transition

- System: $X = (x_1, \dots, x_N) \in [0, L]^{2N}$

  Pair-wise short-range interactive potential $\rho_*(X) = Z_*^{-1} \exp\left(-\frac{\beta}{2N}\sum_{i,j=1}^{N} W(x_i - x_j)\right)$

- Model as mixture of push-forwards: $\rho_p(X) = p\rho_{\text{gaz}}(X) + (1-p)\rho_{\text{liquid}}(X)$

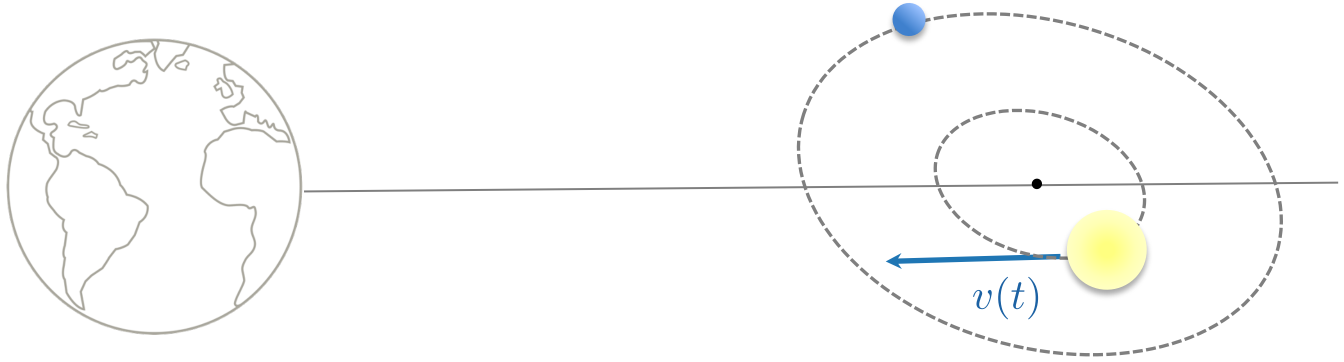




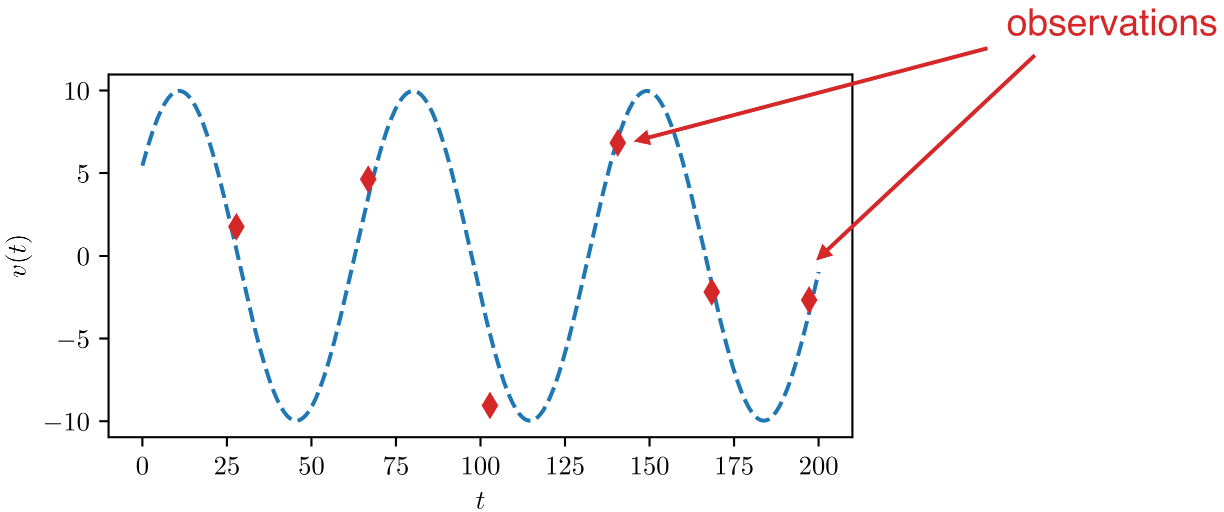
Captures precisely location of the transition as it avoids metastabilities

Bayesian inference:
An example of model selection from astrophysics

▷ Star-exoplanet system orbiting center of mass



▷ Radial velocity along the orbit $v(t;x) = v_0 + K \cos\left(\frac{2\pi}{P}t + \phi_0\right)$



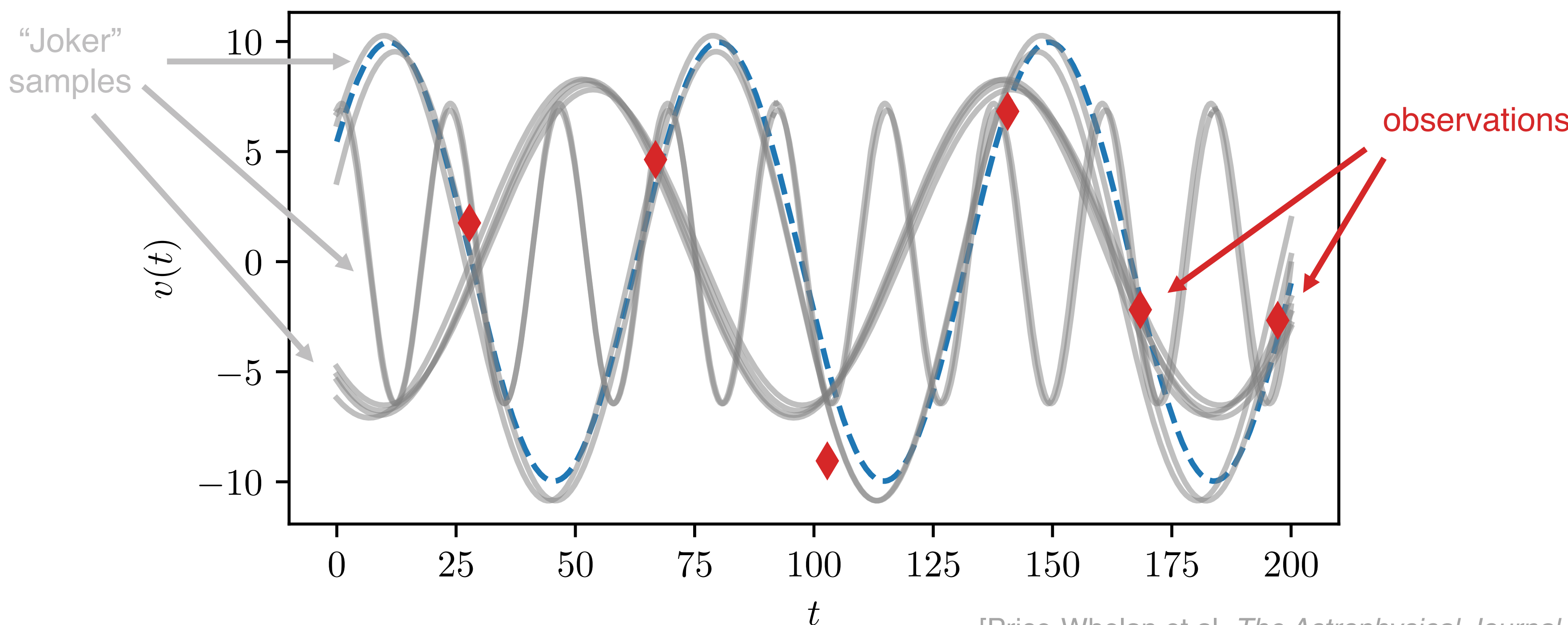
Bayesian model for velocity parameters

▷ Radial velocity $v(t; x) = v_0 + K \cos\left(\frac{2\pi}{P} t + \phi_0\right)$

▷ Parameters $x = (v_0, K, \phi_0, \ln P) \in \Omega \subset \mathrm{R}^4$

▷ Priors
$$\ln P \sim \mathcal{U}(\ln P_{\min}, \ln P_{\max}),$$
$$\phi_0 \sim \mathcal{U}(0, 2\pi),$$
$$K \sim \mathcal{N}(\mu_K, \sigma_K^2),$$
$$v_0 \sim \mathcal{N}(0, \sigma_{v_0}^2).$$

▷ Likelihood from observations $L(x) = \mathcal{N}(v_k; v(t_k; x), \sigma_{\text{obs}}^2)$

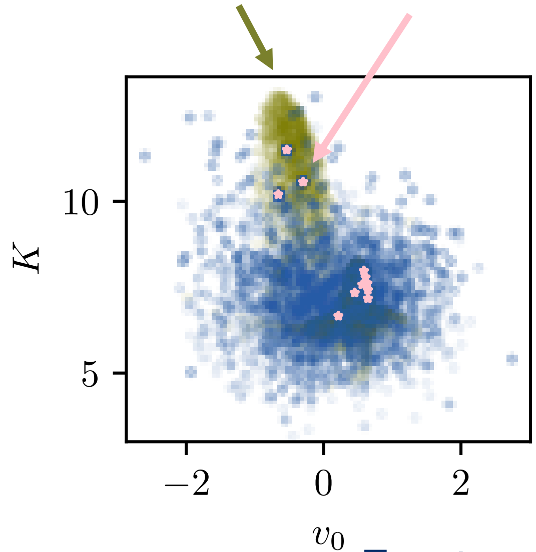


[Price-Whelan et al. *The Astrophysical Journal* 2017]

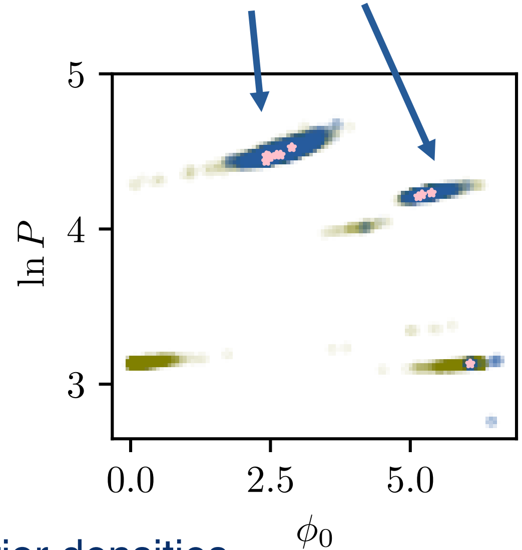
Sampling from the posterior

Joker samples

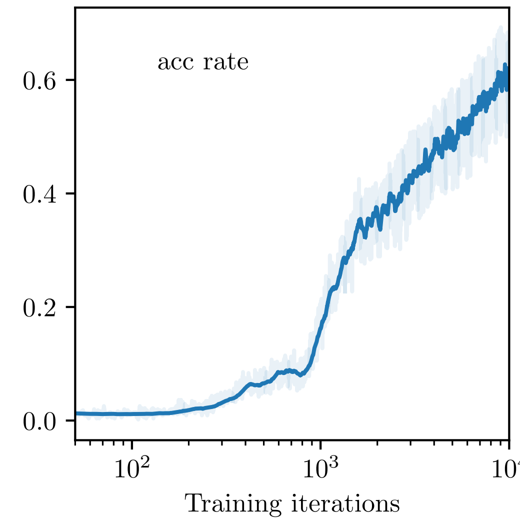
chain initilalizations



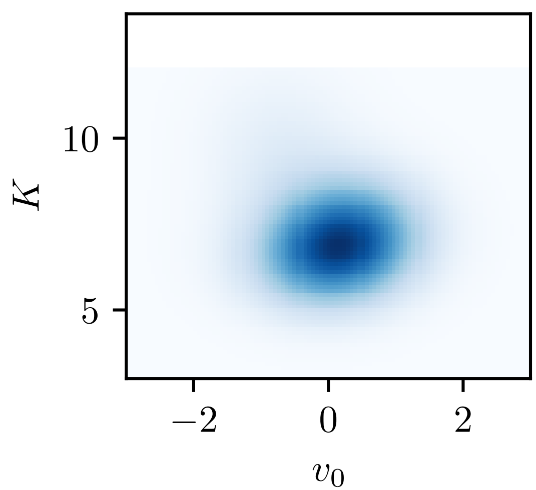
posterior samples from adaptive MCMC with NF

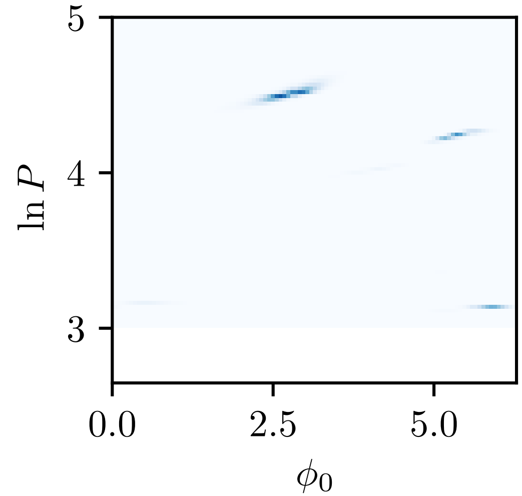


Acceptance along training

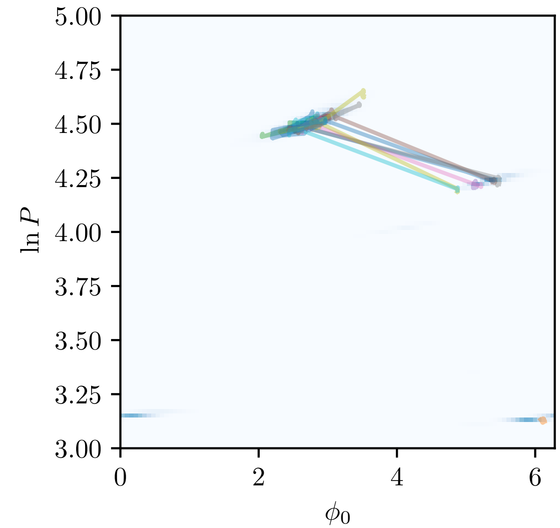


Exact posterior densities





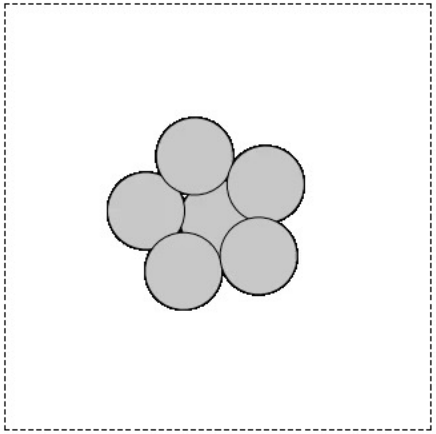
Fast mixing chains



[Gabrié, Rotskoff & Vanden-Eijnden - *ICML workshop on INNF+* - arXiv 2107.08001]

Perspectives

- ▷ Exciting applications ahead
  - Molecular dynamics
    Pilar Cossio (CCM/B, Flatiron Institute),
    Olga Acevedo & Ana Taborda (U. de Antioquia)



  - Bayesian Inference in Astrophysics
    Kaze Wong & David Foreman-Mackey (CCA, Flatiron Institute)

- ▷ An important take away: blending domain knowledge and learning is key!
  - cf Giulio's talk

▷Thank you!

Collaborators:
Grant Rotskoff (Stanford), Éric Vanden-Eijnden (Courant Institute, NYU)
James Brofos & Roy Lederman (Yale University), Marcus Brubacker (York university)
Pilar Cossio (CCM), Olga Lopez Acevedo & Ana Molina Taborda (Universidad de Antioquia)
Kaze Wong & Dan Foreman-Mackey (CCA)