Adaptive Monte Carlo Augmented with Normalizing Flows

On Future Synergies for Stochastic and Learning Algorithms

“Marseille”, Sep. 29, 2021

https://arxiv.org/abs/2107.08001
https://arxiv.org/abs/2105.12603

Grant M. Rotskoff (joint work with Marylou Gabrié and Eric Vanden-Eijnden)

https://statmech.stanford.edu

Design challenges in MCMC

Many problems in the physical sciences require sampling *high-dimensional, multimodal* distributions

Exponential timescales for transitions between known states, nontrivial to design MCMC to accelerate

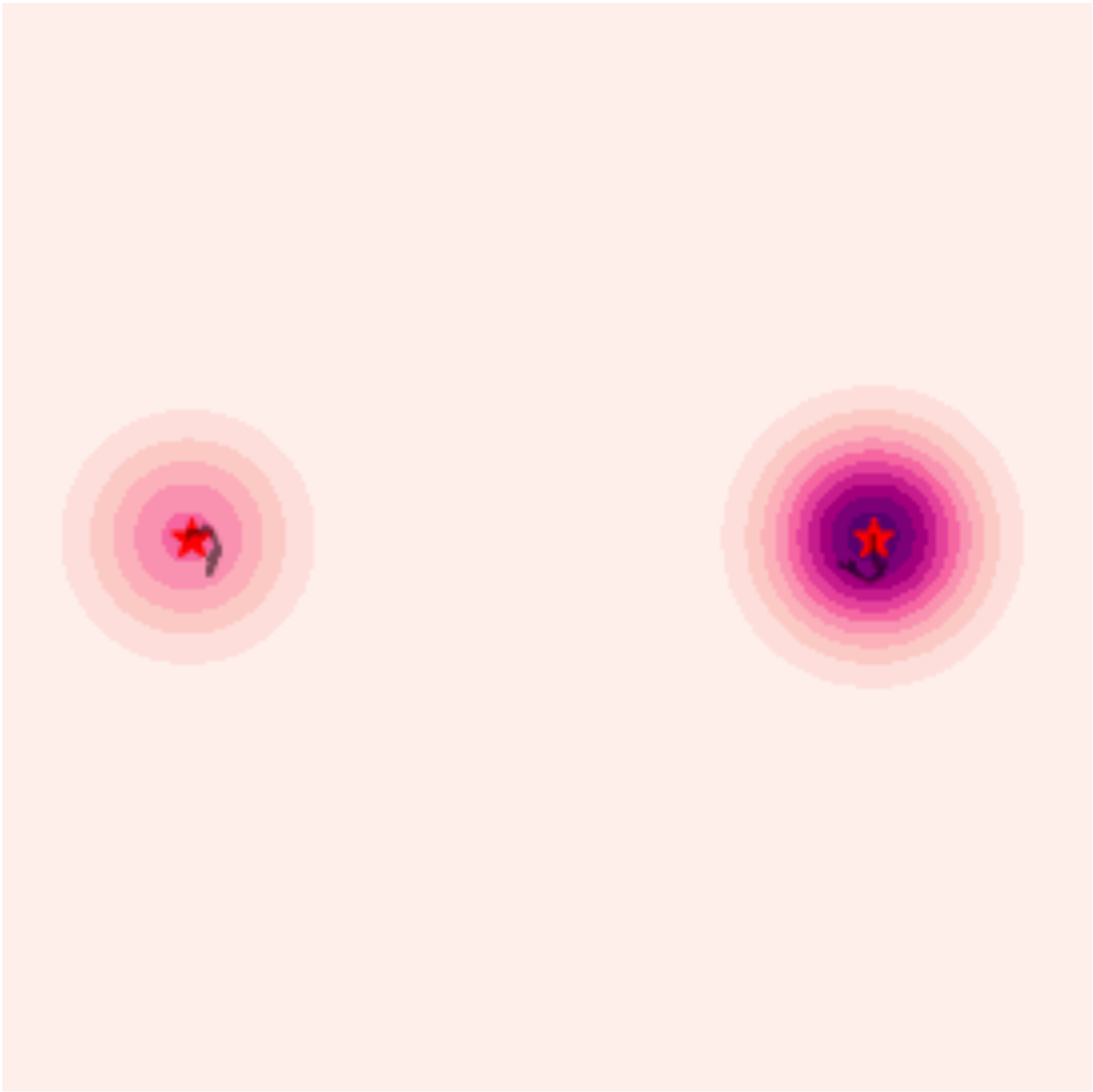
Specialized samplers are not transferrable between physically similar systems

Target distribution:

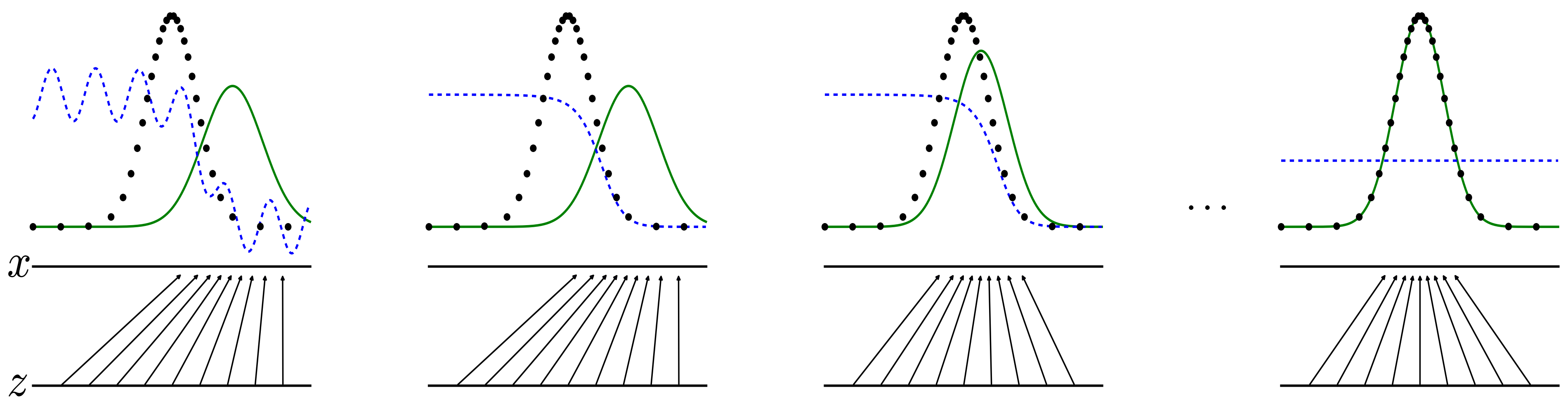
$$\rho_*(x) = Z_*^{-1} e^{-U_*(x)}$$

Typically design transition kernel with detailed balance:

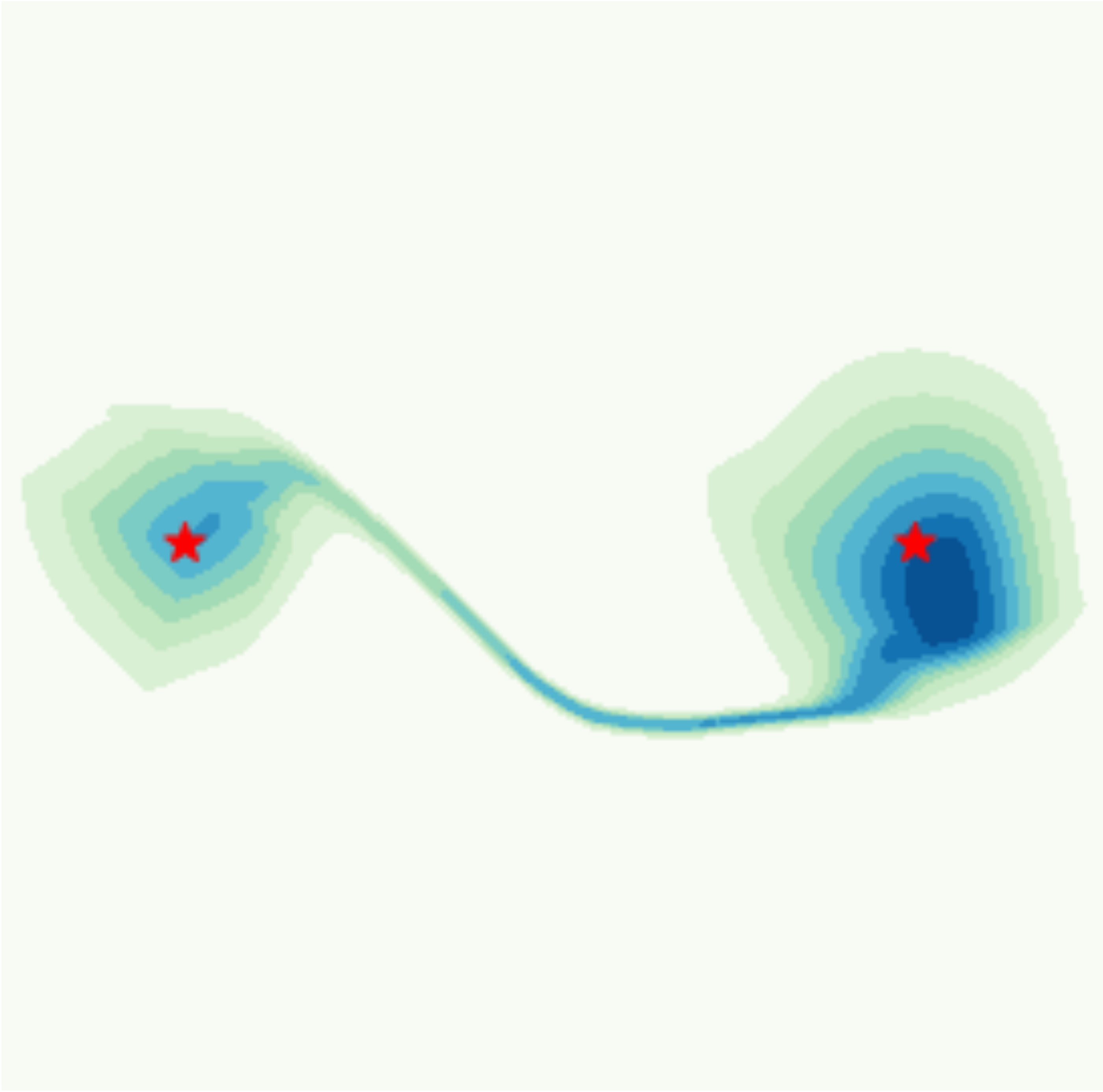
$$\rho_*(x)\pi(x,y) = \rho_*(y)\pi(y,x)$$



Generative models as MCMC samplers



Goodfellow et al., 2014.



Sample independently from learned distribution?

Requires *invertible* architecture
and (potentially) large amounts of data

Many works have considered this paradigm, e.g.,

2011: Andrieu, C.; Jasra, A.; Doucet, A.; Moral, P. D. On Nonlinear Markov Chain Monte Carlo. *Bernoulli* **2011**, *17* (3), 987–1014.

2016: Wang, D.; Liu, Q. Learning to Draw Samples: With Application to Amortized MLE for Generative Adversarial Learning. *arXiv:1611.01722 [cs, stat]* **2016**.

2017: Song, J.; Zhao, S.; Ermon, S. A-Nice-Mc: Adversarial Training for MCMC. In *Advances in neural information processing systems* 2017; Vol. 30.

2019: Albergo, M. S.; Kanwar, G.; Shanahan, P. E. Flow-Based Generative Models for Markov Chain Monte Carlo in Lattice Field Theory. *Phys. Rev. D* **2019**, *100* (3), 034515.

Normalizing flows; composable, invertible

Diffeomorphism (*flow*)

$$T : \mathbb{R}^d \to \mathbb{R}^d \qquad x = T(z)$$

$$\hat{\rho}(x) = \rho_{\mathrm{B}}(T^{-1}(x))\,|\det \nabla T^{-1}(x)\,|$$

“Base” measure — typically a Gaussian

In practice, architectures are built from compositions of many such maps:

$$x_k = T_k \circ \ldots \circ T_0(z)$$

We use “realNVP”

MCMC procedure is straightforward:

1. Generate a new configuration, accept/reject via MH

$$\mathrm{acc}(x, y) = \min\left[1, \frac{\hat{\rho}(x)\rho_*(y)}{\rho_*(x)\hat{\rho}(y)}\right]$$

2. *Optional* intercalate with local sampling (e.g., MALA)

$$\pi_T(x, y) = \mathrm{acc}(x, y)\hat{\rho}(y) + \big(1 - r(x)\big)\delta(x - y)$$

$$\hat{\pi}(x, y) = \int_\Omega \pi(x, z)\pi_T(z, y)dz$$

Local transition kernel

Tabak, E. G.; Vanden-Eijnden, E. Density Estimation by Dual Ascent of the Log-Likelihood. *Communications in Mathematical Sciences* **2010**, *8* (1), 217–233.
Rezende, D.; Mohamed, S. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning*; PMLR, 2015; pp 1530–1538.

Concurrent training and sampling

Sampling with an "imperfect" map T not optimal

Use the "forward" KL as a figure of merit:

$$D_{\mathrm{KL}}(\rho_* \| \hat{\rho}) = C_* - \int_\Omega \log \hat{\rho}(x) \rho_*(x) dx$$

"Self-training" uses the *reverse* KL

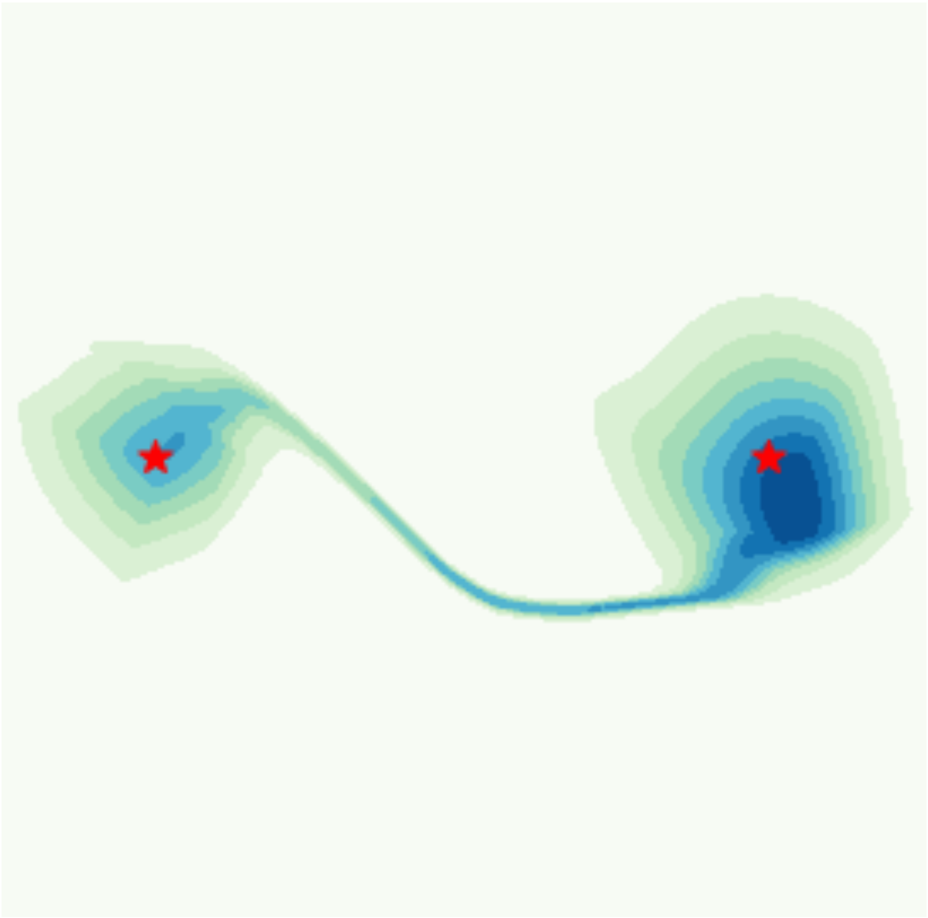
$$D_{\mathrm{KL}}(\hat{\rho} \| \rho_*) = \int_\Omega \log \frac{\hat{\rho}(x)}{\rho_*(x)} \hat{\rho}(x) dx$$

Mode collapse

$$\begin{aligned}\mathscr{L}_n[T] &= -\frac{1}{n}\sum_{i=1}^{n} \log \hat{\rho}(x_i(k)) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(U_{\mathrm{B}}(T^{-1}(x_i(k))) - \log\det|\nabla T^{-1}(x_i(k))| \right)\end{aligned}$$

Initialize with at least one walker in each metastable basin

Nonlocal jumps between basins key for acceleration



Continuous limit and convergence

$$g_t = \rho_t / \rho_* \qquad \hat{g}_t = \hat{\rho}_t / \rho_*$$

$$\partial_t g_t = -\nabla U_* \cdot \nabla g_t + \Delta g_t$$
$$+\alpha \int_\Omega \min(\hat{g}_t(x), \hat{g}_t(y)) \big(g_t(y) - g_t(x) \big) \rho_*(y) dy$$

Pearson χ^2 divergence

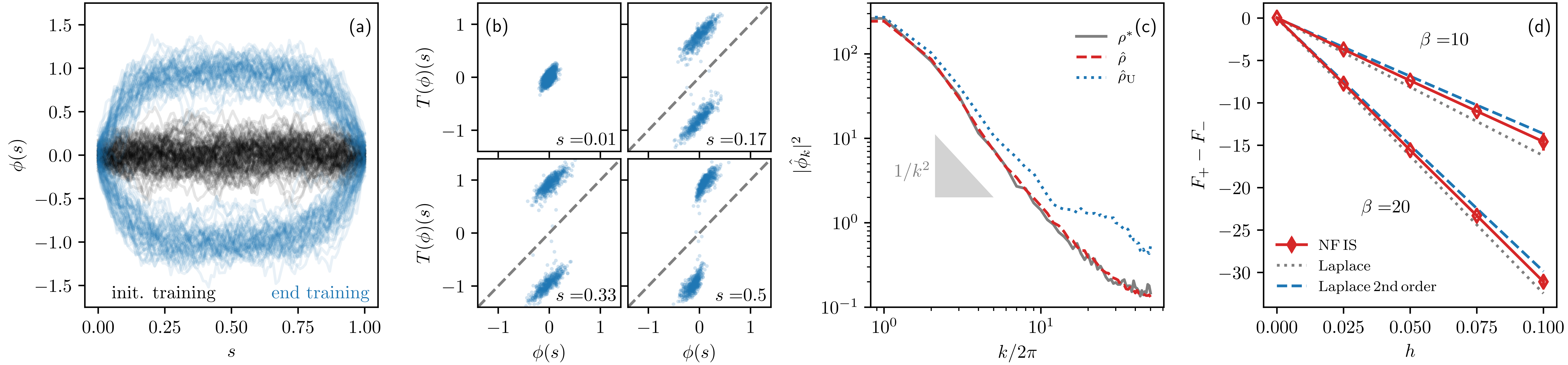
$$D_t = \int_\Omega \frac{\rho_t^2}{\rho_*} dx - 1 = \int_\Omega g_t^2 \rho_* dx - 1 \geq 0.$$

Convergence, provided initial distribution not "too far":

$$\forall t \geq t_0 \quad : \quad D_t \leq \frac{D_{t_0}}{\left(G_{t_0}(e^{\alpha(t-t_0)} - 1) + 1 \right)^2}$$

$$G_t = \inf_x g_t(x)$$

Base measure for rough paths

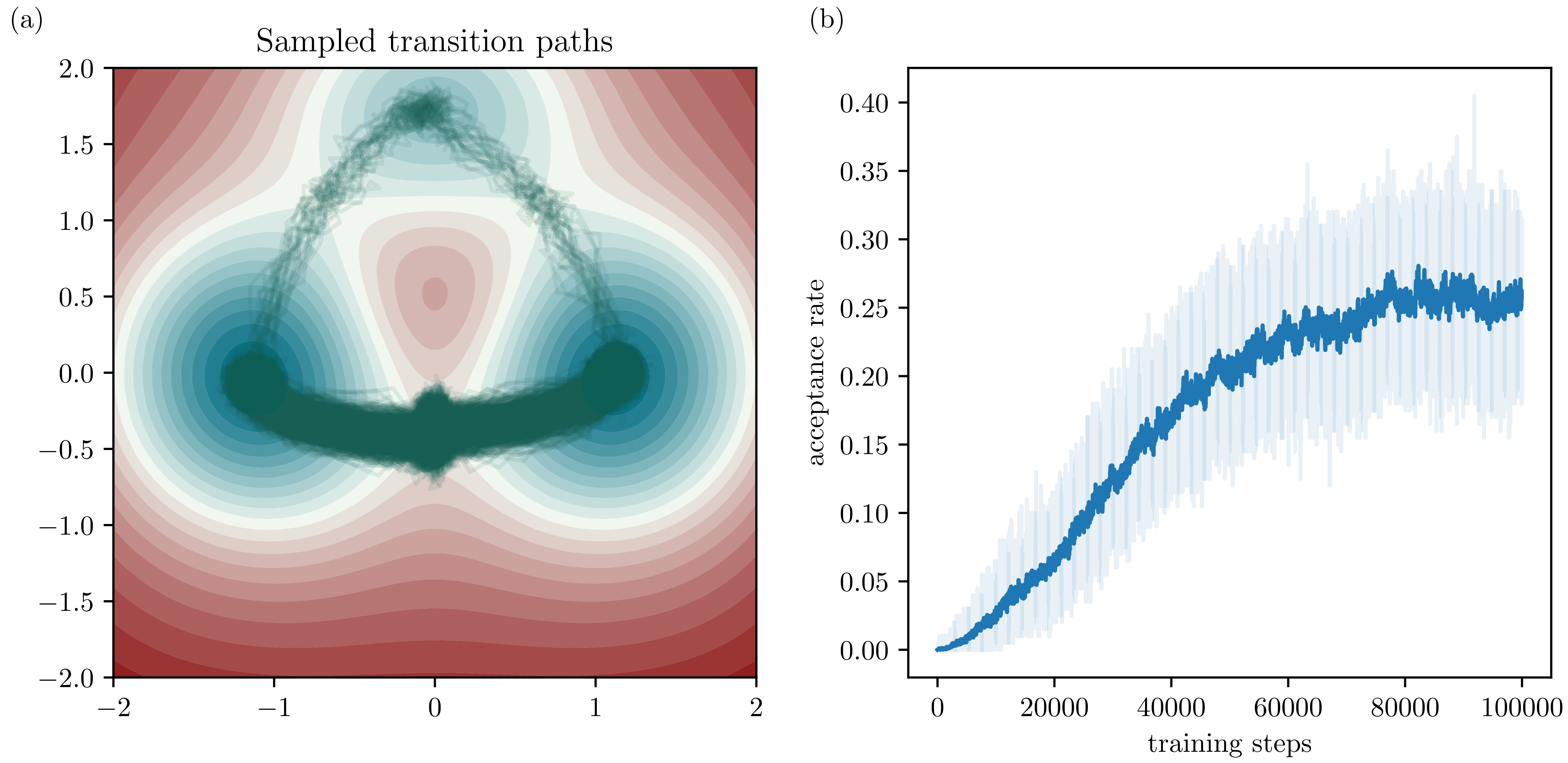


Stochastic Allen-Cahn model: $$\partial_t \phi = a \partial_s^2 \phi + a^{-1}(\phi - \phi)^3 + \sqrt{2\beta^{-1}}\, \eta(t, s)$$

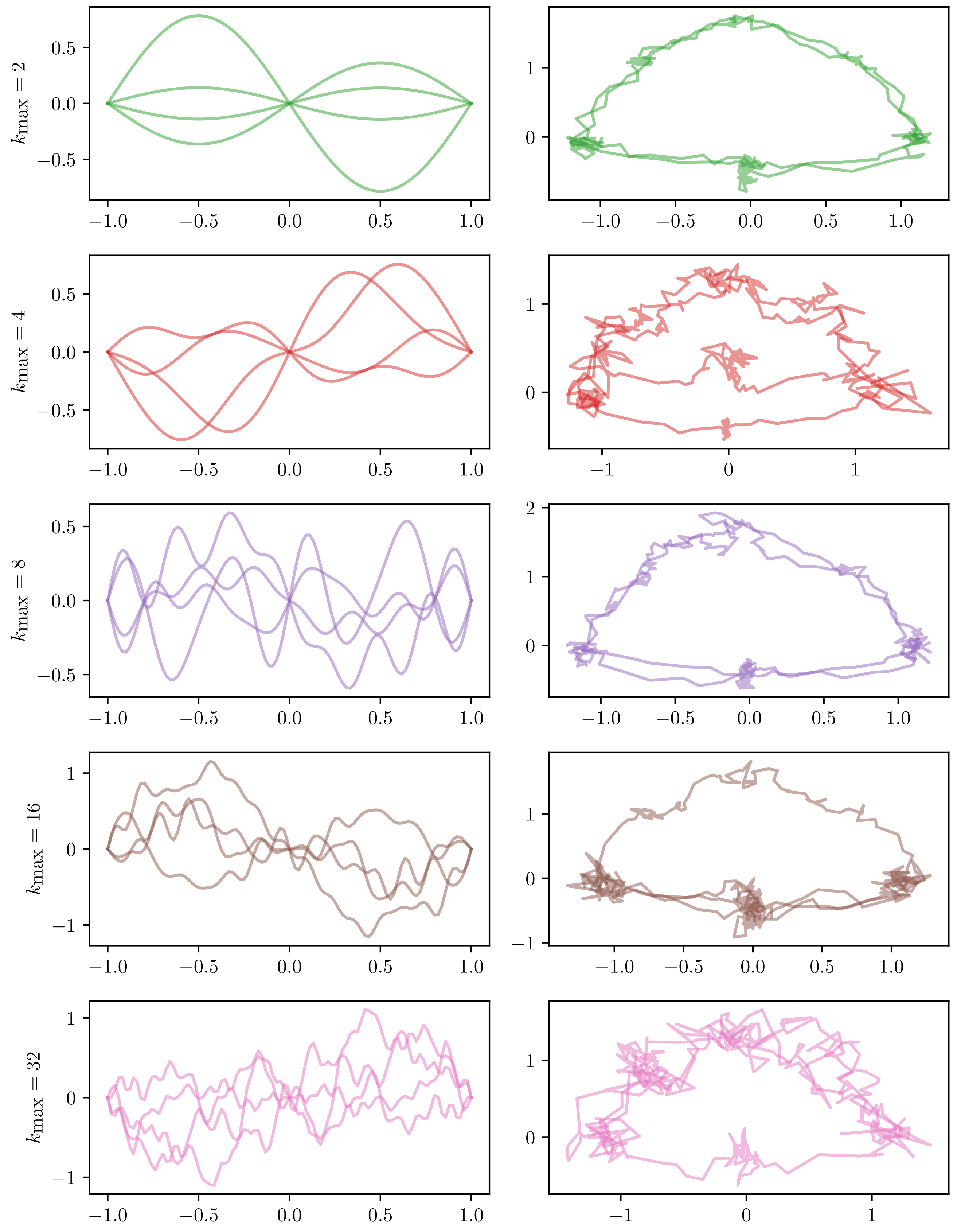
$$U_*[\phi] = \beta \int_0^1 \left[\frac{a}{2} (\partial_s \phi)^2 + \frac{1}{4a} \left(1 - \phi^2(s)\right)^2 \right] ds$$

$$U_\mathrm{B}[\phi] = \beta \int_0^1 \left[\frac{a}{2} (\partial_s \phi)^2 + \frac{1}{2a} \phi^2 \right] ds$$

Nonequilibrium Path Sampling



Karhunen-Loève shows *locality* and *smoothness*



$$\mathbb{P}_*(x_{[0,t_{\max}]}) \propto \exp\left[-\frac{\beta}{4}\int_0^{t_{\max}} |\dot{x}_t - b(x_t)|^2\, dt\right]$$

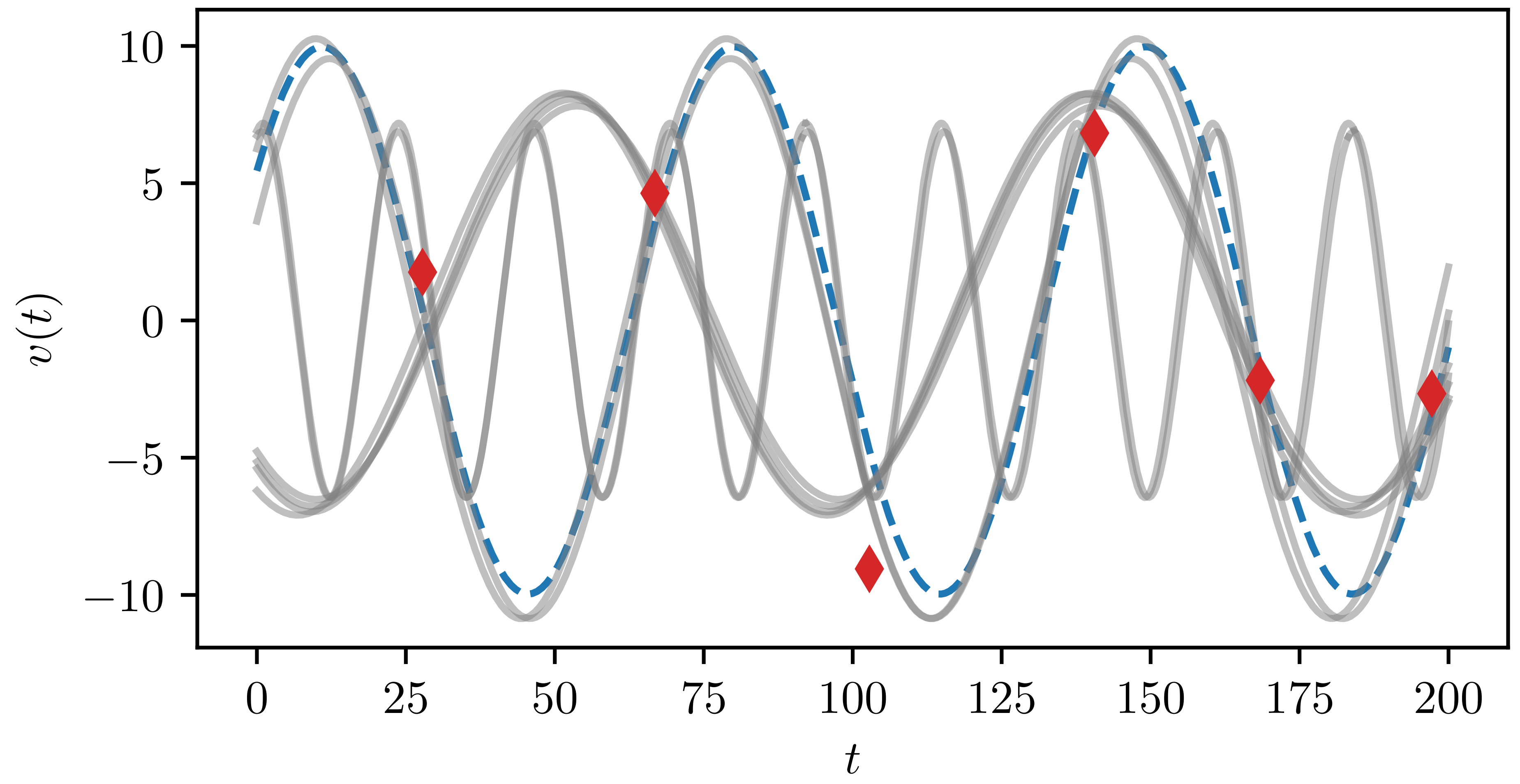
Brownian bridge base measure:

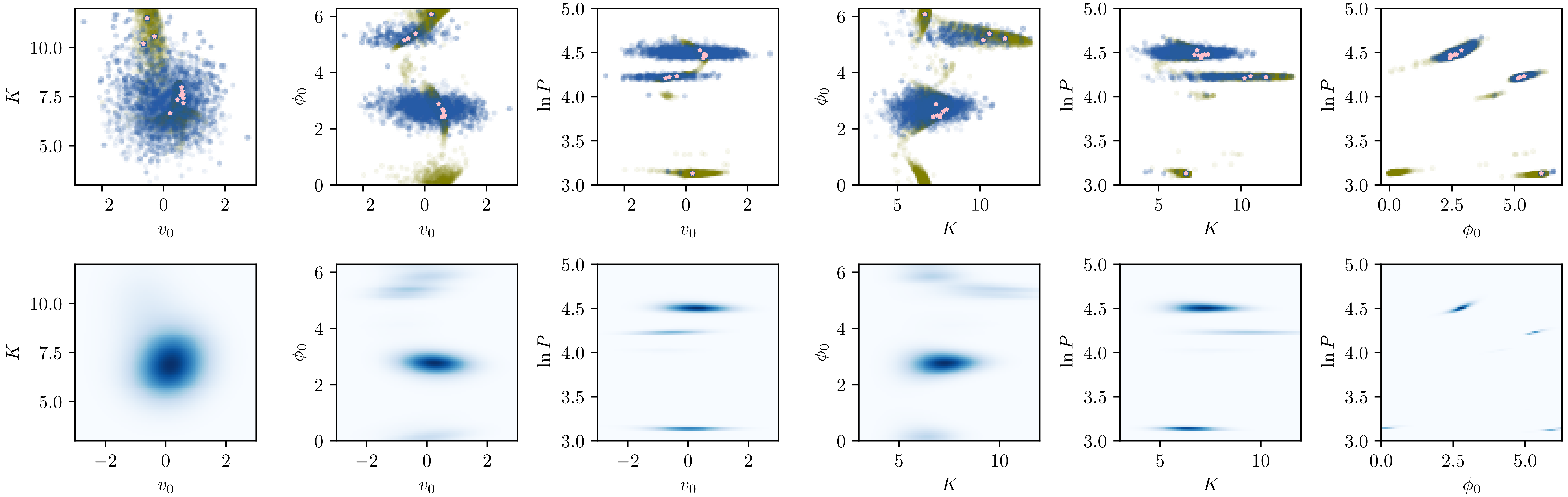
$$\mathbb{P}_{\mathrm{B}}(x_{[0,t_{\max}]}) \propto \exp\left[-\frac{\beta}{4}\int_0^{t_{\max}} |\dot{x}_t|^2\, dt\right]$$

Bayesian sampling

$$\rho_*(\theta) = \rho(\theta\,|\,D)\rho_o(\theta) = Z_*^{-1}L(\theta)\rho_o(\theta)$$

$$Z_* = \int_\Theta L(\theta)\rho_o(\theta)d\theta \qquad Z_* = \mathbb{E}_{\rho_B}\left[\frac{L(T(\theta_B))\rho_o(T(\theta_B))}{\hat{\rho}(T(\theta_B))}\right]$$





Conclusions

- Sampling and training NFs to augment MCMC yields non-local transport
- Flexible, generalizable, even in high-dimensions
- Training is non-trivial, but local dynamics helps explore basins
- *Not a method for discovery* (at least, not yet)
- Good convergence properties, provided an appropriate base measure
- Much work to be done to adapt base measures in cases where no *a priori* data exists

Thanks!