Will be related to Lorenz's talk! arxiv:2407.07873

NETS

A Non-Equilibrium Transport Sampler

aka an Answer to Tej's question of the connection between flows and diffusions for sampling

aka a continuous time algorithm for what Alessandro is doing

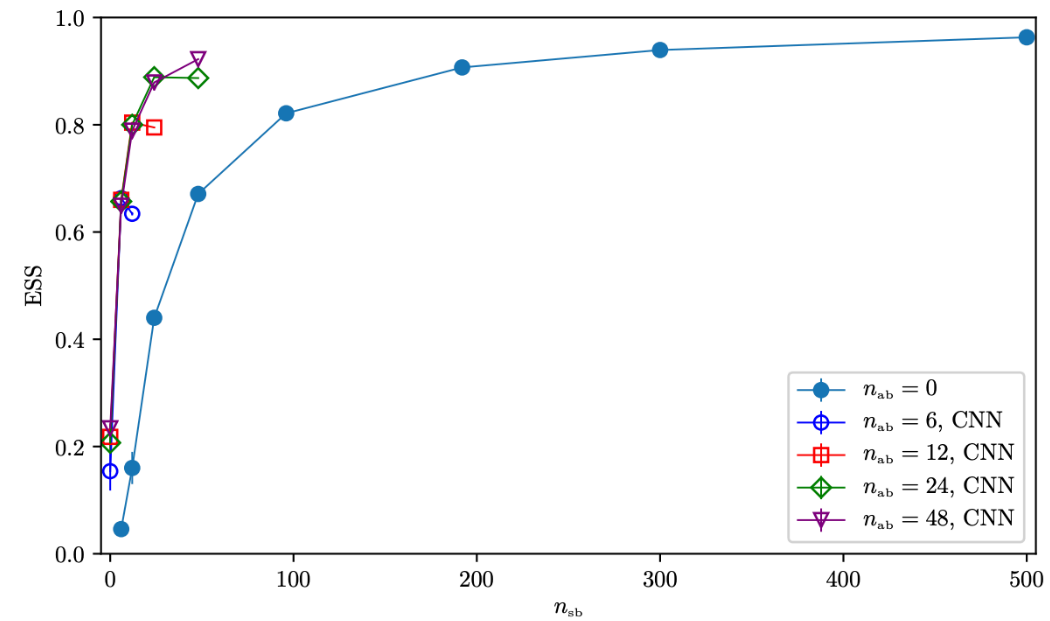
Michael Albergo **Bonn, Germany** **October 24 2024**

Remember this from yesterday?

Stochastic normalizing flows as non-equilibrium transformations

Michele Caselle[1,2,*] Elia Cellini[1,2,†] Alessandro Nada[1‡] and Marco Panero[1,2§]

Improved ESS by growing the # of discrete affine flows + stochastic steps



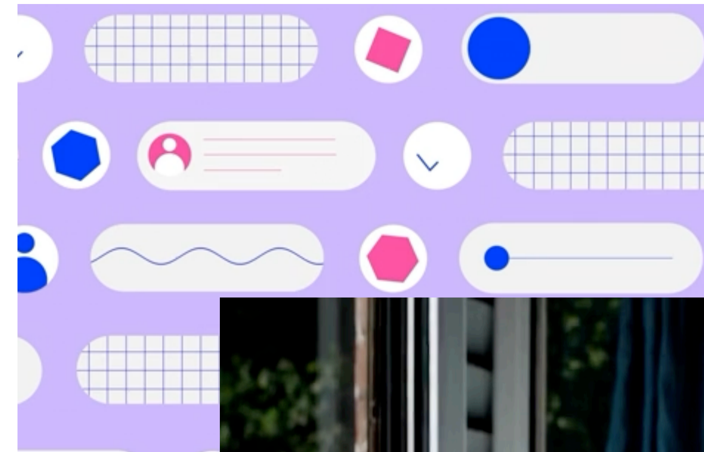
NETS is a continuous time limit of SNFs

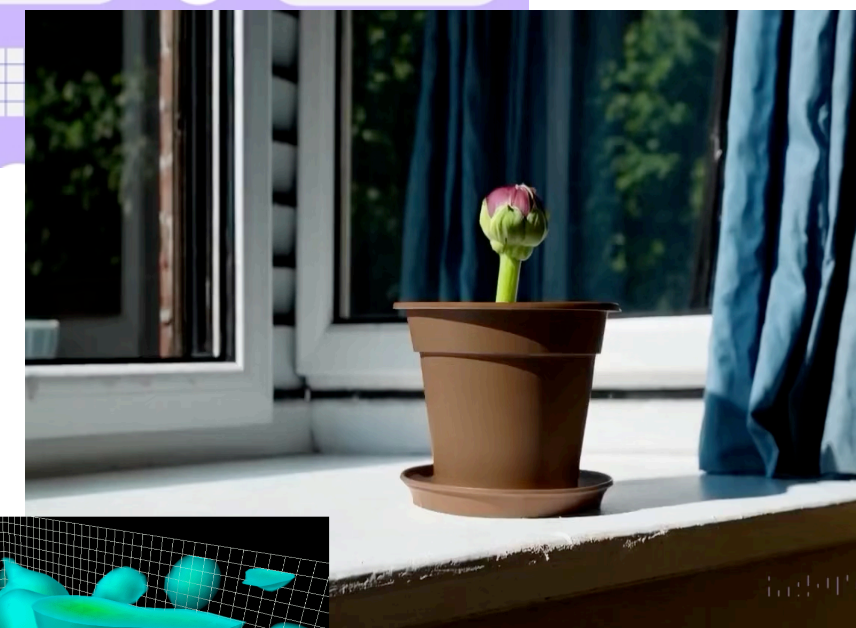
- Can choose how many steps + diffusion ***after training***
- Knob to explicitly get more performance from more compute

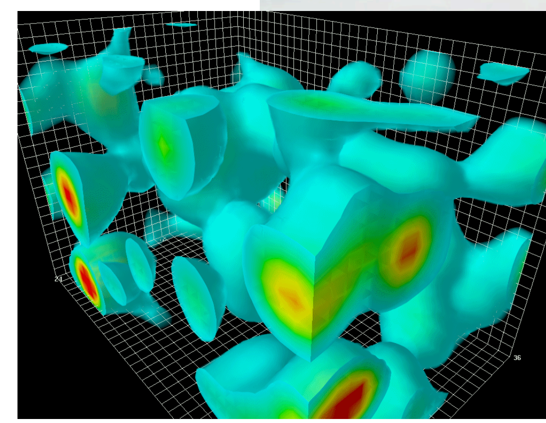
Advertisement: New Research group

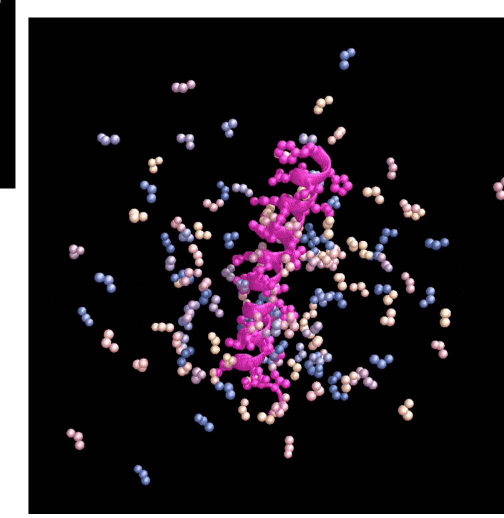
In 2026 I will be starting a group at Harvard in Applied mathematics + Kempner Institute

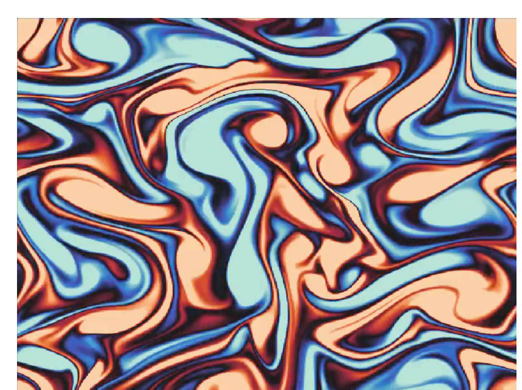
- *Theme: Nature and Computation*
- *Interdisciplinary! Computationally inclined, mathematically inclined welcome*
- *Current undergraduates, master's students, graduating PhDs, and postdocs, please reach out if interested*
- *Advisors, please forward your students :)*











Kempner INSTITUTE

HARVARD UNIVERSITY

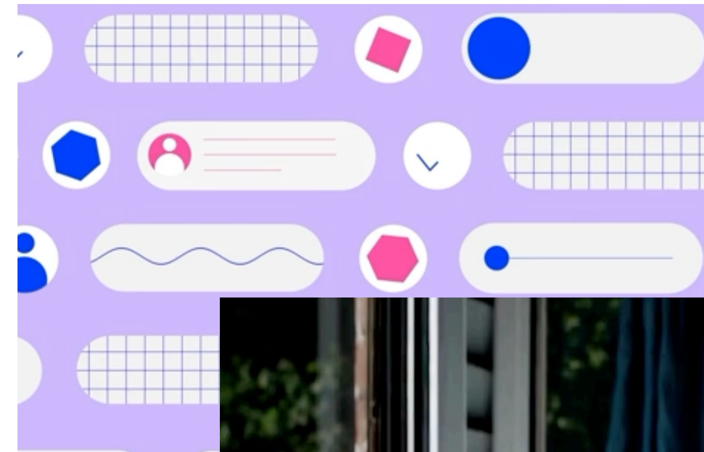
Advertisement: New Research group

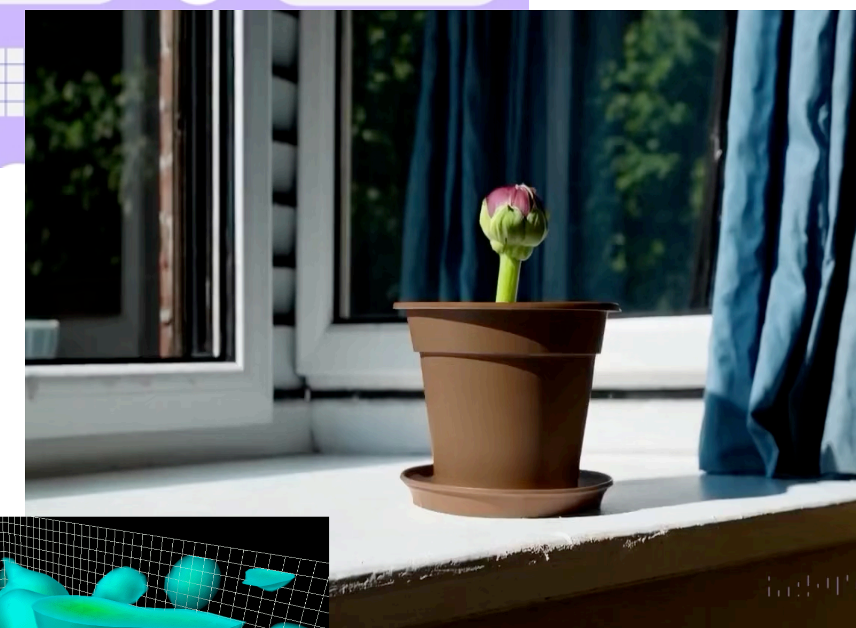
In 2026 I will be starting a group at Harvard in Applied mathematics + Kempner Institute

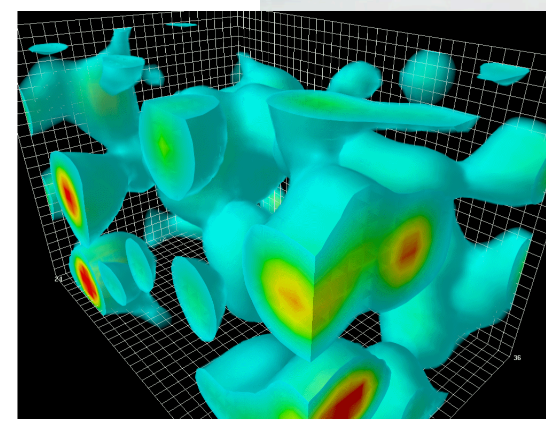
- *Theme: Nature and Computation*
- *Interdisciplinary! Computationally inclined, mathematically inclined welcome*
- *Current undergraduates, master's students, graduating PhDs, and postdocs, please reach out if interested*
- *Advisors, please forward your students :)*

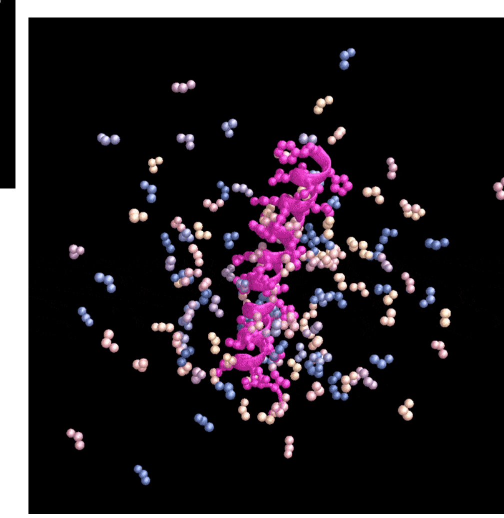
Kempner INSTITUTE

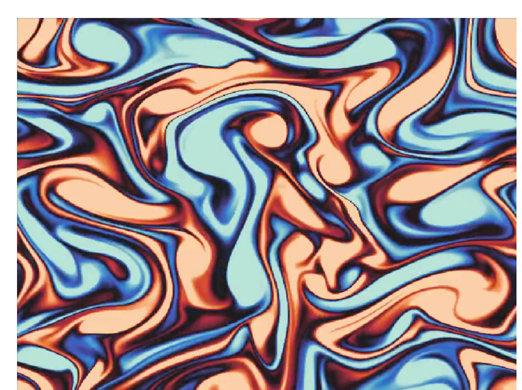
HARVARD UNIVERSITY











Agenda

Annealed Importance Sampling and Jarzynski's equality

Problem statement

Much related work!

Dynamical Measure Transport

Recent methods for learning maps between distributions

Combining the two!

New learning algorithms

Applications, e.g. field theory

Agenda

Annealed Importance Sampling and Jarzynski's equality

Problem stat

Dynamical Measur

Recent metho

Main motivation for this work:

Can we explicitly get a machine learning-augmented sampling setup for which "when I pay more from using my model, I get more from my model"?

Combining the two!

New learning algorithms

Applications, e.g. field theory

Agenda

Annealed Importance Sampling and Jarzynski's equality

Problem stat

Dynamical Measur

Recent metho

Joint work with Eric Vanden-Eijnden

Combining the two!

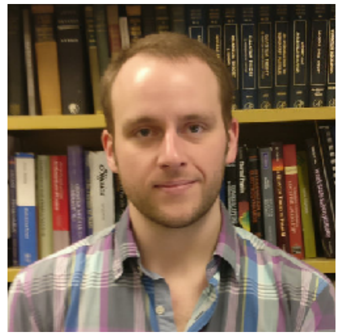
New learning algorithms

Applications, e.g. field theory

Thanks to all collaborators!

Massachusetts Institute of Technology


P. Shanahan


D. Hackett


F. Romero-


R. Abbott


D. Boyda

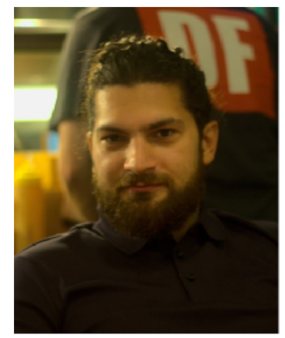

J. Urban

DeepMind


D. Rezende


S. Racanière


A. Razavi


A. Botev

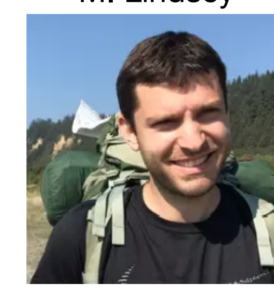

M. Lindsey

u^b

UNIVERSITÄT BERN

AEC ALBERT EINSTEIN CENTER FOR FUNDAMENTAL PHYSICS


G. Kanwar

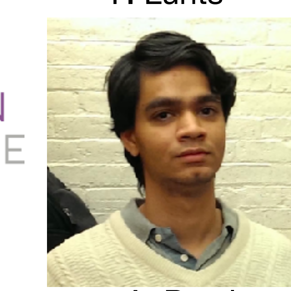

P. Lunts

NEW YORK UNIVERSITY


K. Cranmer


N. Boffi

FLATIRON INSTITUTE Center for Computational Quantum Physics


A. Patel

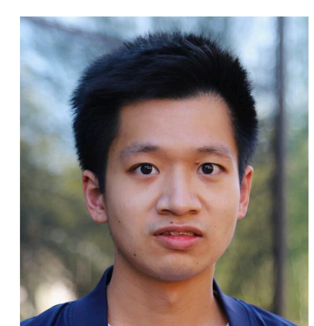

M. Goldstein


R. Ranganath


E. Vanden-Eijnden


Y. LeCun


S. Xie


Y. Chen

Problem Setup

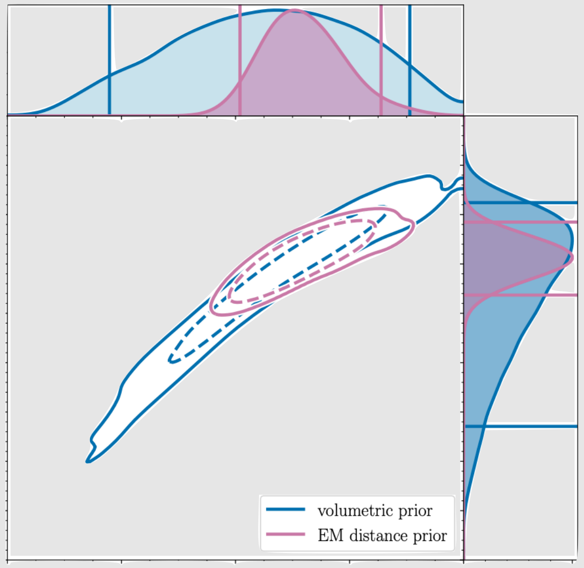
Goal: estimate the unknown *probability density function* $\rho_1 \in \mathcal{D}(\Omega)$ either through:

1. sample data $\{x_i\}_{i=1}^n$
2. **query access to the unnormalized log likelihood (energy function)**

Sampling problem ubiquitous!

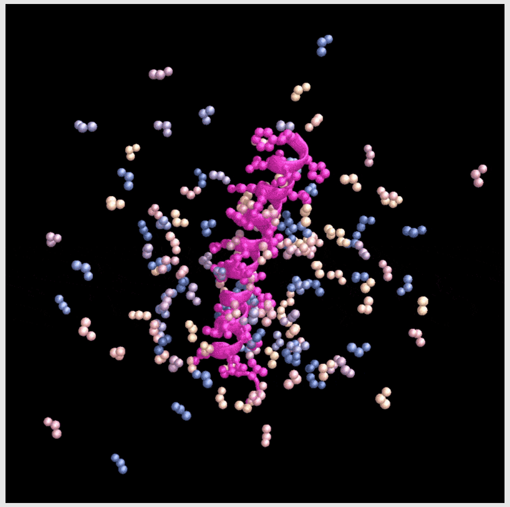
(obviously, to this audience)

energy function $U_1(x)$

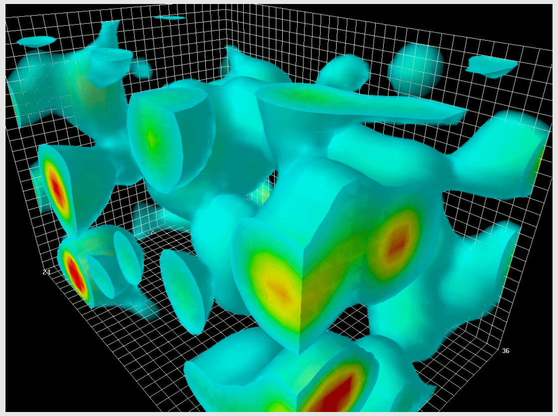


bayesian inference in GW astronomy

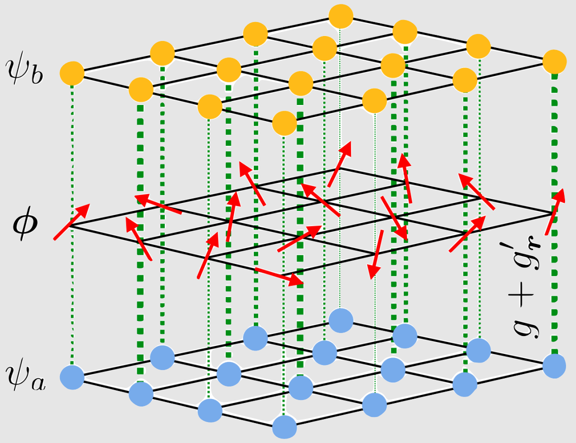
1809.02293



MD simulations



quantum field theory



condensed matter

Problem Setup

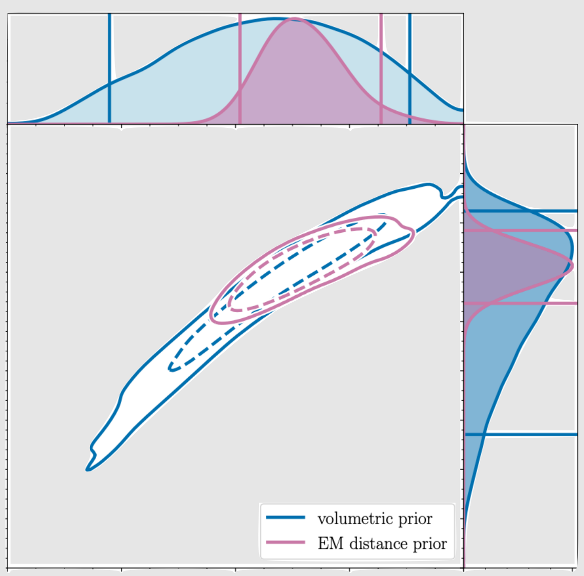
Goal: estimate the unknown *probability density function* $\rho_1 \in \mathscr{D}(\Omega)$ either through:

1. sample data $\{x_i\}_{i=1}^n$
2. **query access to the unnormalized log likelihood (energy function)**

Sampling problem ubiquitous!

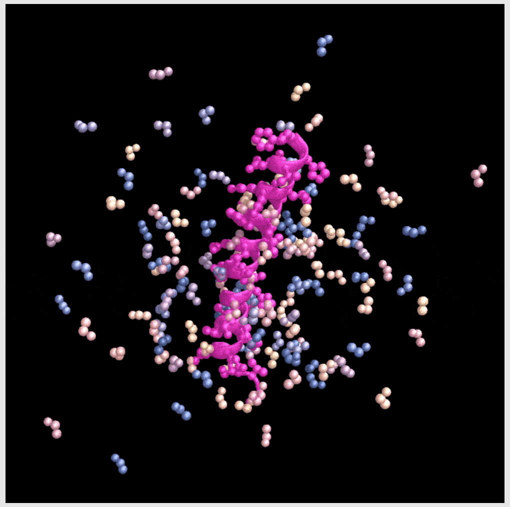
(obviously, to this audience)

energy function $U_1(x)$

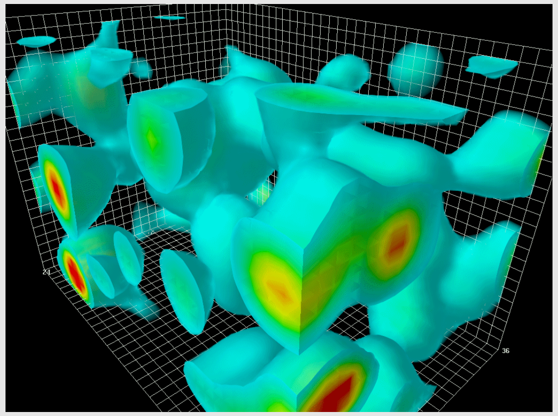


bayesian inference in GW astronomy

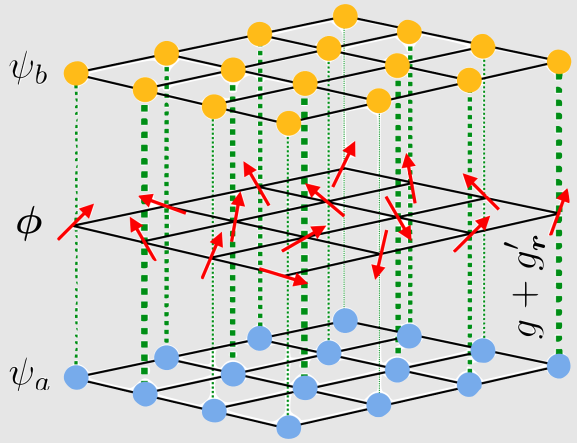
1809.02293



MD simulations



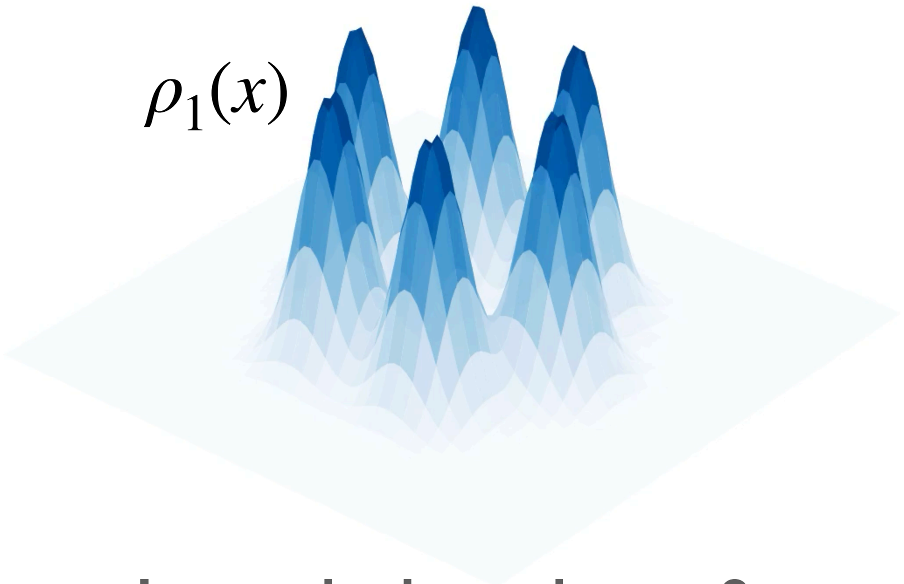
quantum field theory



condensed matter

Approaches to sampling

Markov Chain Monte Carlo build randomized sequence of samples $\{x_i\}_{i=1}^N$ so that



Langevin dynamics on 2-dimensional distribution

$$\lim_{N\to\infty} \mathbb{E}[h(x)]_N \to \mathbb{E}[h(x)]$$

Common tool: Langevin Dynamics

$$dX_t = -\epsilon \nabla U_1(X_t)dt + \sqrt{2\epsilon}dW_t$$

gradient drift

incremental brownian motion

Importance Sampling

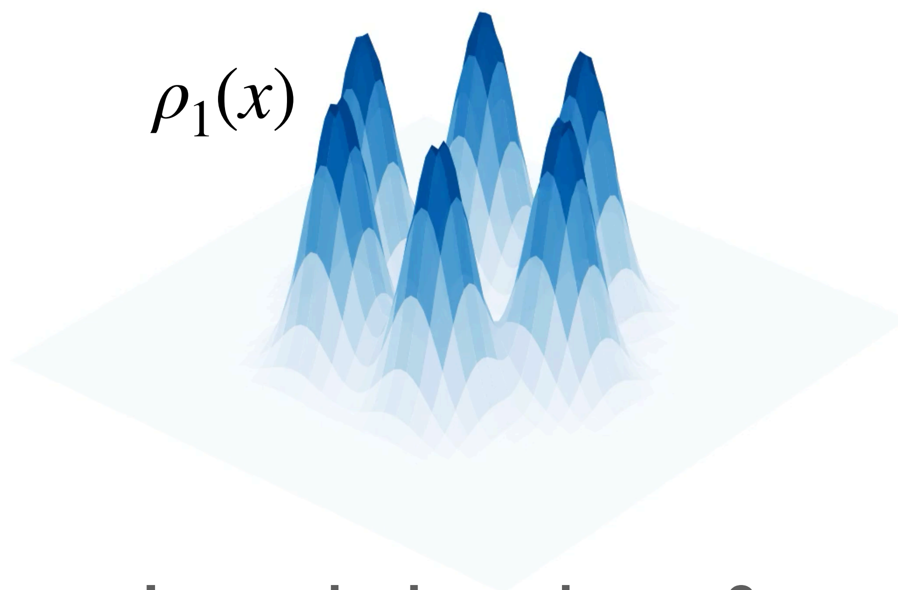
Re-weight samples from cheap surrogate model

$$\mathbb{E}_{\rho_1}[h(x)] = \mathbb{E}_{\hat{\rho}_1}\left[h(x)\frac{\rho_1(x)}{\hat{\rho}_1(x)}\right]$$

Effective when $\rho_1, \hat{\rho}_1$ overlap

Approaches to sampling

Markov Chain Monte Carlo build randomized sequence of samples $\{x_i\}_{i=1}^N$ so that



Langevin dynamics on 2-dimensional distribution

$$\lim_{N\to\infty} \mathbb{E}[h(x)]_N \to \mathbb{E}[h(x)]$$

Common tool: Langevin Dynamics

$$dX_t = -\epsilon \nabla U_1(X_t)dt + \sqrt{2\epsilon}dW_t$$

gradient drift

incremental brownian motion

Importance Sampling

Re-weight samples from cheap surrogate model

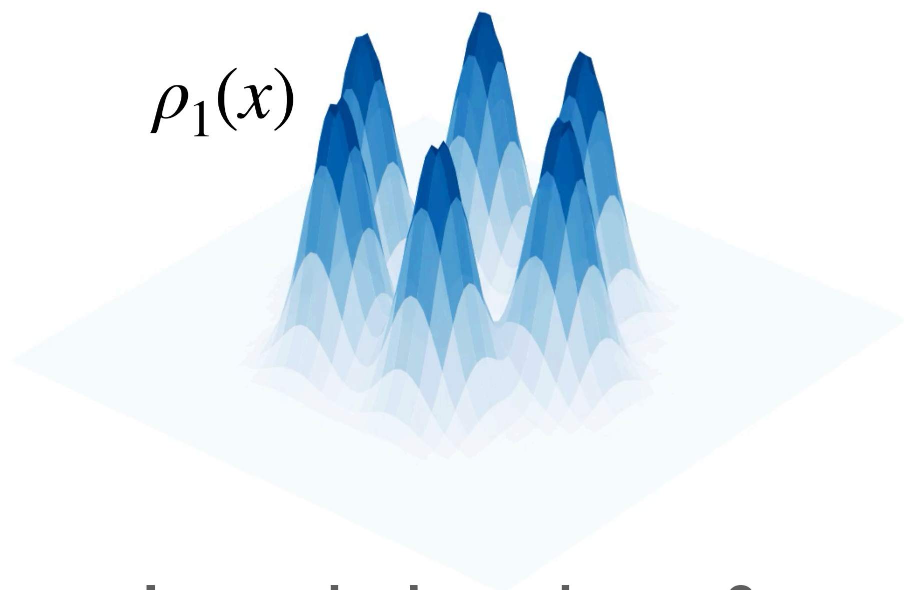
$$\mathbb{E}_{\rho_1}[h(x)] = \mathbb{E}_{\hat{\rho}_1}\left[h(x)\frac{\rho_1(x)}{\hat{\rho}_1(x)}\right]$$

Effective when $\rho_1, \hat{\rho}_1$ overlap

Approaches to sampling

Markov Chain Monte Carlo build randomized sequence of samples $\{x_i\}_{i=1}^{N}$ so that

$\rho_1(x)$



Langevin dynamics on 2-dimensional distribution

$$\lim_{N\to\infty} \mathbb{E}[h(x)]_N \to \mathbb{E}[h(x)]$$

Common tool: Langevin Dynamics

$$dX_t = -\epsilon \nabla U_1(X_t)dt + \sqrt{2\epsilon}\, dW_t$$

gradient drift

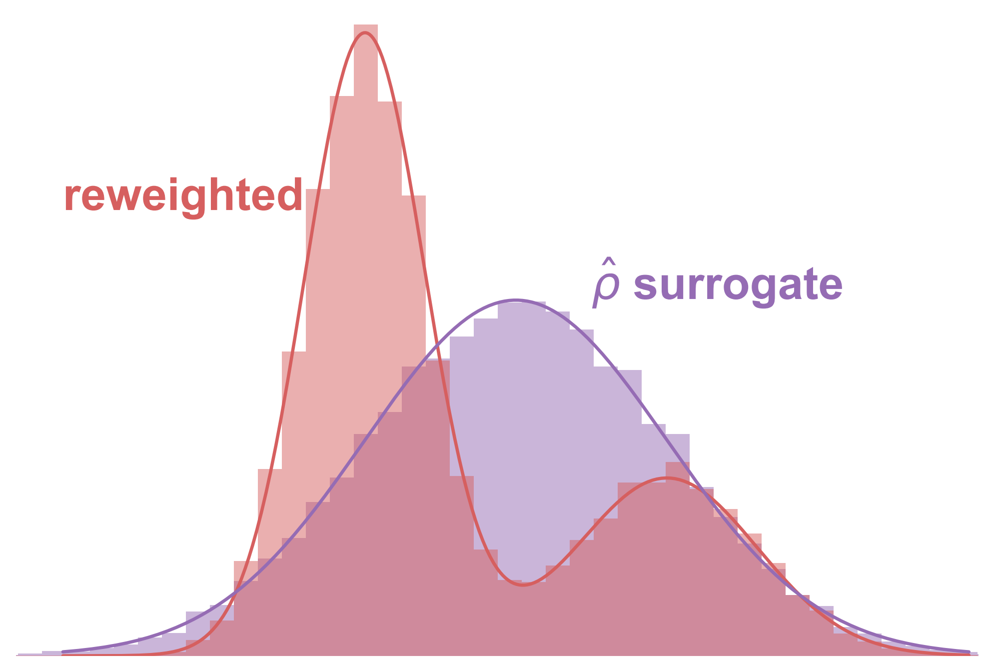
incremental brownian motion

Importance Sampling

Re-weight samples from cheap surrogate model

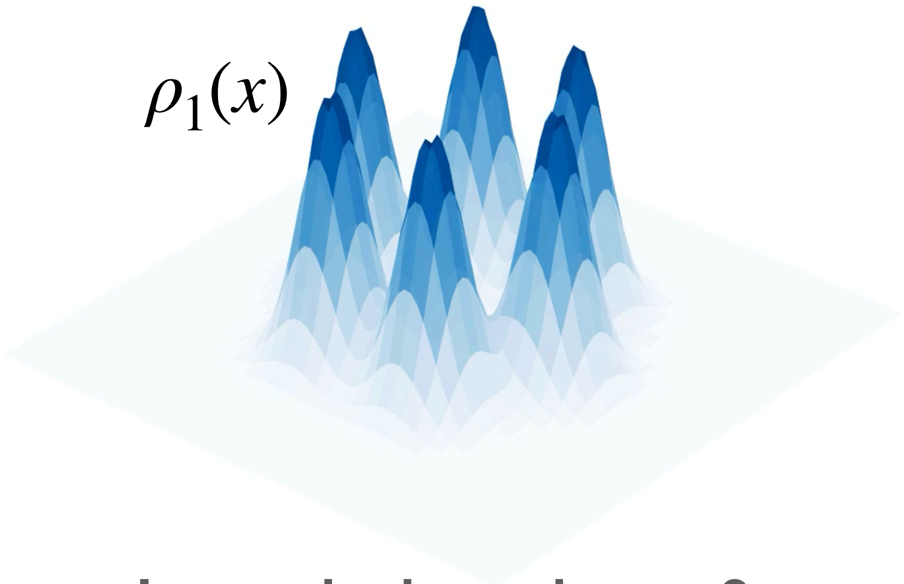
$$\mathbb{E}_{\rho_1}[h(x)] = \mathbb{E}_{\hat{\rho}_1}\left[h(x)\frac{\rho_1(x)}{\hat{\rho}_1(x)}\right]$$

Effective when $\rho_1, \hat{\rho}_1$ overlap



Approaches to sampling

Markov Chain Monte Carlo build randomized sequence of samples $\{x_i\}_{i=1}^{N}$ so that



Langevin dynamics on 2-dimensional distribution

$$\lim_{N\to\infty} \mathbb{E}[h(x)]_N \to \mathbb{E}[h(x)]$$

Common tool: Langevin Dynamics

$$dX_t = -\epsilon \nabla U_1(X_t)dt + \sqrt{2\epsilon}\, dW_t$$

gradient drift

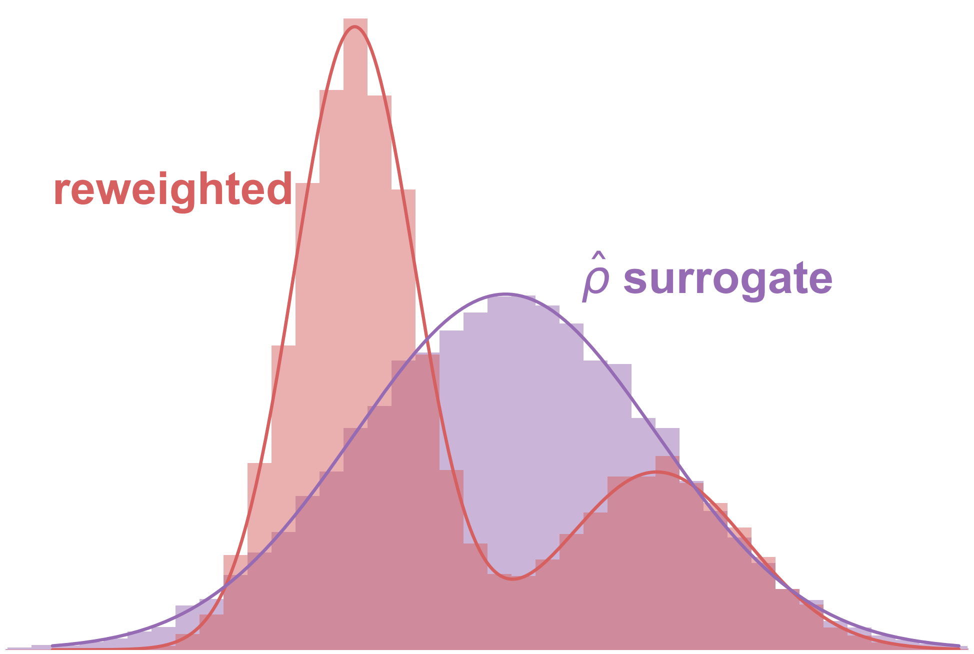
incremental brownian motion

Importance Sampling

Re-weight samples from cheap surrogate model

$$\mathbb{E}_{\rho_1}[h(x)] = \mathbb{E}_{\hat{\rho}_1}\left[h(x)\frac{\rho_1(x)}{\hat{\rho}_1(x)}\right]$$

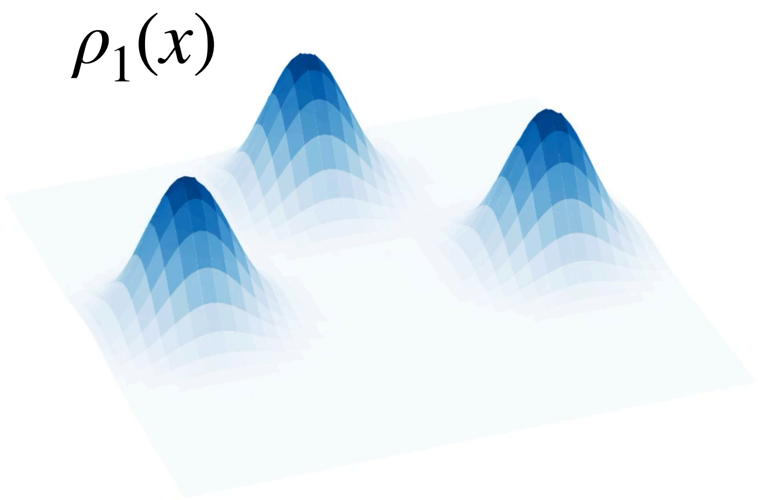
Effective when $\rho_1, \hat{\rho}_1$ *overlap*



Limitations of MCMC and IS

Markov Chain Monte Carlo

build randomized sequence of samples $\{x_i\}_{i=1}^N$ so that



Non-log concave target, exponentially slow mixing

Convergence can be exponentially slow

Common tool: Langevin Dynamics

$$dX_t = -\epsilon \nabla U_1(X_t)dt + \sqrt{2\epsilon} dW_t$$

gradient drift

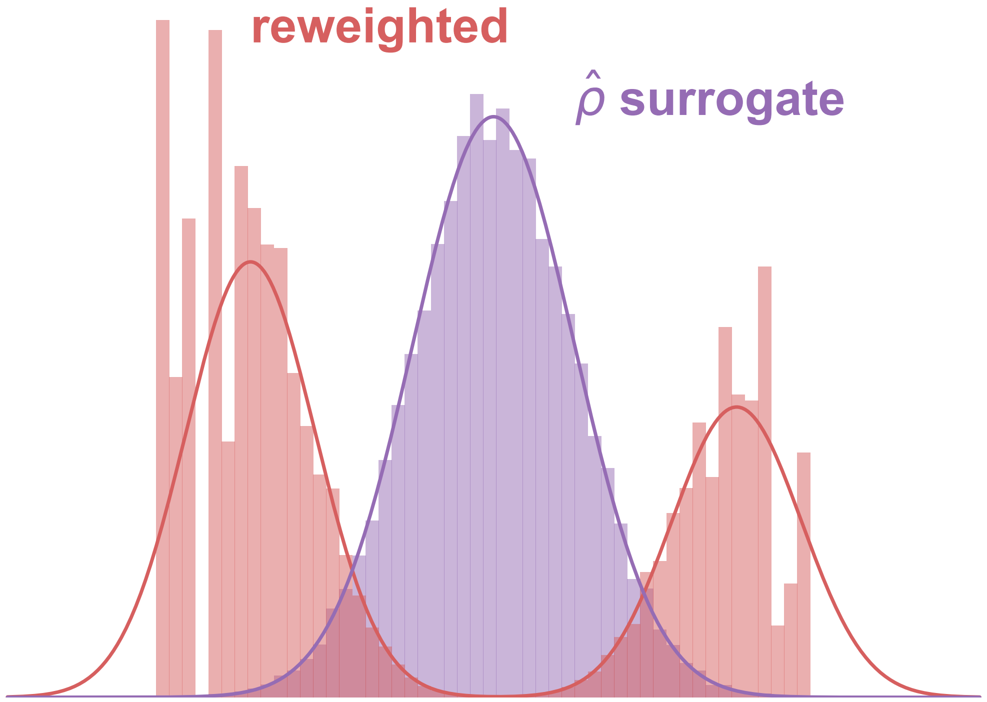
incremental brownian motion

Importance Sampling

Re-weight samples from cheap surrogate model

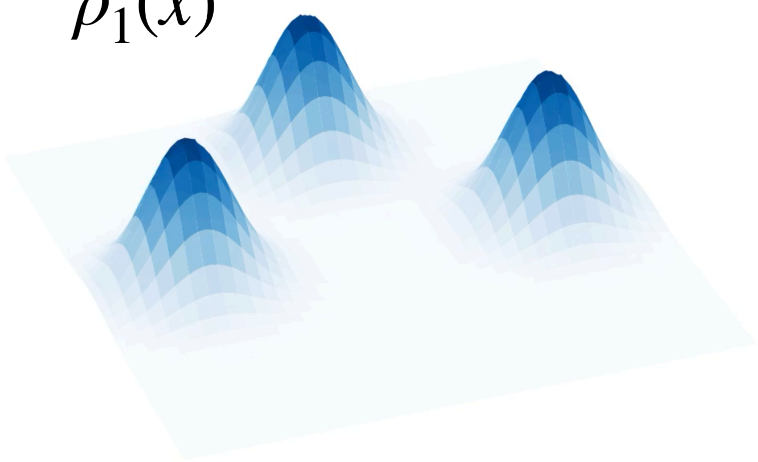
$$\mathbb{E}_{\rho_1}[h(x)] = \mathbb{E}_{\hat{\rho}_1}\left[h(x)\frac{\rho_1(x)}{\hat{\rho}_1(x)}\right]$$

Variance can be exponentially bad, especially in high dimension



Limitations of MCMC and IS

Markov Chain Monte Carlo build randomized sequence of samples $\{x_i\}_{i=1}^N$ so that



Non-log concave target, exponentially slow mixing

Convergence can be exponentially slow

Common tool: Langevin Dynamics

$$dX_t = -\epsilon \nabla U_1(X_t)dt + \sqrt{2\epsilon}dW_t$$

gradient drift

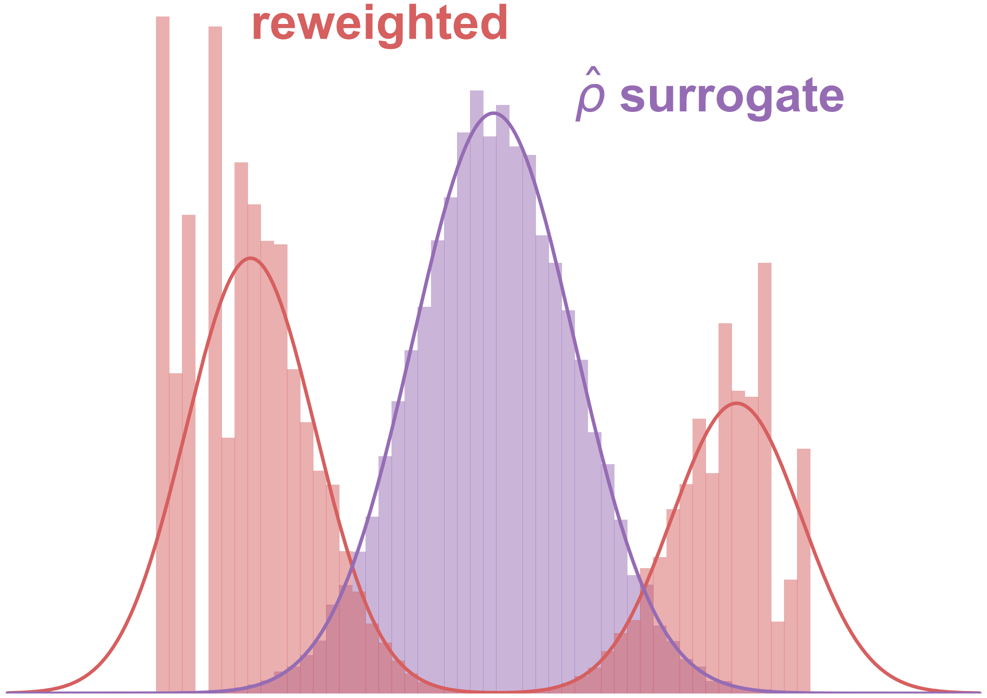
incremental brownian motion

Importance Sampling

Re-weight samples from cheap surrogate model

$$\mathbb{E}_{\rho_1}[h(x)] = \mathbb{E}_{\hat{\rho}_1}\left[h(x)\frac{\rho_1(x)}{\hat{\rho}_1(x)}\right]$$

Variance can be exponentially bad, especially in high dimension



Common Augmentation: Annealed Langevin Dynamics

Introduce dynamics which anneal to $U_1(x)$ ***from some*** $U_0(x)$

$U_t(x) = (1-t)U_0 + tU_1$ **PDF:** $\rho_t(x) = e^{-U_t(x)+F_t}, \quad F_t = -\log Z_t$

Common Augmentation: Annealed Langevin Dynamics

Introduce dynamics which anneal to $U_1(x)$ ***from some*** $U_0(x)$

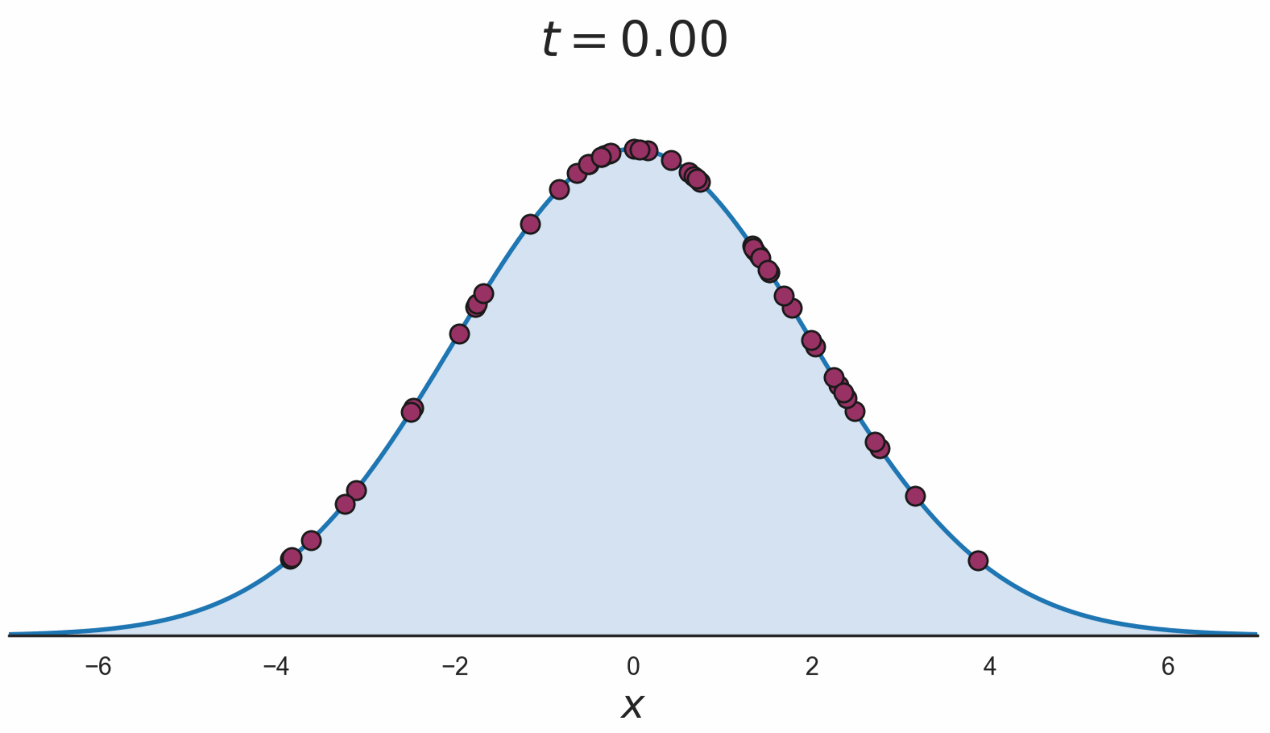
$$U_t(x) = (1-t)U_0 + tU_1$$

PDF: $\rho_t(x) = e^{-U_t(x)+F_t}, \quad F_t = -\log Z_t$

SDE:

$$d\tilde{X}_t = -\epsilon_t \nabla U_t(\tilde{X}_t)dt + \sqrt{2\epsilon_t}dW_t$$

- Time evolving potential
- ϵ_t sets speed of walkers per time step
- high temperature -> low temperature helps with multimodality



Common Augmentation: Annealed Langevin Dynamics

Introduce dynamics which anneal to $U_1(x)$ from some $U_0(x)$

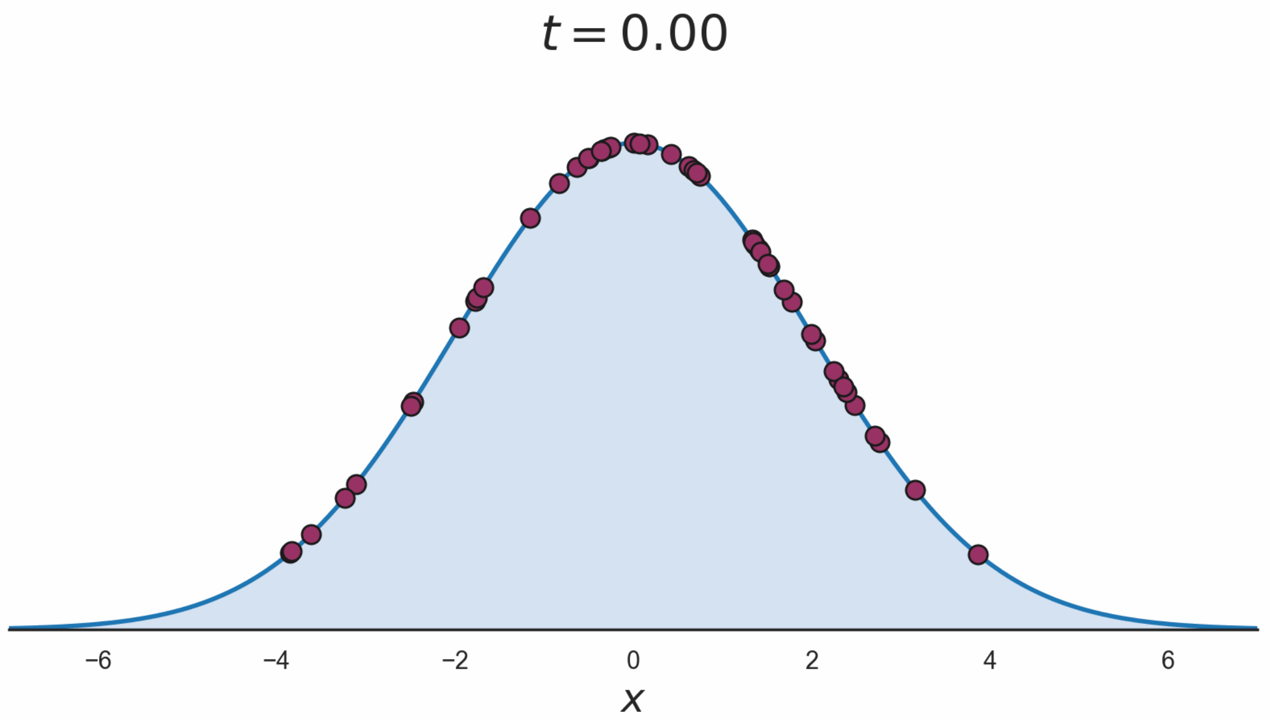
$$U_t(x) = (1-t)U_0 + tU_1$$

PDF: $\rho_t(x) = e^{-U_t(x)+F_t}, \quad F_t = -\log Z_t$

SDE:

$$d\tilde{X}_t = -\epsilon_t \nabla U_t(\tilde{X}_t)dt + \sqrt{2\epsilon_t}dW_t$$

- Time evolving potential
- ϵ_t sets speed of walkers per time step
- high temperature -> low temperature helps with multimodality



Common Augmentation: Annealed Langevin Dynamics

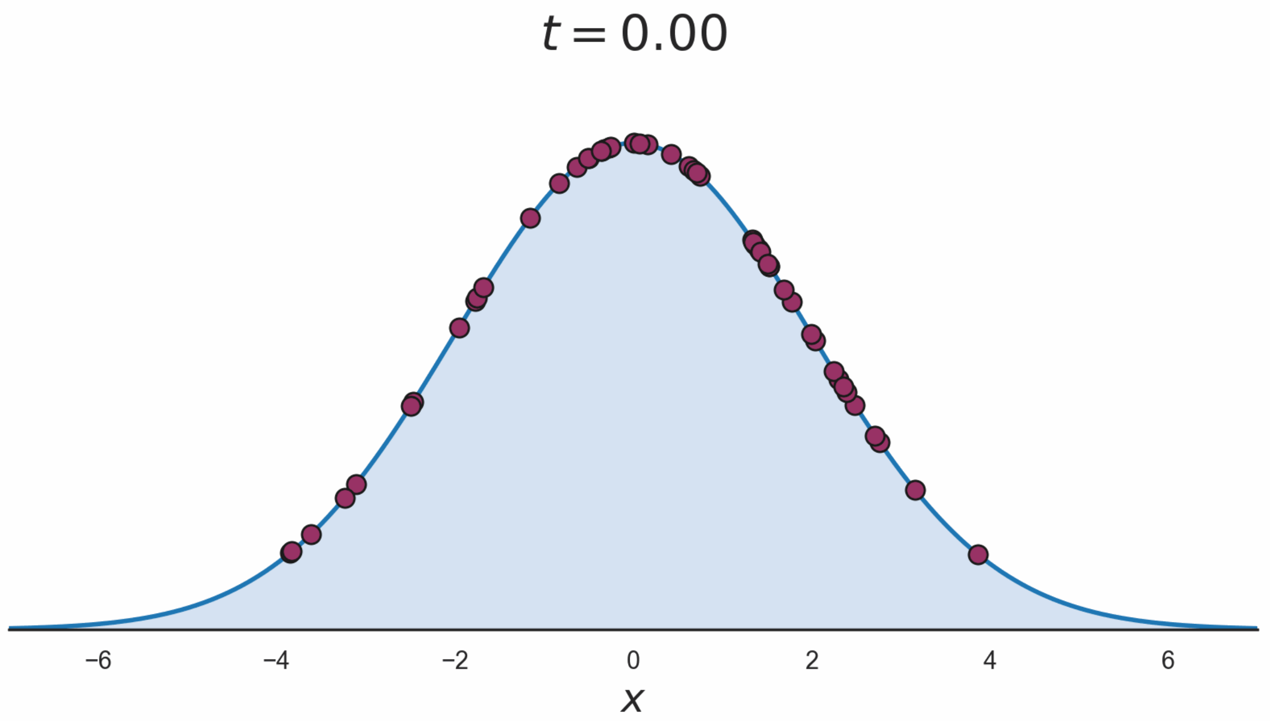
Introduce dynamics which anneal to $U_1(x)$ from some $U_0(x)$

$$U_t(x) = (1-t)U_0 + tU_1 \qquad \textbf{PDF: } \rho_t(x) = e^{-U_t(x)+F_t}, \quad F_t = -\log Z_t$$

SDE:

$$d\tilde{X}_t = -\epsilon_t \nabla U_t(\tilde{X}_t)dt + \sqrt{2\epsilon_t}dW_t$$

- Time evolving potential
- ϵ_t sets speed of walkers per time step
- high temperature -> low temperature helps with multimodality



Question: does the solution $\tilde{X}_t$ to this SDE have ρ_t as its density?

Common Augmentation: Annealed Langevin Dynamics

Introduce dynamics which anneal to $U_1(x)$ from some $U_0(x)$

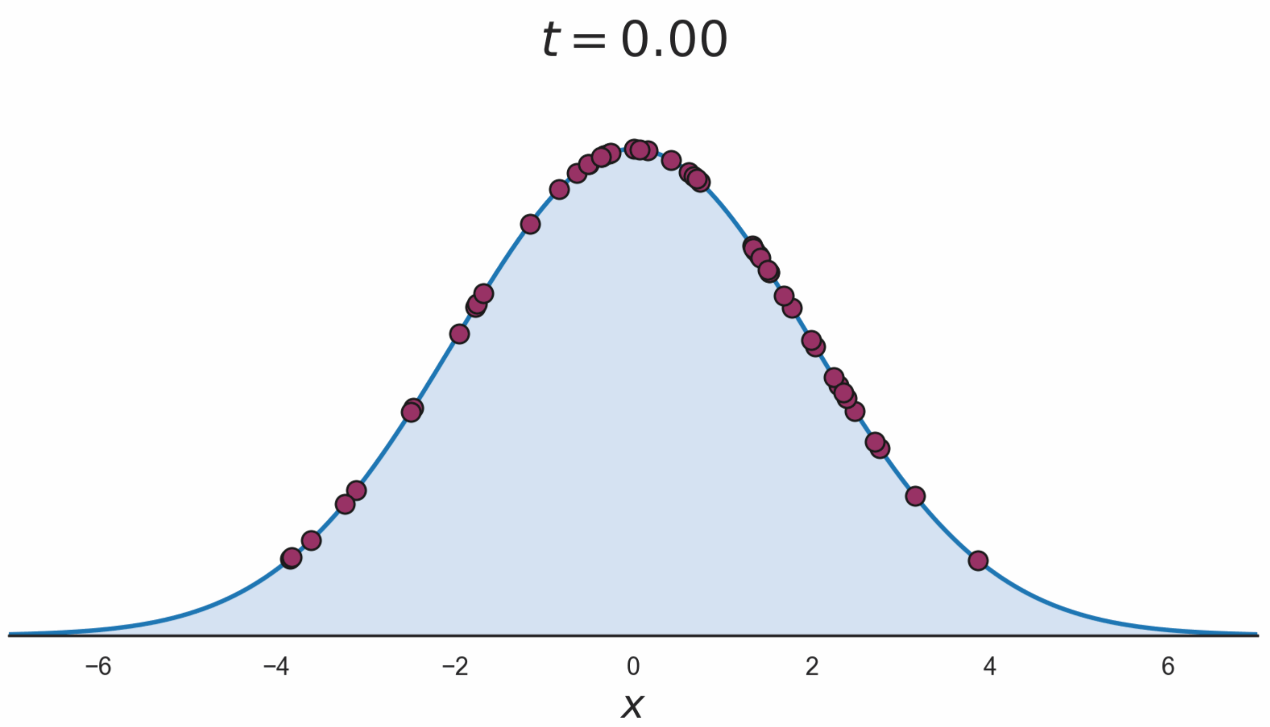
$$U_t(x) = (1 - t)U_0 + tU_1$$

PDF: $\rho_t(x) = e^{-U_t(x)+F_t}, \quad F_t = -\log Z_t$

SDE:

$$d\tilde{X}_t = -\epsilon_t \nabla U_t(\tilde{X}_t)dt + \sqrt{2\epsilon_t}dW_t$$

- Time evolving potential
- ϵ_t sets speed of walkers per time step
- high temperature -> low temperature helps with multimodality



Question: does the solution $\tilde{X}_t$ to this SDE have ρ_t as its density?

Common Augmentation: Annealed Langevin Dynamics

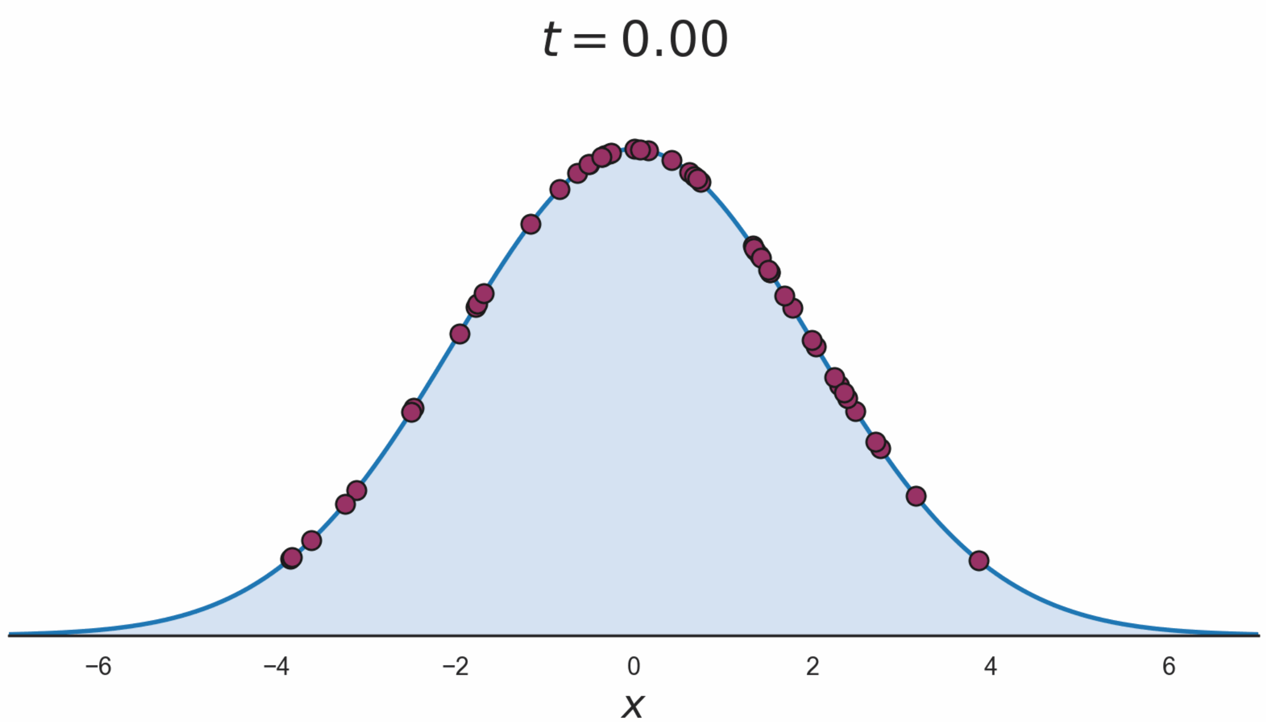
Introduce dynamics which anneal to $U_1(x)$ from some $U_0(x)$

$$U_t(x) = (1-t)U_0 + tU_1 \qquad \textbf{PDF: } \rho_t(x) = e^{-U_t(x)+F_t}, \quad F_t = -\log Z_t$$

SDE:

$$d\tilde{X}_t = -\epsilon_t \nabla U_t(\tilde{X}_t)dt + \sqrt{2\epsilon_t}dW_t$$

- Time evolving potential
- ϵ_t sets speed of walkers per time step
- high temperature -> low temperature helps with multimodality



Question: does the solution $\tilde{X}_t$ to this SDE have ρ_t as its density?

NO! only if $\epsilon_t \to \infty$ and $dt \to 0$.....Why?

Common Augmentation: Annealed Langevin Dynamics

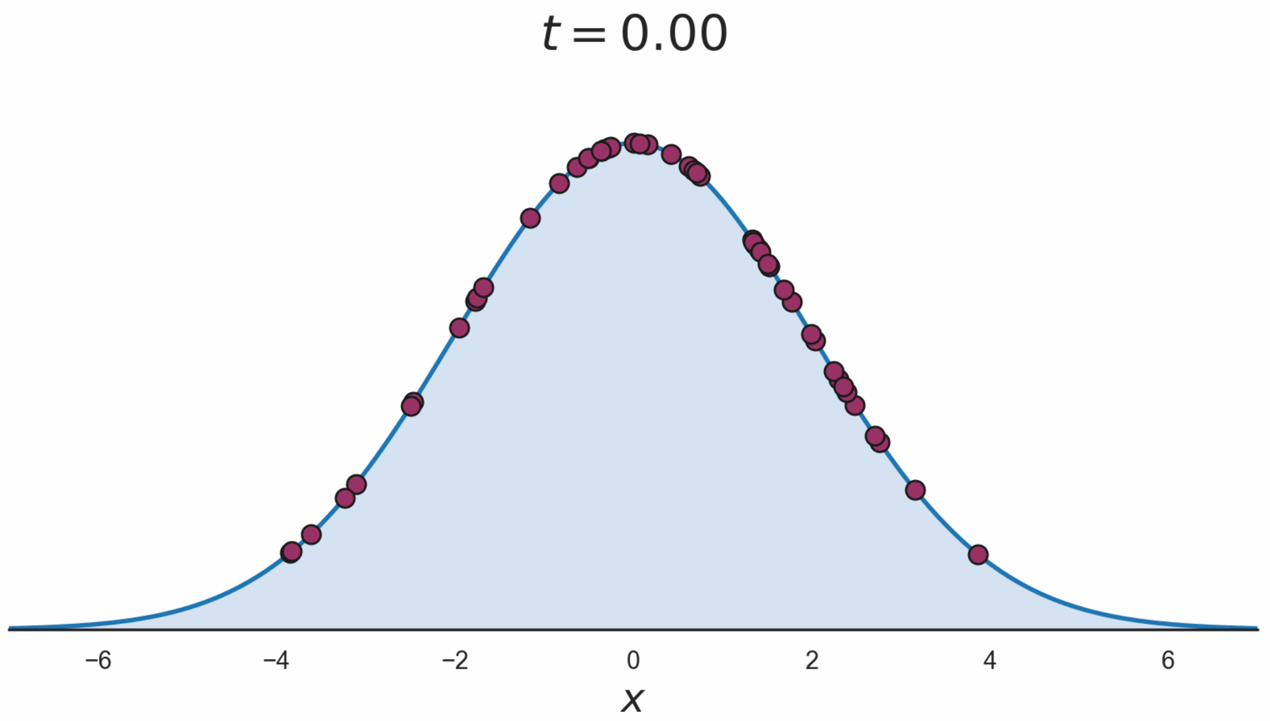
Introduce dynamics which anneal to $U_1(x)$ from some $U_0(x)$

$$U_t(x) = (1-t)U_0 + tU_1 \qquad \textbf{PDF: } \rho_t(x) = e^{-U_t(x)+F_t}, \quad F_t = -\log Z_t$$

SDE:

$$d\tilde{X}_t = -\epsilon_t \nabla U_t(\tilde{X}_t)dt + \sqrt{2\epsilon_t}dW_t$$

- Time evolving potential
- ϵ_t sets speed of walkers per time step
- high temperature -> low temperature helps with multimodality



Question: does the solution $\tilde{X}_t$ to this SDE have ρ_t as its density?

NO! only if $\epsilon_t \to \infty$ and $dt \to 0$.....Why?

$\tilde{\rho}_t \neq \rho_t$!

Compare the Fokker-Planck to $\partial_t \rho_t$

$$\rho_t(x) = e^{-U_t(x)+F_t}$$

SDE:

$$d\tilde{X}_t = -\epsilon_t \nabla U_t(\tilde{X}_t)dt + \sqrt{2\epsilon_t}dW_t$$

FPE:

$$\partial_t \tilde{\rho}_t = \epsilon \nabla \cdot (\nabla U_t \tilde{\rho}_t + \nabla \tilde{\rho}_t)$$

$\tilde{\rho}_t \neq \rho_t$!

Compare the Fokker-Planck to $\partial_t \rho_t$

$$\rho_t(x) = e^{-U_t(x)+F_t}$$

SDE:

$$d\tilde{X}_t = -\epsilon_t \nabla U_t(\tilde{X}_t)dt + \sqrt{2\epsilon_t}dW_t$$

FPE:

$$\partial_t \tilde{\rho}_t = \epsilon \nabla \cdot (\nabla U_t \tilde{\rho}_t + \nabla \tilde{\rho}_t)$$

Direct calculation:

$$\partial_t \rho_t = \frac{\partial}{\partial t}\left[e^{-U_t(x)+F_t}\right] - (\partial_t U_t - \partial_t F_t)\rho_t$$

$$= \epsilon_t \nabla \cdot (\nabla U_t \rho_t + \nabla \rho_t) + (\partial_t U_t - \partial_t F_t)\rho_t$$

since $\nabla \rho_t = -\nabla U_t \rho_t$

$\tilde{\rho}_t \neq \rho_t$!

Compare the Fokker-Planck to $\partial_t \rho_t$

$$\rho_t(x) = e^{-U_t(x)+F_t}$$

SDE:

$$d\tilde{X}_t = -\epsilon_t \nabla U_t(\tilde{X}_t)dt + \sqrt{2\epsilon_t}dW_t$$

FPE:

$$\partial_t \tilde{\rho}_t = \epsilon \nabla \cdot (\nabla U_t \tilde{\rho}_t + \nabla \tilde{\rho}_t)$$

Direct calculation:

$$\begin{aligned}\partial_t \rho_t &= \frac{\partial}{\partial t}\left[e^{-U_t(x)+F_t}\right] - (\partial_t U_t - \partial_t F_t)\rho_t \\ &= \epsilon_t \nabla \cdot (\nabla U_t \rho_t + \nabla \rho_t) + (\partial_t U_t - \partial_t F_t)\rho_t\end{aligned}$$

since $\nabla \rho_t = -\nabla U_t \rho_t$

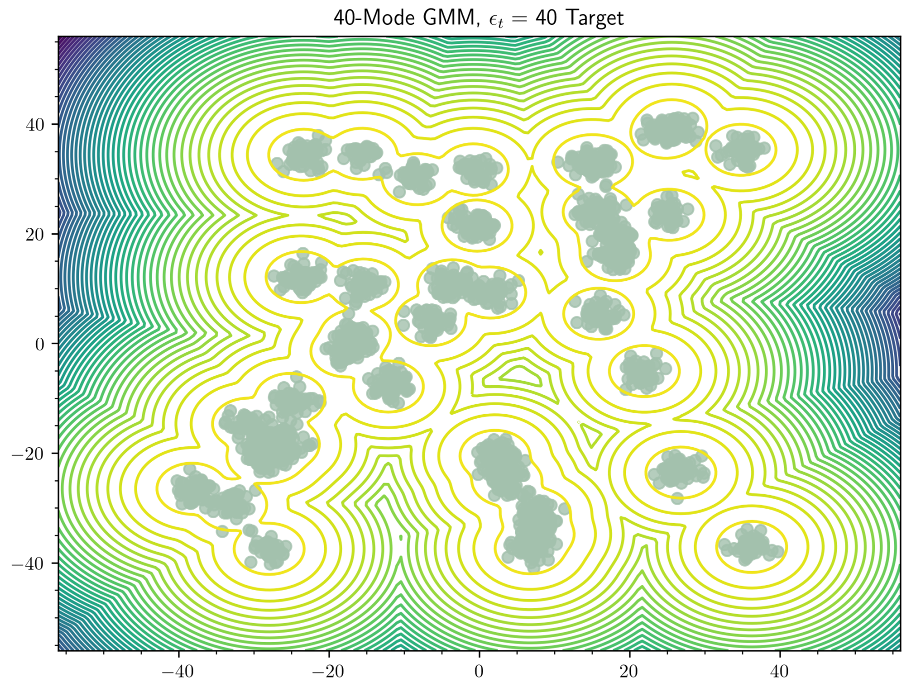
$\partial_t \rho_t$ and $\partial_t \tilde{\rho}_t$ differ by factor arising from time dynamics of U_t

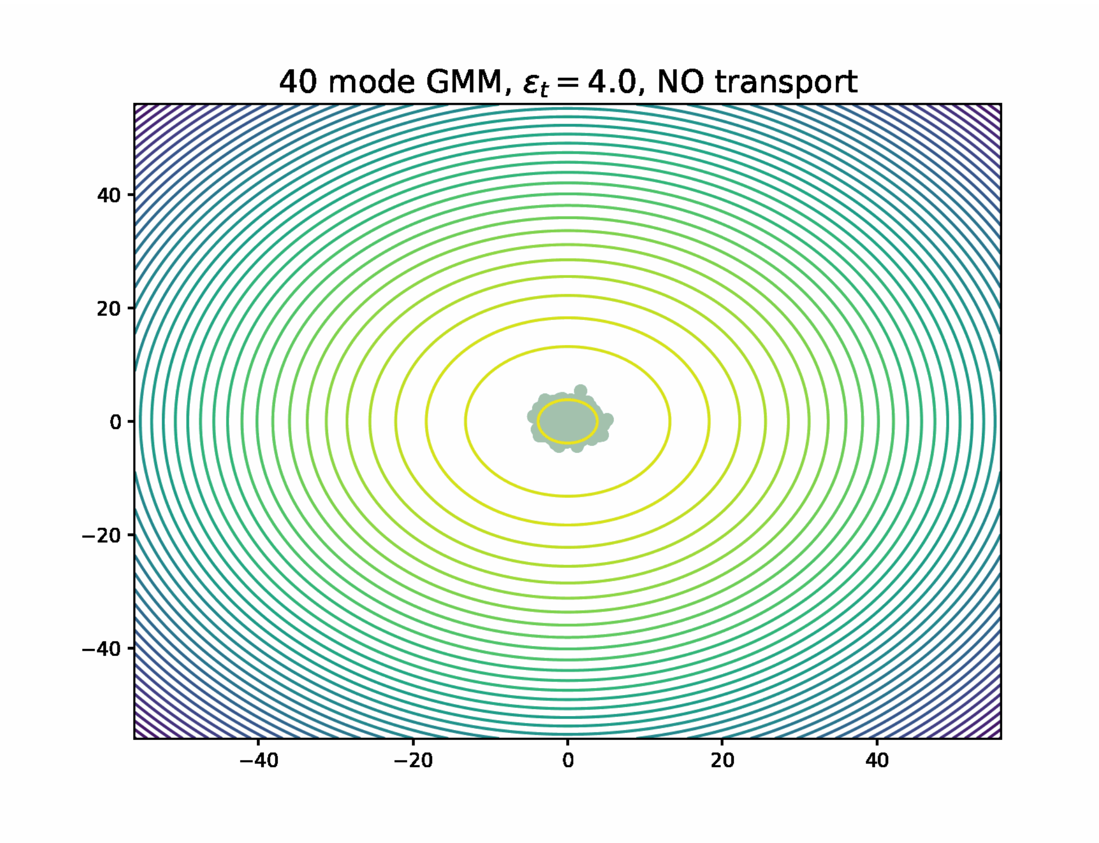
In practice, the walkers $\tilde{X}_t$ "lag behind" the intended evolution of ρ_t

The lag of $\tilde{X}_t$ in practice

Compare the Fokker-Planck to $\partial_t \rho_t$

$$\rho_t(x) = e^{-U_t(x)+F_t}$$





$\partial_t \rho_t$ and $\partial_t \tilde{\rho}_t$ differ by factor arising from time dynamics of U_t

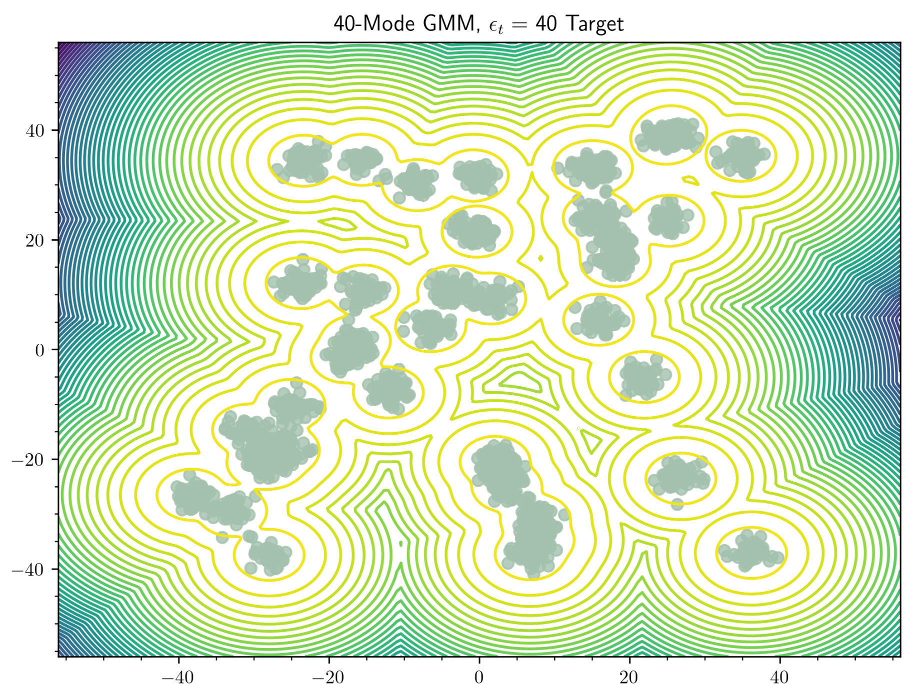
In practice, the walkers $\tilde{X}_t$ "lag behind" the intended evolution of ρ_t

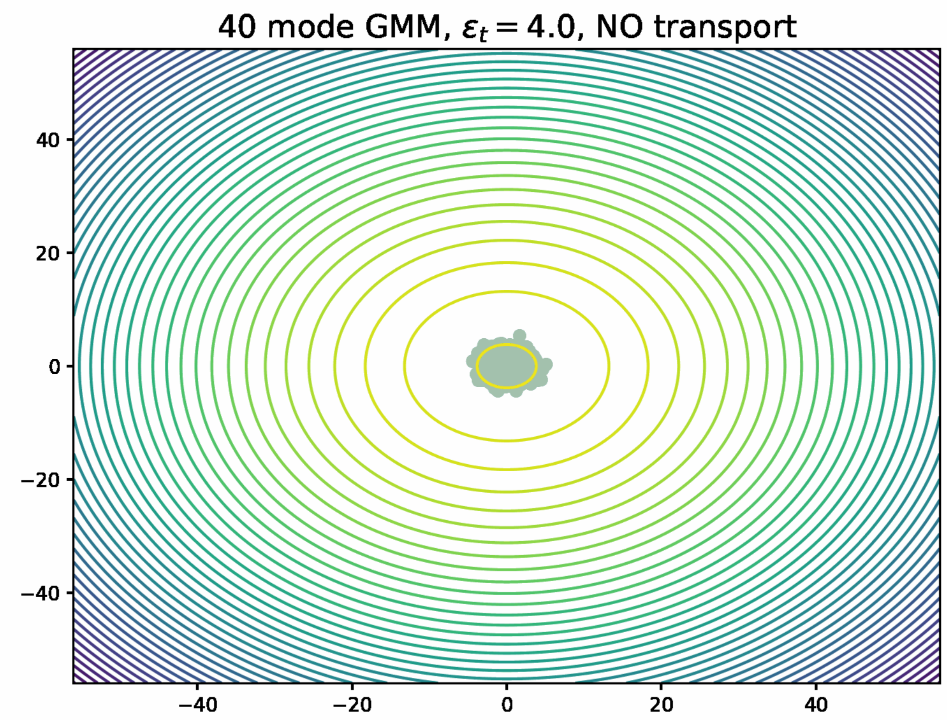
This can in theory be fixed with re-weighting

The lag of $\tilde{X}_t$ in practice

Compare the Fokker-Planck to $\partial_t \rho_t$

$$\rho_t(x) = e^{-U_t(x)+F_t}$$





$\partial_t \rho_t$ and $\partial_t \tilde{\rho}_t$ differ by factor arising from time dynamics of U_t

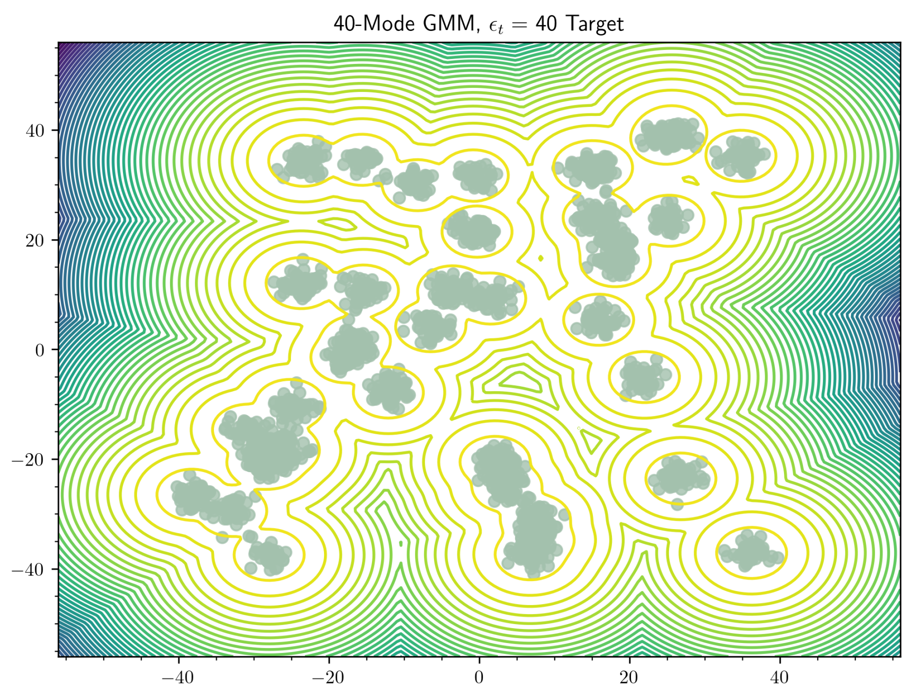
In practice, the walkers $\tilde{X}_t$ "lag behind" the intended evolution of ρ_t

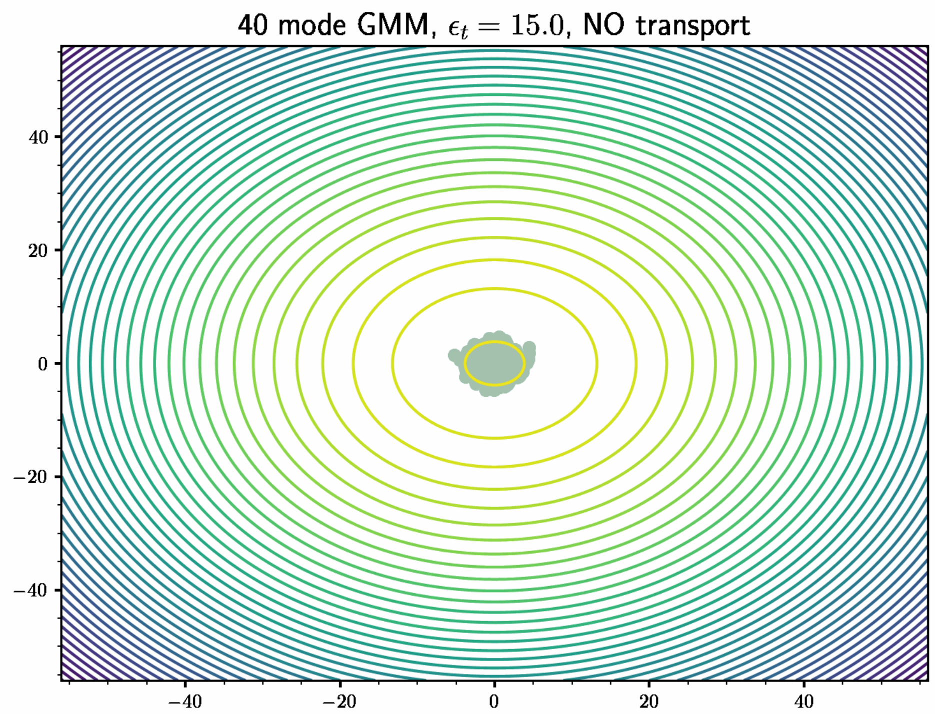
This can in theory be fixed with re-weighting

The lag of $\tilde{X}_t$ in practice

Compare the Fokker-Planck to $\partial_t \rho_t$

$$\rho_t(x) = e^{-U_t(x)+F_t}$$





$\partial_t \rho_t$ and $\partial_t \tilde{\rho}_t$ differ by factor arising from time dynamics of U_t

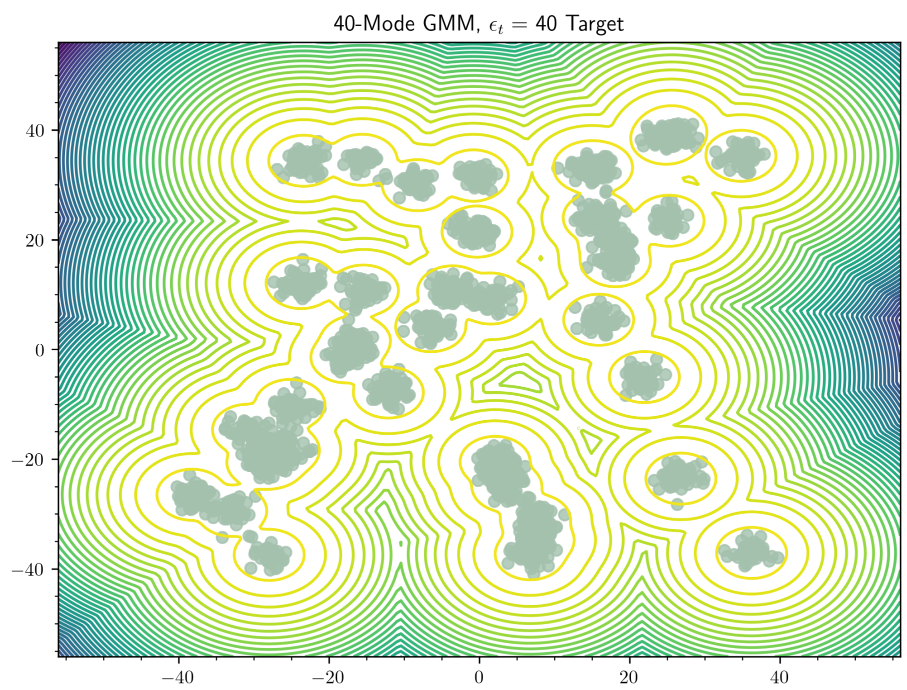
In practice, the walkers $\tilde{X}_t$ "lag behind" the intended evolution of ρ_t

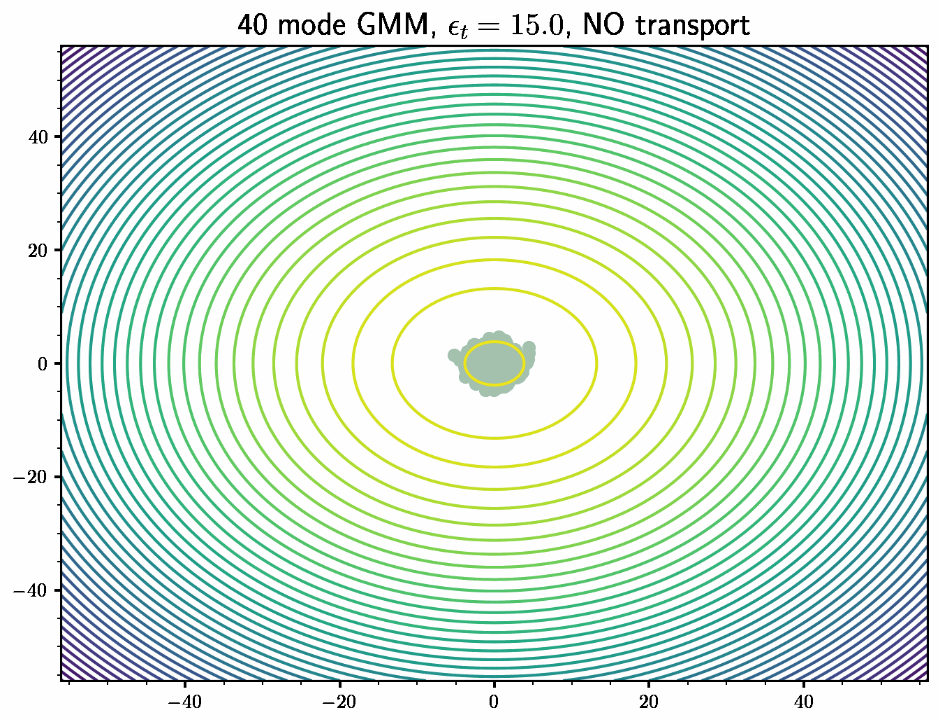
This can in theory be fixed with re-weighting

The lag of $\tilde{X}_t$ in practice

Compare the Fokker-Planck to $\partial_t \rho_t$

$$\rho_t(x) = e^{-U_t(x)+F_t}$$





$\partial_t \rho_t$ and $\partial_t \tilde{\rho}_t$ differ by factor arising from time dynamics of U_t

In practice, the walkers $\tilde{X}_t$ "lag behind" the intended evolution of ρ_t

This can in theory be fixed with re-weighting

Jarzynski Equality:

MSA & Vanden-Eijnden arXiv:2410.02711 (2024);
*Jarzynski, PRL **78**, 2690 (1997)*

Introduce weights A_t to account for the lag of the walkers

Proposition

Let (X_t, A_t) be the solution to the coupled SDE/ODE

$$dX_t = -\epsilon_t \nabla U_t(X_t)dt + \sqrt{2\epsilon_t}dW_t, \qquad X_0 \sim \rho_0$$
$$dA_t = -\partial_t U_t(X_t)dt \qquad A_0 = 0$$

then for all test functions $h(x)$, we have

$$\int_{\mathbb{R}^d} h(x)\rho_t(x)dx = \frac{\mathbb{E}[e^{A_t}h(x)]}{\mathbb{E}[e^{A_t}]} \qquad Z_t/Z_0 = e^{-F_t+F_0} = \mathbb{E}\left[e^{A_t}\right]$$

Jarzynski!

change in free energy ↔ **average work!**

- Can be proven by looking at the FPE for the joint pdf $f_t(x, a) : \mathbb{R}^{d+1} \to \mathbb{R}$

Jarzynski Equality:

***MSA** & Vanden-Eijnden arXiv:2410.02711 (2024);*
*Jarzynski, PRL **78**, 2690 (1997)*

Introduce weights A_t to account for the lag of the walkers

Proposition

Let (X_t, A_t) be the solution to the coupled SDE/ODE

$$dX_t = -\epsilon_t \nabla U_t(X_t)dt + \sqrt{2\epsilon_t}dW_t, \qquad X_0 \sim \rho_0$$
$$dA_t = -\partial_t U_t(X_t)dt \qquad A_0 = 0$$

then for all test functions $h(x)$, we have

$$\int_{\mathbb{R}^d} h(x)\rho_t(x)dx = \frac{\mathbb{E}[e^{A_t}h(x)]}{\mathbb{E}[e^{A_t}]} \qquad Z_t/Z_0 = e^{-F_t+F_0} = \mathbb{E}\left[e^{A_t}\right]$$

- Can be proven by looking at the FPE for the joint pdf $f_t(x,a) : \mathbb{R}^{d+1} \to \mathbb{R}$
- **Problem**: variance of e^{A_t} may be so large that re-weighting not useful

Jarzynski Equality:

MSA *& Vanden-Eijnden arXiv:2410.02711 (2024);*
Jarzynski, PRL ***78****, 2690 (1997)*

Introduce weights A_t to account for the lag of the walkers

Proposition

Let (X_t, A_t) be the solution to the coupled SDE/ODE

$$dX_t = -\epsilon_t \nabla U_t(X_t)dt + \sqrt{2\epsilon_t}dW_t, \qquad X_0 \sim \rho_0$$

Can we fix this with measure transport?

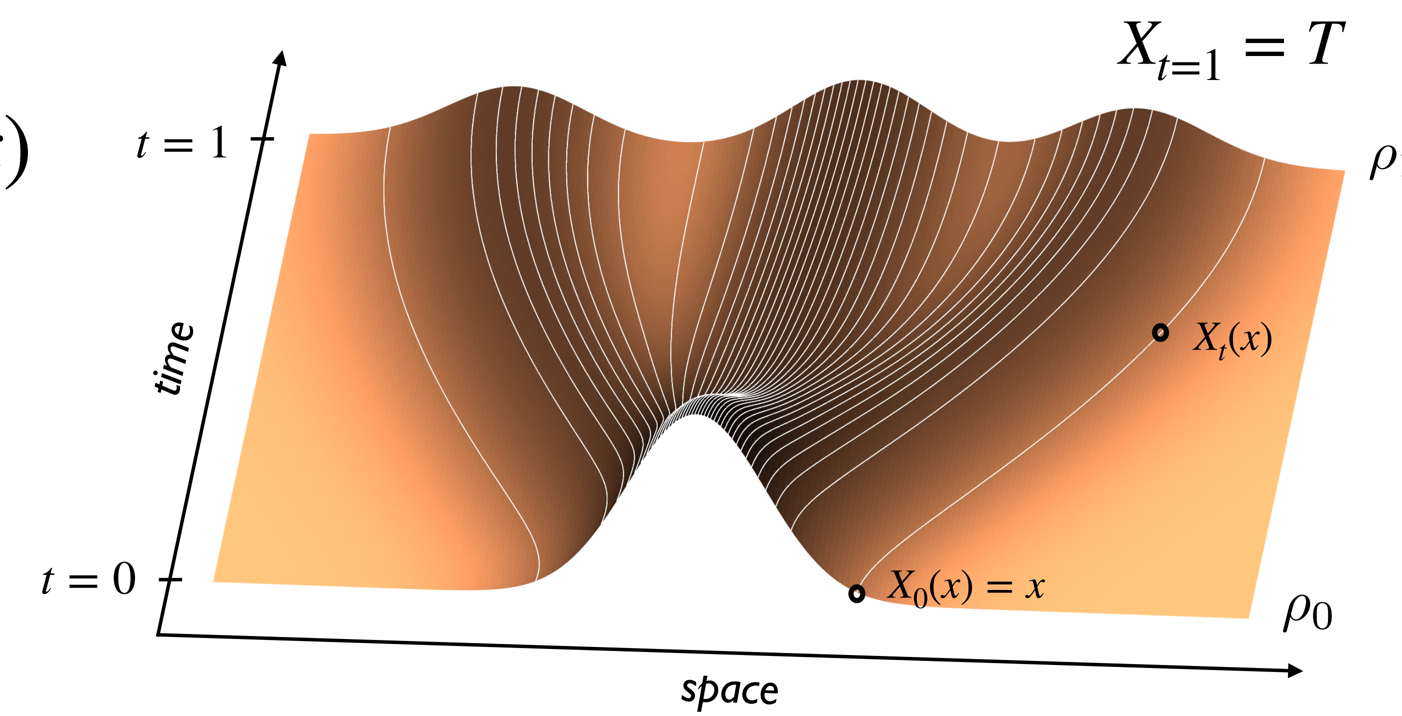
- Can be proven by looking at the FPE for thejoint pdf $f_t(x, a) : \mathbb{R}^{d+1} \to \mathbb{R}$
- **Problem**: variance of e^{A_t} may be so large that re-weighting not useful

Measure Transport

X_t flow map given by velocity field $b(t, x)$

$$X_{t=0}(x) = x \in \mathbb{R}^d$$

$$\dot{X}_t(x) = b_t(X_t(x))$$

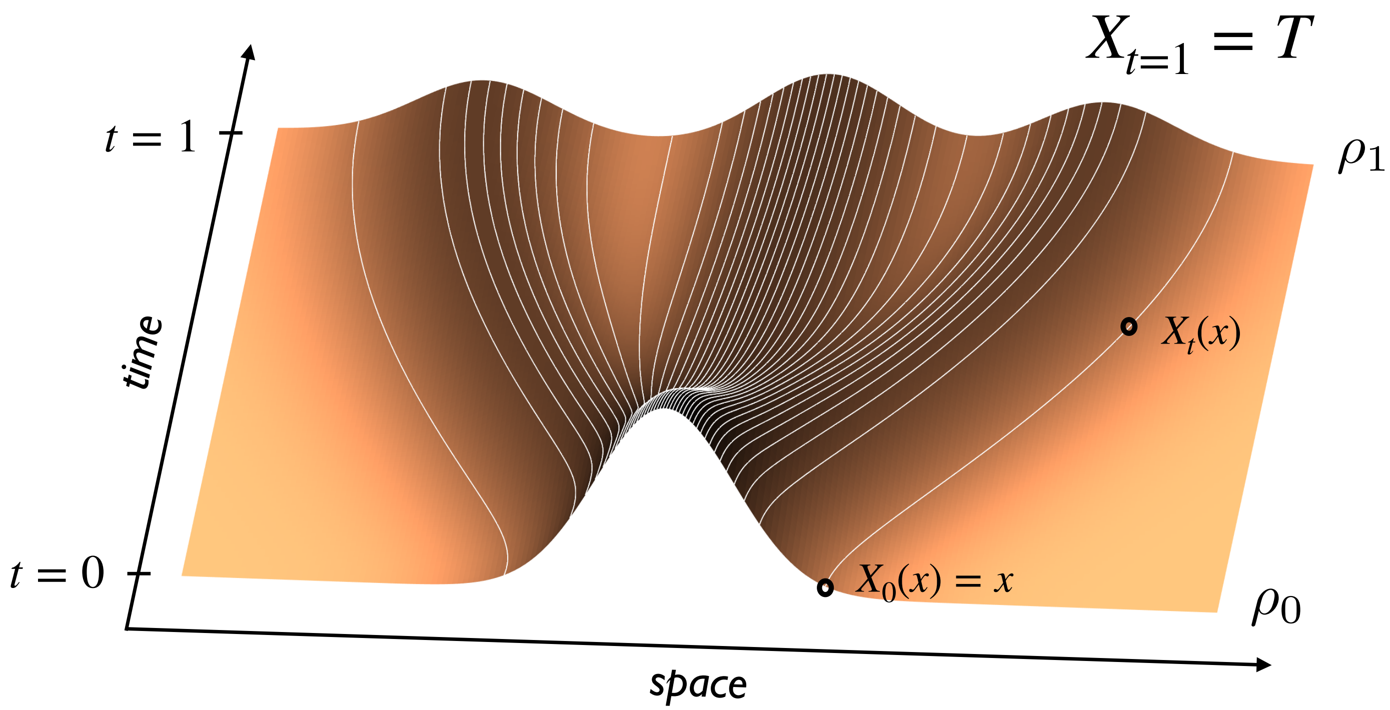


Measure Transport

X_t flow map given by velocity field $b(t, x)$

$$X_{t=0}(x) = x \in \mathbb{R}^d$$

$$\dot{X}_t(x) = b_t(X_t(x))$$



At the level of the of the distribution, how does $\rho(t, x)$ evolve?

Transport equation

$$\partial_t \rho_t + \nabla \cdot \left(b_t \rho_t\right) = 0, \quad \rho_{t=0} = \rho_0$$

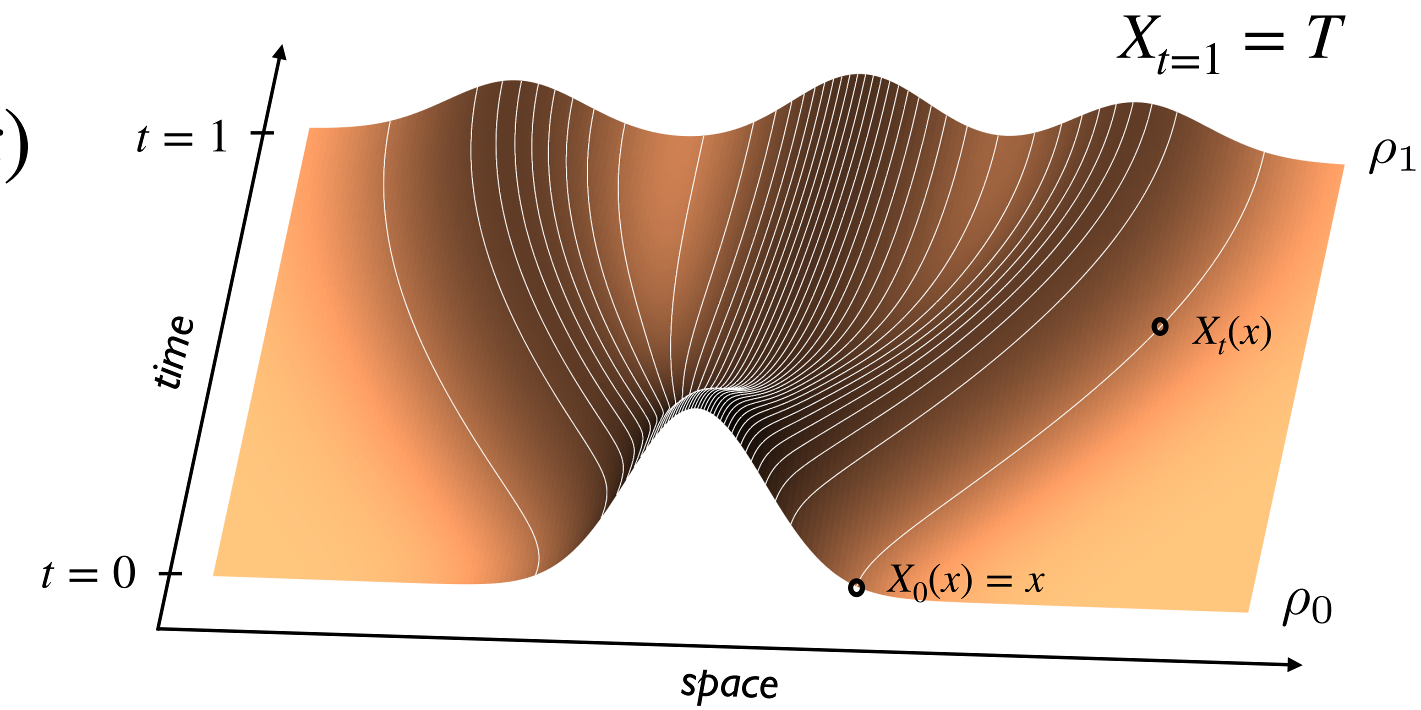
If $\rho(t)$ solves TE, **then** $\rho_{t=1} = \rho_1$

Measure Transport

X_t flow map given by velocity field $b(t, x)$

$$X_{t=0}(x) = x \in \mathbb{R}^d$$

$$\dot{X}_t(x) = b_t(X_t(x))$$



At the level of the of the distribution, how does $\rho(t, x)$ evolve?

Transport equation

$$\partial_t \rho_t + \nabla \cdot \left(b_t \rho_t\right) = 0, \quad \rho_{t=0} = \rho_0$$

If $\rho(t)$ solves TE, **then** $\rho_{t=1} = \rho_1$

Fokker-Planck Equation

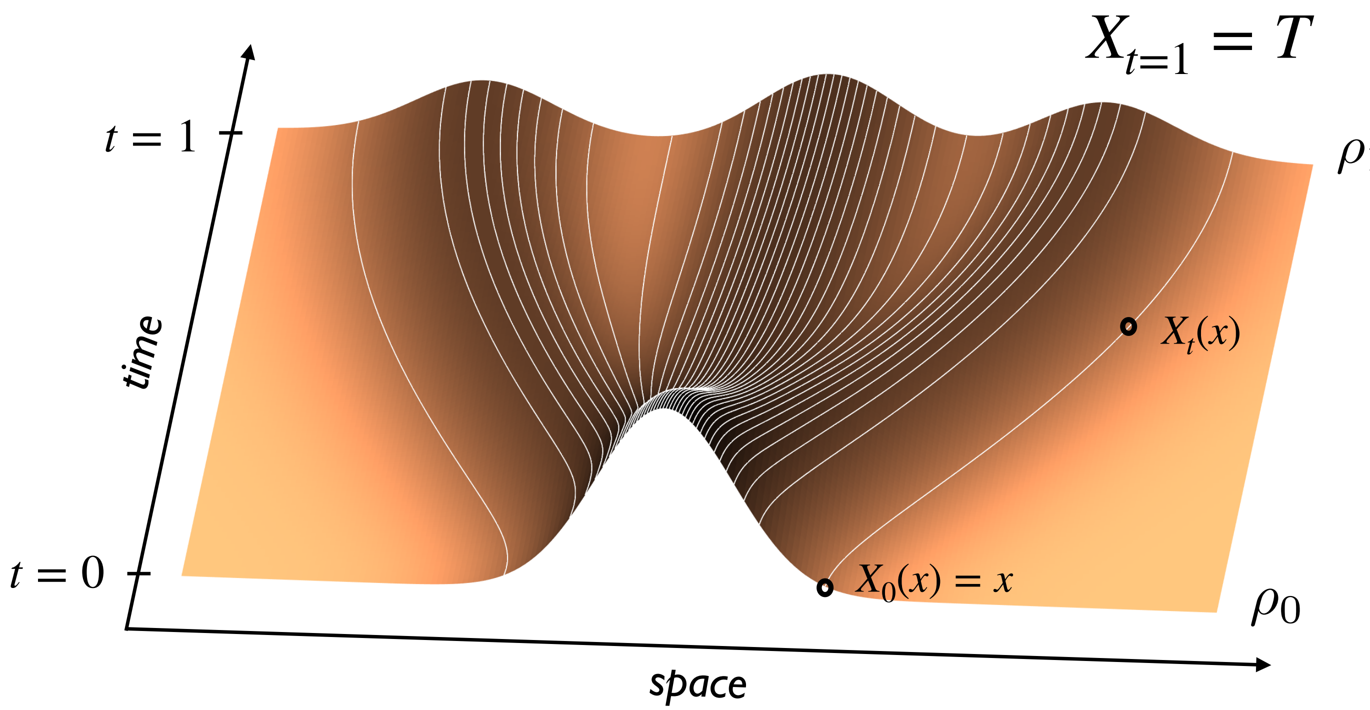
$$\partial_t \rho_t + \nabla \cdot \left(b_t \rho_t\right) = \epsilon \nabla \cdot (\nabla U_t \rho_t + \nabla \rho_t)$$

Measure Transport

X_t flow map given by velocity field $b(t, x)$

$$X_{t=0}(x) = x \in \mathbb{R}^d$$

$$\dot{X}_t(x) = b_t(X_t(x))$$



At the level of the of the distribution, how does $\rho(t, x)$ evolve?

Transport equation

$$\partial_t \rho_t + \nabla \cdot \left(b_t \rho_t\right) = 0, \quad \rho_{t=0} = \rho_0$$

If $\rho(t)$ solves TE, **then** $\rho_{t=1} = \rho_1$

Exact sampling

If X_t solves $dX_t = -\epsilon_t \nabla U_t(X_t)dt + b_t(X_t)dt + \sqrt{2\epsilon_t}dW_t$ **Then** $X_t \sim \rho_t$

Non-equilibrium transport sampler

What if you don't have the perfect b_t ?

MSA *& Vanden-Eijnden arXiv:2410.02711 (2024);*
Vargas et al ICLR (2024);
Vaikuntanathan and Jarzynski, PRL ***78****, 2690 (2008)*

Using $\nabla \cdot (\hat{b}_t \rho_t) = \nabla \cdot \hat{b}_t \rho_t - \nabla U_t \cdot b_t \rho_t$

FPE:

New non-eq term!

$$\partial_t \rho_t + \nabla \cdot (\hat{b}_t \rho_t) = \epsilon_t \nabla \cdot (\nabla U_t \rho_t + \nabla \rho_t) + (\nabla \cdot \hat{b}_t - \nabla U_t \cdot \hat{b}_t - \partial_t U_t + \partial_t F_t)\rho_t$$

Proposition

Let (X_t, A_t) be the solution to the coupled SDE/ODE

$$dX_t = -\epsilon_t \nabla U_t(X_t) dt + \sqrt{2\epsilon_t} dW_t, \qquad X_0 \sim \rho_0$$

$$dA_t = (\nabla \cdot \hat{b}_t(X_t) - \nabla U_t(X_t) \cdot \hat{b}_t(X_t) - \partial_t U_t(X_t)) dt \qquad A_0 = 0$$

then for all test functions $h(x)$, we have

$$\int_{\mathbb{R}^d} h(x) \rho_t(x) dx = \frac{\mathbb{E}[e^{A_t} h(x)]}{\mathbb{E}[e^{A_t}]} \qquad Z_t / Z_0 = e^{-F_t + F_0} = \mathbb{E}\left[e^{A_t}\right]$$

Non-equilibrium transport sampler

What if you don't have the perfect b_t ?

***MSA** & Vanden-Eijnden arXiv:2410.02711 (2024);*
Vargas et al ICLR (2024);
*Vaikuntanathan and Jarzynski, PRL **78**, 2690 (2008)*

Proposition

Let (X_t, A_t) be the solution to the coupled SDE/ODE

$$dX_t = -\epsilon_t \nabla U_t(X_t)dt + \sqrt{2\epsilon_t}dW_t, \qquad X_0 \sim \rho_0$$

$$dA_t = (\nabla \cdot \hat{b}_t(X_t) - \nabla U_t(X_t) \cdot \hat{b}_t(X_t) - \partial_t U_t(X_t))dt \qquad A_0 = 0$$

then for all test functions $h(x)$, we have

$$\int_{\mathbb{R}^d} h(x)\rho_t(x)dx = \frac{\mathbb{E}[e^{A_t}h(x)]}{\mathbb{E}[e^{A_t}]} \qquad Z_t/Z_0 = e^{-F_t+F_0} = \mathbb{E}\left[e^{A_t}\right]$$

Correctable dynamical transport for sampling

Valid for any diffusion ϵ_t which we will exploit

Strict augmentation of annealed Langevin dynamics

Learning b:

FPE:

$$\underbrace{\partial_t \rho_t + \nabla \cdot (\hat{b}_t \rho_t)}_{=0} = \epsilon_t \nabla \cdot (\nabla U_t \rho_t + \nabla \rho_t) + \underbrace{(\nabla \cdot \hat{b}_t - \nabla U_t \cdot \hat{b}_t - \partial_t U_t + \partial_t F_t)\rho_t}_{=0}$$

=0 ***Need either*** **=0**

solves the transport *removes the non-equilibrium lag*

Learning b: Physics Informed Neural Network Loss

MSA *& Vanden-Eijnden arXiv:2410.02711 (2024);*
Tian et. al ICML (2024);

FPE:

$$\underbrace{\partial_t \rho_t + \nabla \cdot (\hat{b}_t \rho_t)}_{=0} = \epsilon_t \nabla \cdot (\nabla U_t \rho_t + \nabla \rho_t) + \underbrace{(\nabla \cdot \hat{b}_t - \nabla U_t \cdot \hat{b}_t - \partial_t U_t + \partial_t F_t)\rho_t}_{=0}$$

=0 ***Need either*** **=0**

solves the transport *removes the non-equilibrium lag*

PINN Loss

Valid for any $\hat{\rho}_t$!
Controls the KL !

All minimizers (b_t, F_t) of the objective

$$L_{PINN}[\hat{b}, \hat{F}] = \int_0^1 \int_{\mathbb{R}^d} \left| \nabla \cdot \hat{b}_t(x) - \nabla U_t(x) \cdot \hat{b}_t(x) - \partial_t U_t(x) + \partial_t \hat{F}_t \right|^2 \hat{\rho}_t(x) dx dt$$

are such that $L_{PINN}[b, F] = 0$, F_t is the free energy, and b_t solves the transport

Learning b: Action Matching Loss

FPE:

$$\underbrace{\partial_t \rho_t + \nabla \cdot (\hat{b}_t \rho_t)}_{=0} = \epsilon_t \nabla \cdot (\nabla U_t \rho_t + \nabla \rho_t) + \underbrace{(\nabla \cdot \hat{b}_t - \nabla U_t \cdot \hat{b}_t - \partial_t U_t + \partial_t F_t)}_{=0} \rho_t$$

=0 ***Need either*** **=0**

solves the transport *removes the non-equilibrium lag*

Action matching loss

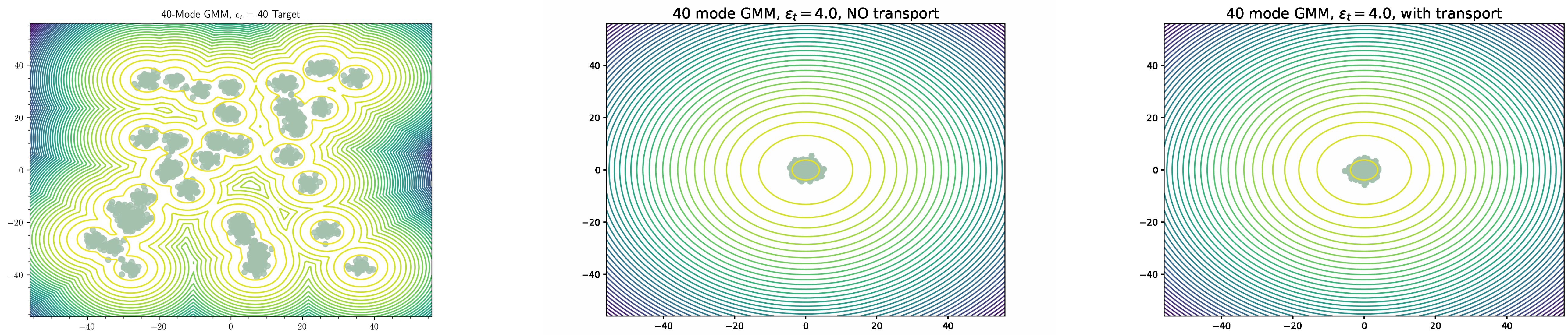
Needs reweighted samples from ρ_t

The minimizer $b_t = \nabla \phi_t$ of the objective

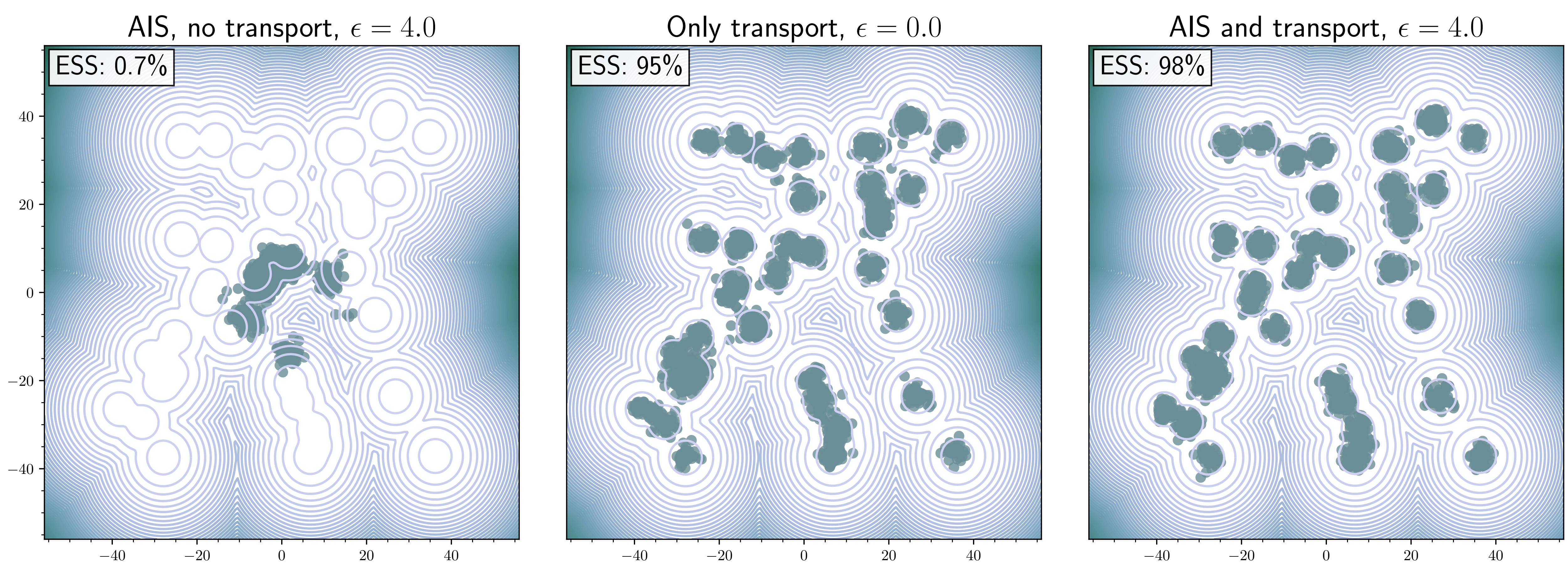
$$L_{AM}^T[\hat{\phi}] = \int_0^T \int_{\mathbb{R}^d} \left[\frac{1}{2} \left| \nabla \hat{\phi}_t(x) \right|^2 + \partial_t \hat{\phi}_t(x) \right] \rho_t(x) dx dt + \int_{\mathbb{R}^d} \left[\hat{\phi}_0(x) \rho_0(x) - \hat{\phi}_T(x) \rho_T(x) \right] dx$$

is unique up to a constant, and solves the transport.

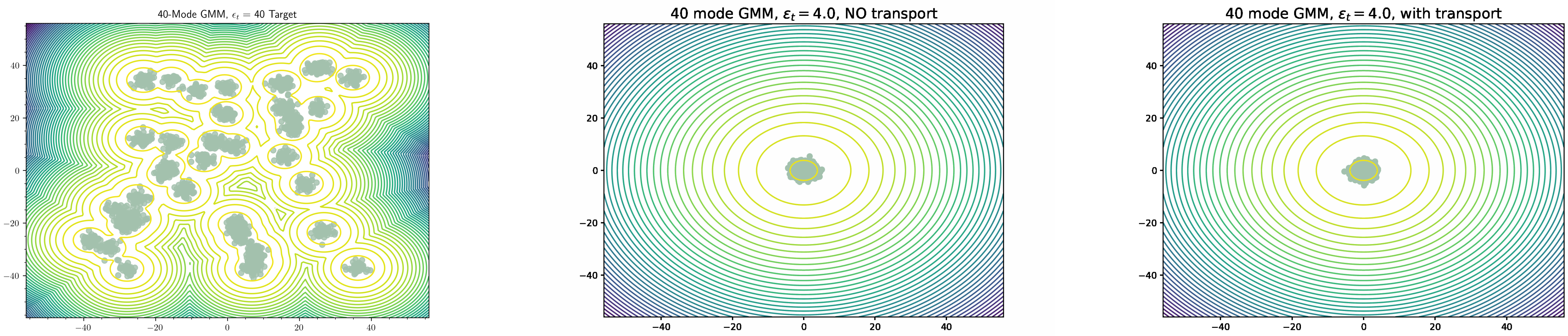
Numerical Example: Painfully multimodal GMM



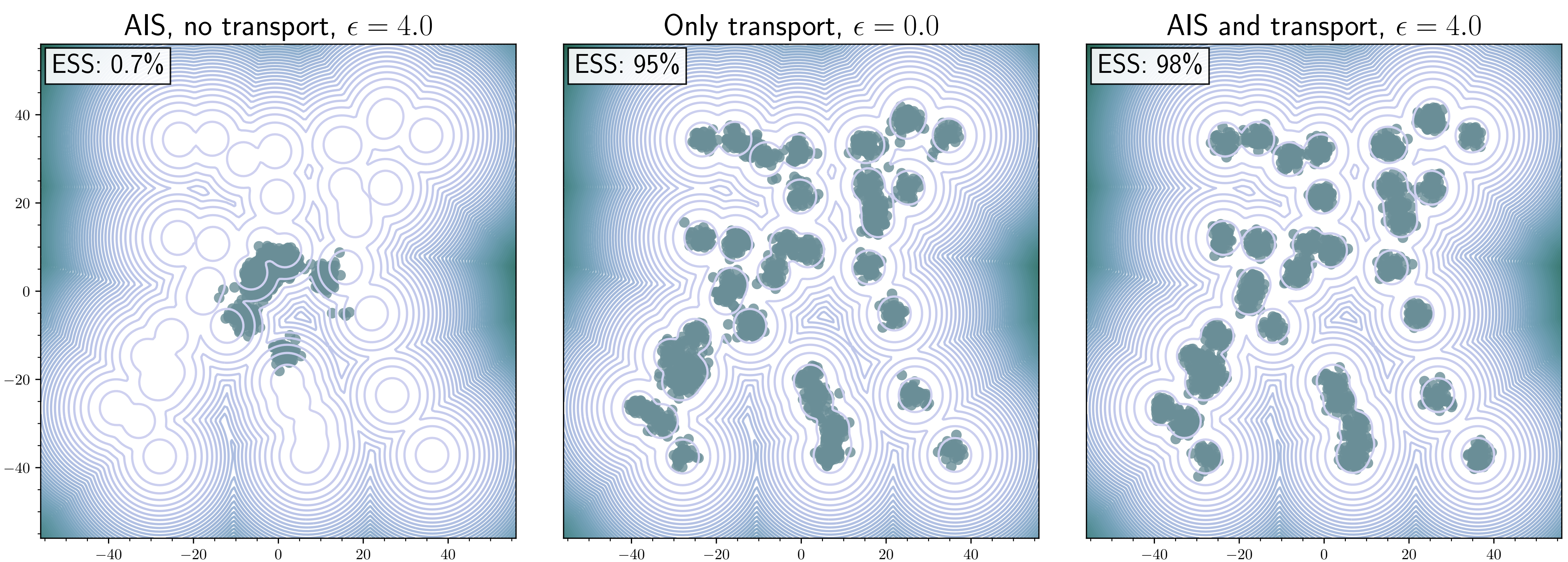
Turning on the diffusion improves ESS



Numerical Example: Painfully multimodal GMM

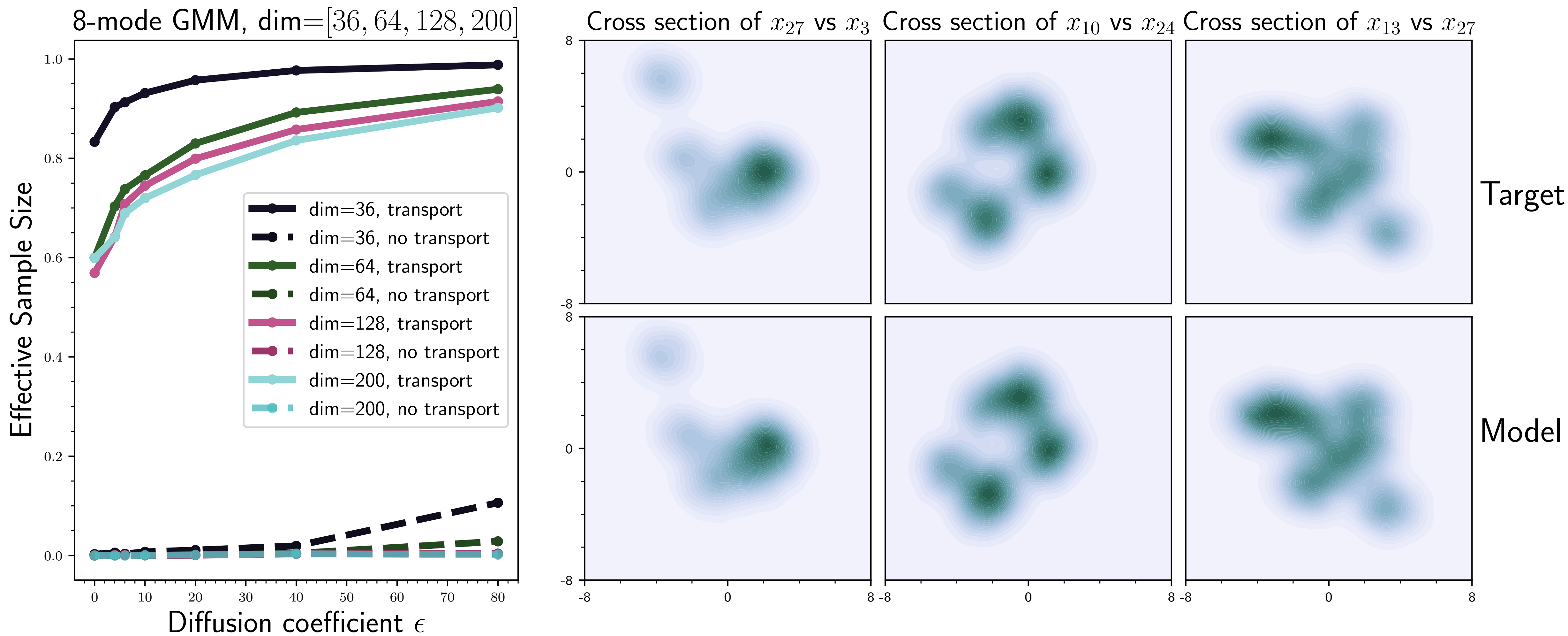


Turning on the diffusion improves ESS



More diffusion helps more with transport than without

Scaling study: Use same neural network for multimodal GMM of growing dimension



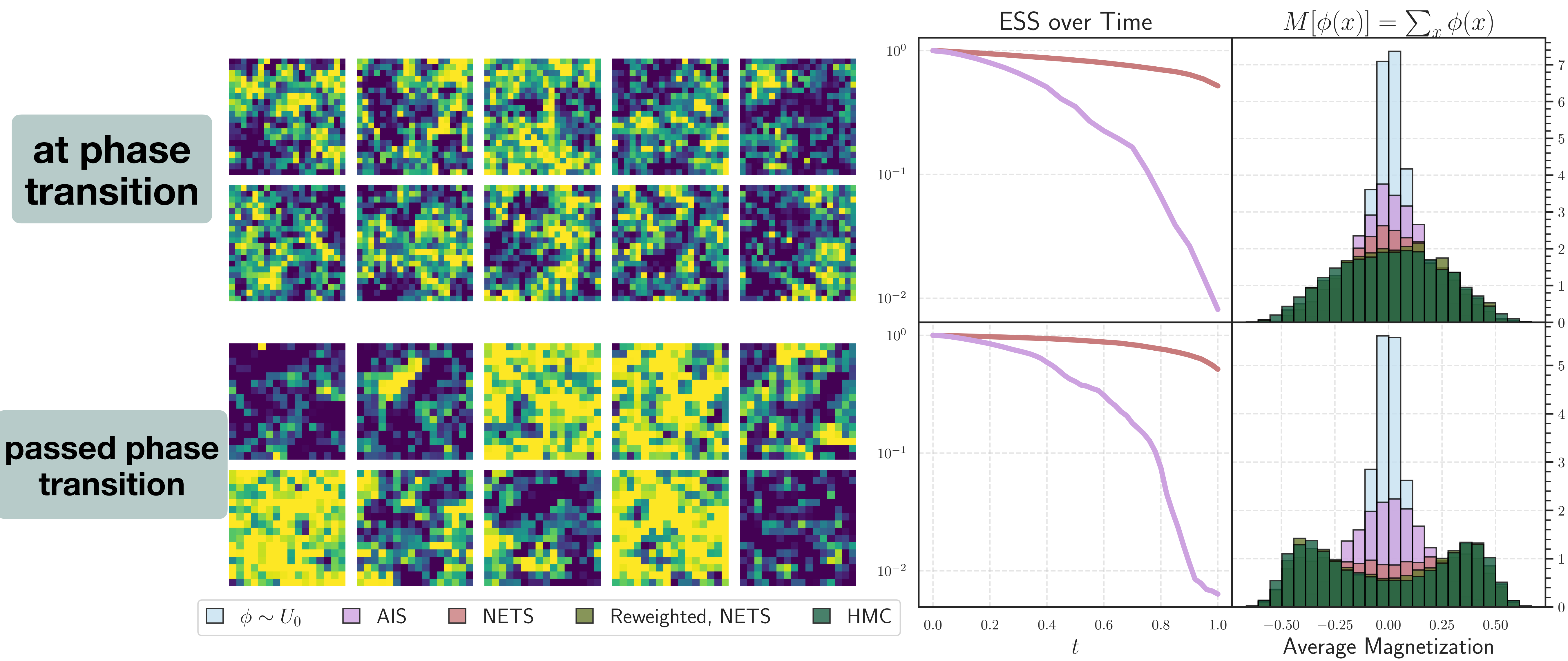
Drop in ESS for deterministic flow $\epsilon_t = 0$ can be alleviated by growing ϵ_t

Less apparent in practice if you just use annealed Langevin along!

Standard test: ϕ^4 theory

Choose energy interpolation in $m^2(t), \lambda(t)$ for the action given by

$$U_t(\varphi) = \sum_x \left[-2\sum_\mu \varphi_x \varphi_{x+\mu}\right] + \left(2D + m_t^2\right)\varphi_x^2 + \lambda_t \varphi_x^4$$



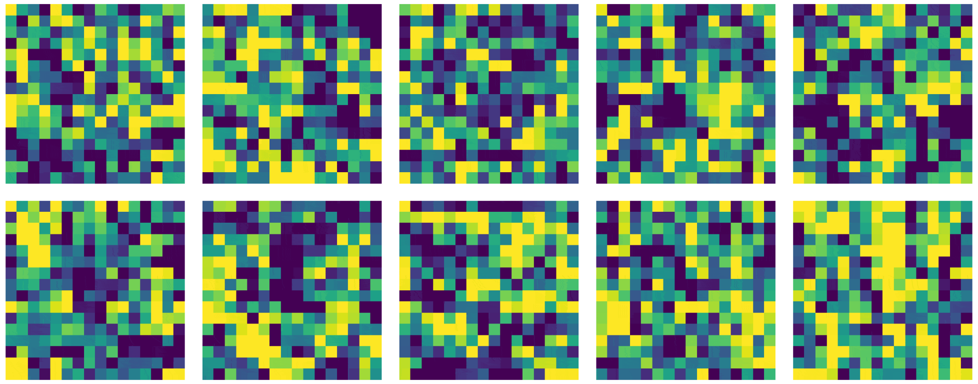
Conclusion

Dynamical formulation of unbiased sampling with transport based on Jarzynski equality

Loss functions do not require backpropagating through SDE

PINN loss is an off-policy loss! (See also Lorenz' previous talk)

Here's one more fun gif! Thanks!



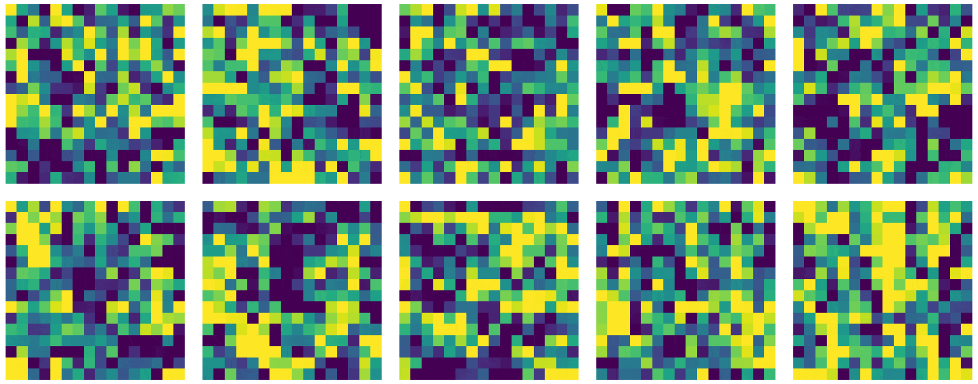
Conclusion

Dynamical formulation of unbiased sampling with transport based on Jarzynski equality

Loss functions do not require backpropagating through SDE

PINN loss is an off-policy loss! (See also Lorenz' previous talk)

Here's one more fun gif! Thanks!



Backup Slides

Computationally cheaper weights!

Note that the weights do not need a divergence if you use $b_t = \nabla \phi_t$

Proposition

Let (X_t, A_t) be the solution to the coupled SDE/ODE

$$dX_t = -\epsilon_t \nabla U_t(X_t)dt + \nabla \phi_t(X_t)dt + \sqrt{2\epsilon_t}dW_t, \qquad X_0 \sim \rho_0$$

$$dB_t = \partial_t U_t(X_t)\, dt + \frac{1}{\varepsilon_t}\partial_t \hat{\phi}_t(X_t) + \frac{1}{\varepsilon_t}\left|\nabla \hat{\phi}_t(X_t)\right|^2 dt + \sqrt{\frac{2}{\varepsilon_t}}\nabla \hat{\phi}_t(X_t)\cdot dW_t \qquad A_0 = 0$$

then for all test functions $h(x)$, we have

$$\int_{\mathbb{R}^d} h(x)\rho_t(x)dx = \frac{\mathbb{E}[e^{A_t}h(x)]}{\mathbb{E}[e^{A_t}]} \qquad Z_t/Z_0 = e^{-F_t+F_0} = \mathbb{E}\left[e^{A_t}\right]$$

where $A_t = \frac{1}{\varepsilon_t}\left[\hat{\phi}_t(X_t) - \hat{\phi}_0(X_0)\right] - B_t$

works by using expanding $d\phi_t$ with Ito formula.

Proof:

Definition of the SDE/ODE for X_t, A_t with $b_t = \nabla \phi_t$

$$dX_t = -\varepsilon_t \nabla U(X_t)dt + \hat{\nabla}\phi_t(X_t)dt + \sqrt{2\varepsilon_t}dW_t, \qquad \hat{X}_0 \sim \rho_0,$$

$$dA_t = \Delta\hat{\phi}_t(X_t)dt - \nabla U_t(X_t)\cdot\nabla\hat{\phi}(X_t)dt - \partial_t U_t(X_t)dt, \qquad A_0 = 0,$$

Ito formula says

$$d\hat{\phi}_t(X_t) = \partial_t\hat{\phi}_t(X_t)dt - \varepsilon_t\nabla\hat{\phi}_t(X_t)\cdot\nabla U(X_t)dt + \left|\nabla\hat{\phi}_t(X_t)\right|^2 dt$$
$$+\sqrt{2\varepsilon_t}\nabla\hat{\phi}_t(X_t)\cdot dW_t + \varepsilon_t\Delta\hat{\phi}_t(X_t)dt,$$

Solving for $\Delta\phi_t$ allows us to write the relation

$$dA_t = \frac{1}{\varepsilon_t}d\hat{\phi}_t(X_t)dt + dB_t$$

where $dB_t = \partial_t U_t(X_t)\,dt + \frac{1}{\varepsilon_t}\partial_t\hat{\phi}_t(X_t) + \frac{1}{\varepsilon_t}\left|\nabla\hat{\phi}_t(X_t)\right|^2 dt + \sqrt{\frac{2}{\varepsilon_t}}\nabla\hat{\phi}_t(X_t)\cdot dW_t$

Proposition 5 (KL control)**.** *Let $\hat{\rho}_t$ be the solution to the transport equation*

$$\partial_t \hat{\rho}_t = -\nabla \cdot (\hat{b}_t \rho_t), \qquad \hat{\rho}_{t=0} = \rho_0 \tag{27}$$

where $\hat{b}_t(x)$ is some predefined velocity field. Then, given any estimate $\hat{F}_t$ of the exact free energy F_t, we have

$$D_{KL}(\hat{\rho}_{t=1} || \rho_1) \leq \sqrt{L_{PINN}^{T=1}(\hat{b}, \hat{F})}. \tag{28}$$

This proposition is proven in Appendix 5.1.