Celebrate Paul Dupuis and his contributions

A Mean-field games laboratory for generative modeling: implications for robust generative algorithms

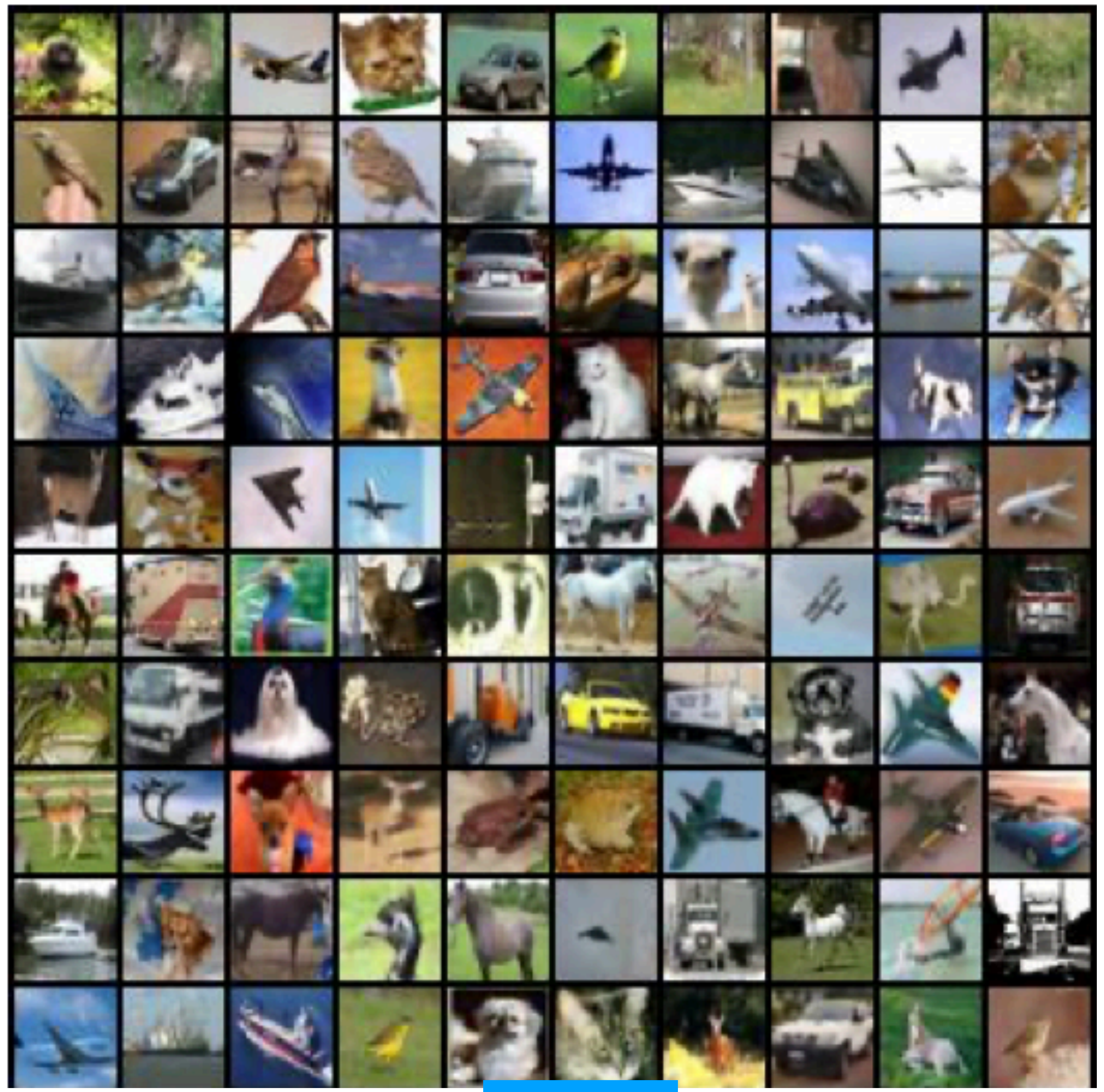
Markos Katsoulakis
Mathematics & Statistics
UMass Amherst

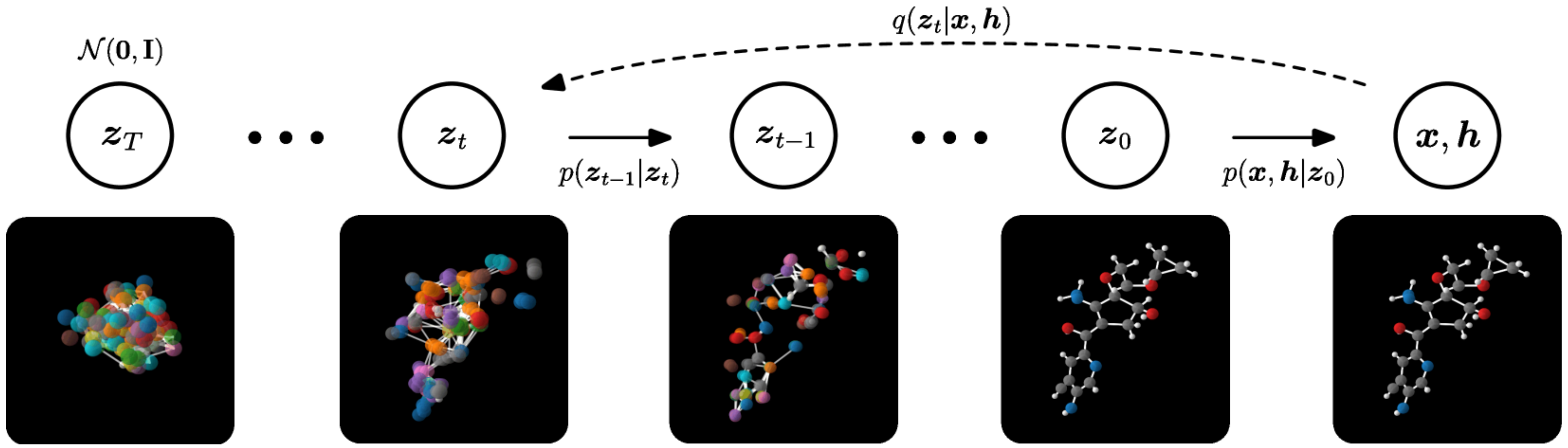
ICERM 2024 — Robust Optimization and Simulation of Complex Stochastic Systems
September 2024

Generative modeling

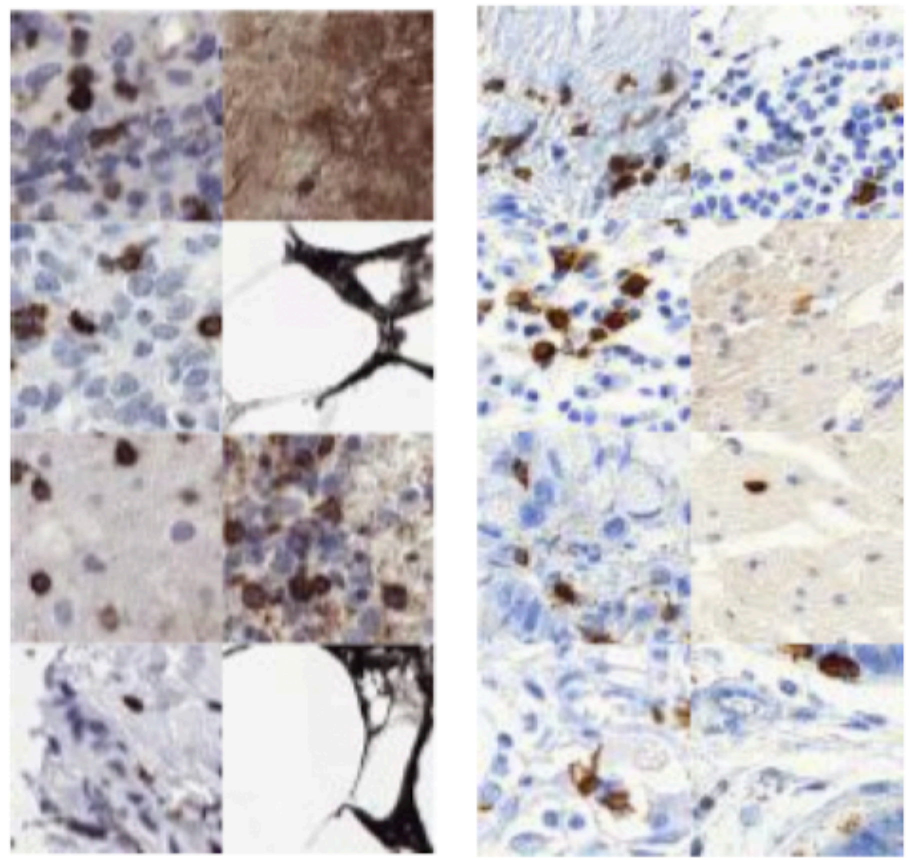

Stable diffusion

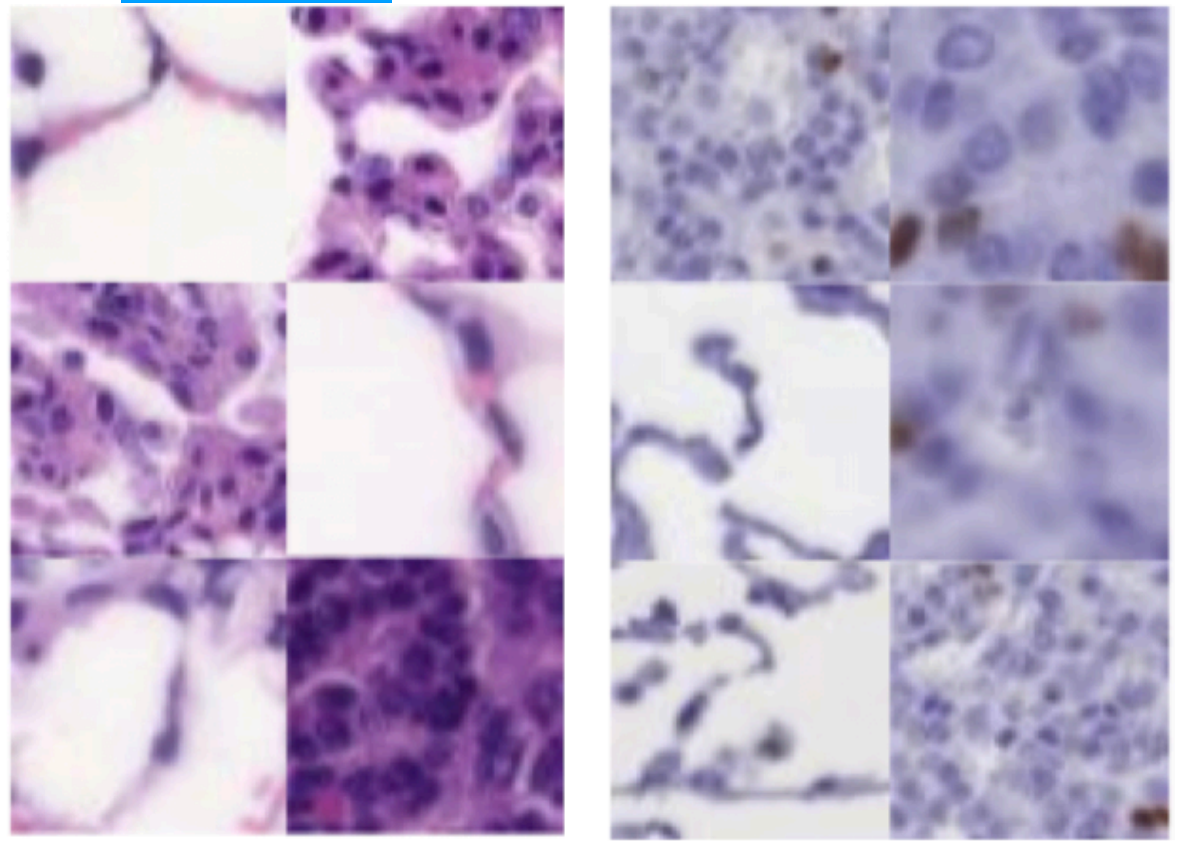
ChatGPT


CIFAR10



Molecular generation
Hoogeboom et al. 2022


LYSTO[1]


ANHIR[2]

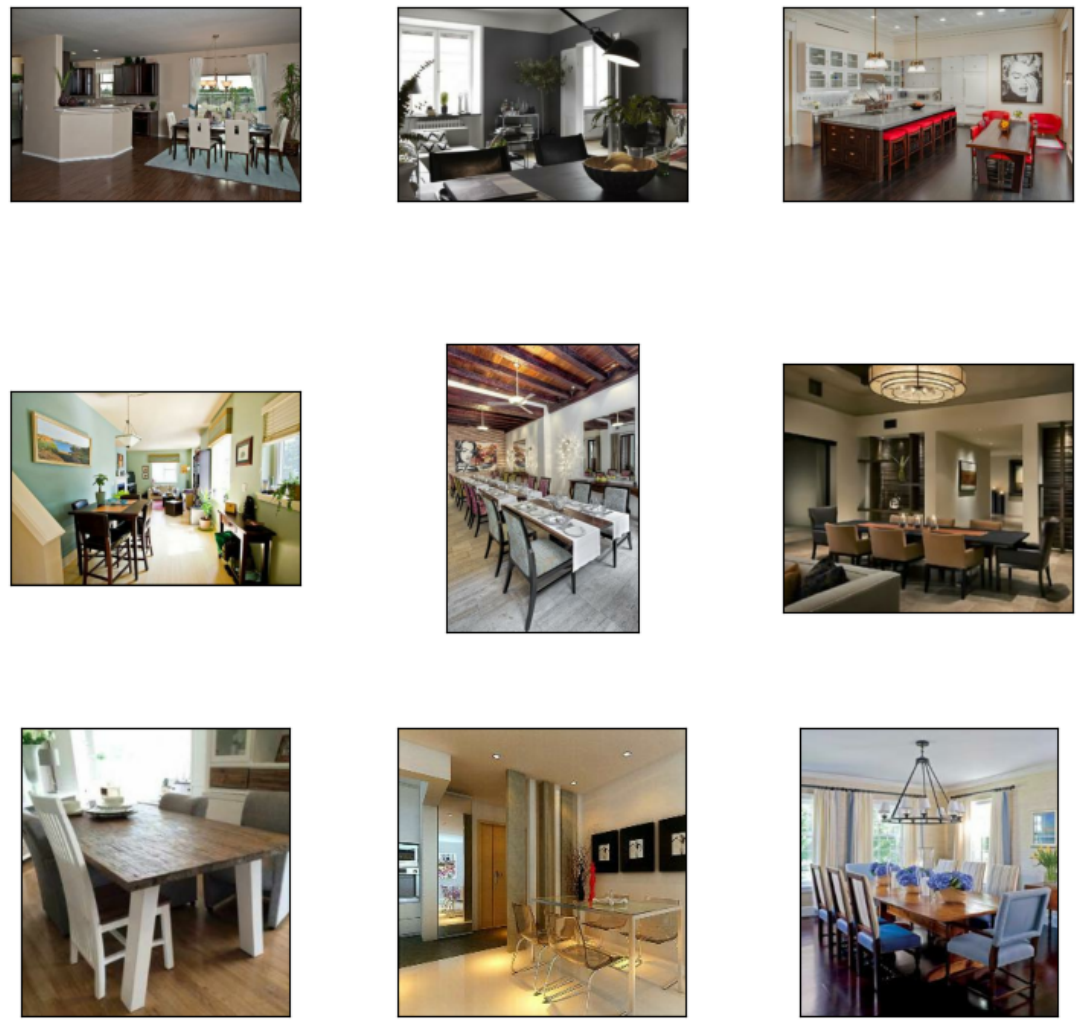
1. *Ciompi et al., Zenodo 2019*
2. *Borovec et al., IEEE Transactions on Medical Imaging 2020*

What is Generative Modeling?

Given a dataset, e.g., images of bedrooms (LSUN dataset), create more data

Assume an *unknown* **data distribution**

$p_{data}(x) = \pi(x)$ $\longrightarrow$

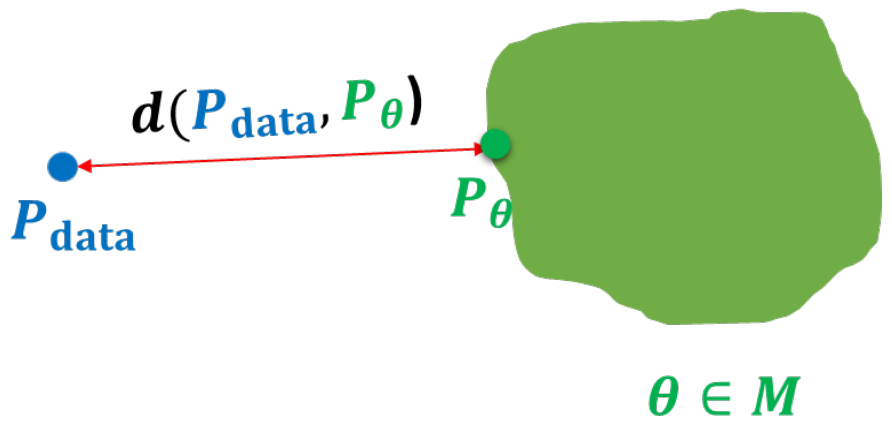


$\longrightarrow$

Generative model distribution

$p_g(x) = p_\theta(x)$

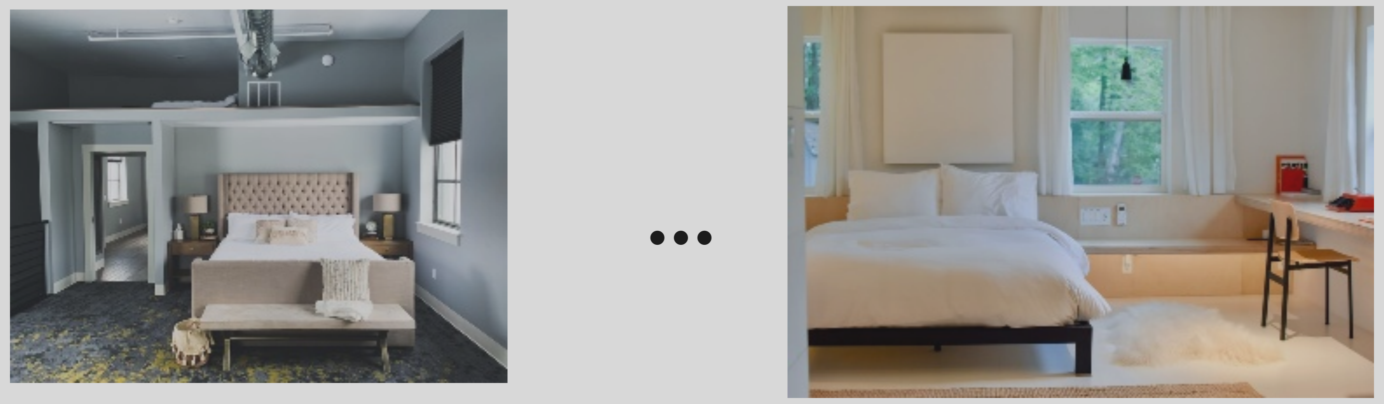
Parametrized family of distributions; use Deep Learning/ neural networks



Model family

Goal: Find $\theta \in \theta$ such that $p_\theta(x) \approx p_{data}(x) = \pi(x)$

Generative because sampling from $p_\theta(x)$ produces new unseen images



$\sim \quad p_\theta(x)$

Generative modeling

$\mathscr{D} = \{x_1, x_2, \ldots, x_n\} \longrightarrow p_{\theta^*}(x) \approx \pi(x) \longrightarrow \widetilde{\mathscr{D}} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$

Generative modeling approaches

Main Goal: Given data $X_i \sim \pi, i = 1,\cdots,N$ from an (unknown) **target distribution** π **,** reproduce *new* samples from π

- Pick a **source distribution** ρ, easy to simulate (e.g. Gaussian).
- **Generative map (one-shot)**: Learn a transport map Φ such that:

$$\Phi_{\#}\rho \approx \pi$$ (e.g. GANs, distillation methods)

Or

- **Generative flow**: Learn a transport flow via a vector field $v(x,t)$ such that

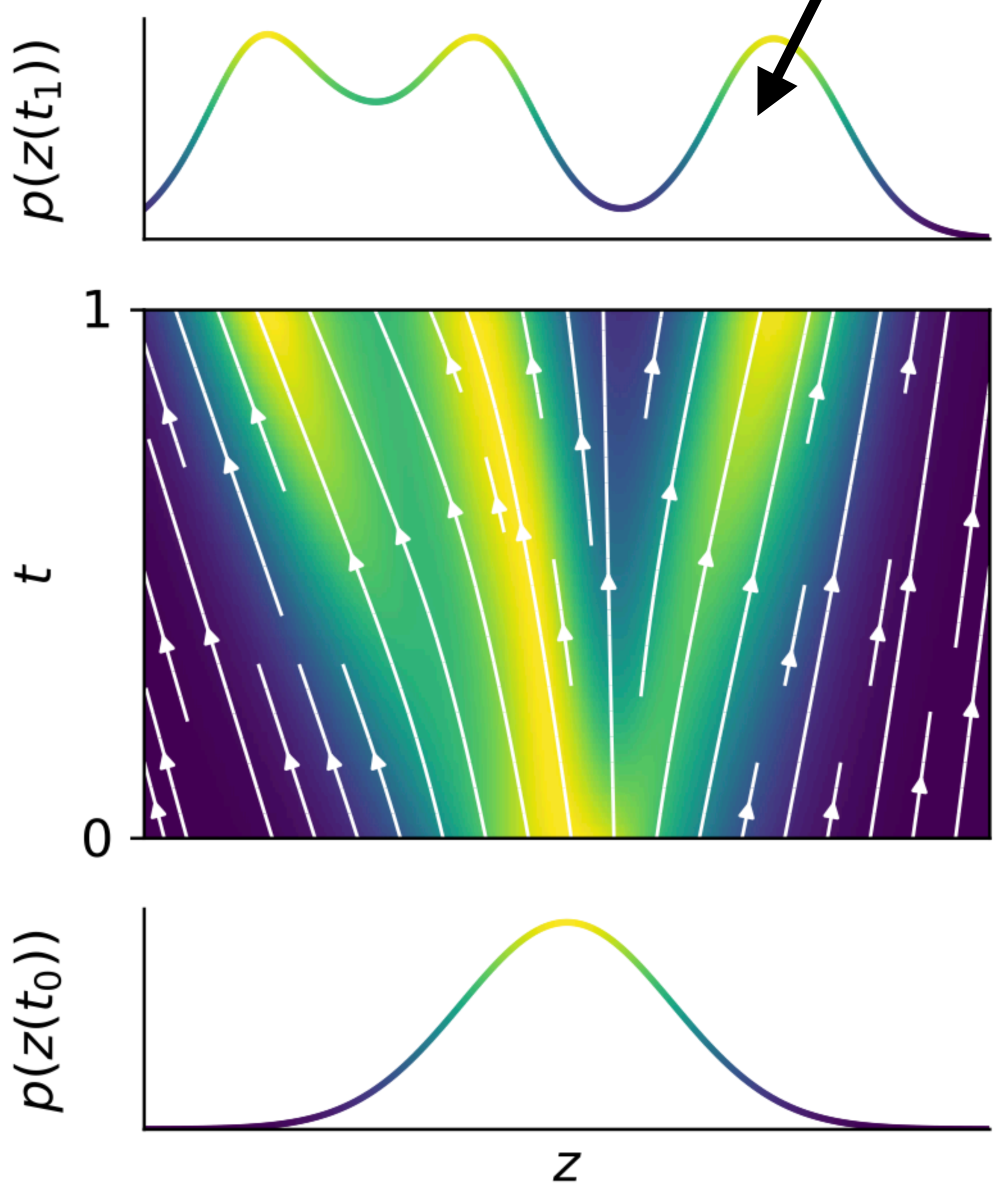
$$dx(t) = v(x(t),t)\,dt + \sigma\,dW(t) \quad \text{s.t.} \quad x(0) \sim \rho \qquad x(T) \sim \pi$$

Normalizing flows $(\sigma = 0)$ learn an ODE

Diffusion models $(\sigma > 0)$ learn a SDE

Flow-based generative modeling

- Given dataset $\{X_i\} \sim \pi$ data distribution viewed as **particles**
- Flow-based generative modeling: based on ODE or SDE for **transport of probability measures**



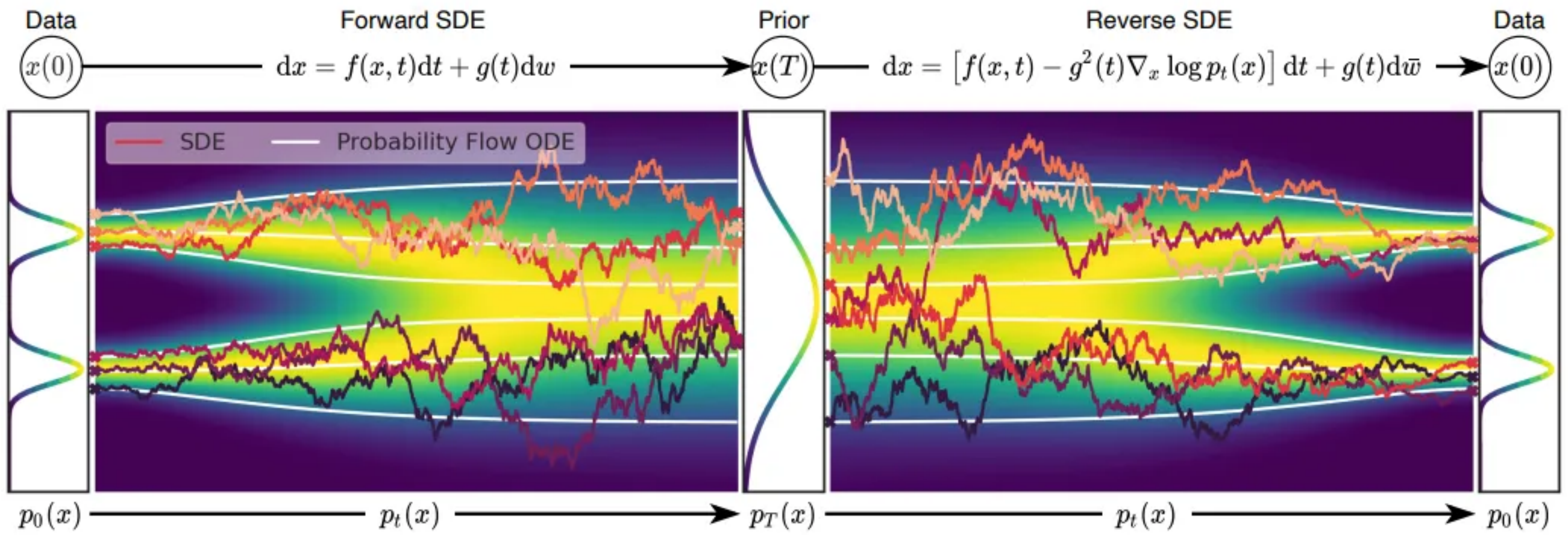
Continuous normalizing flows

Grathwohl et al. '18

Score-based generative models

Song et al. '20

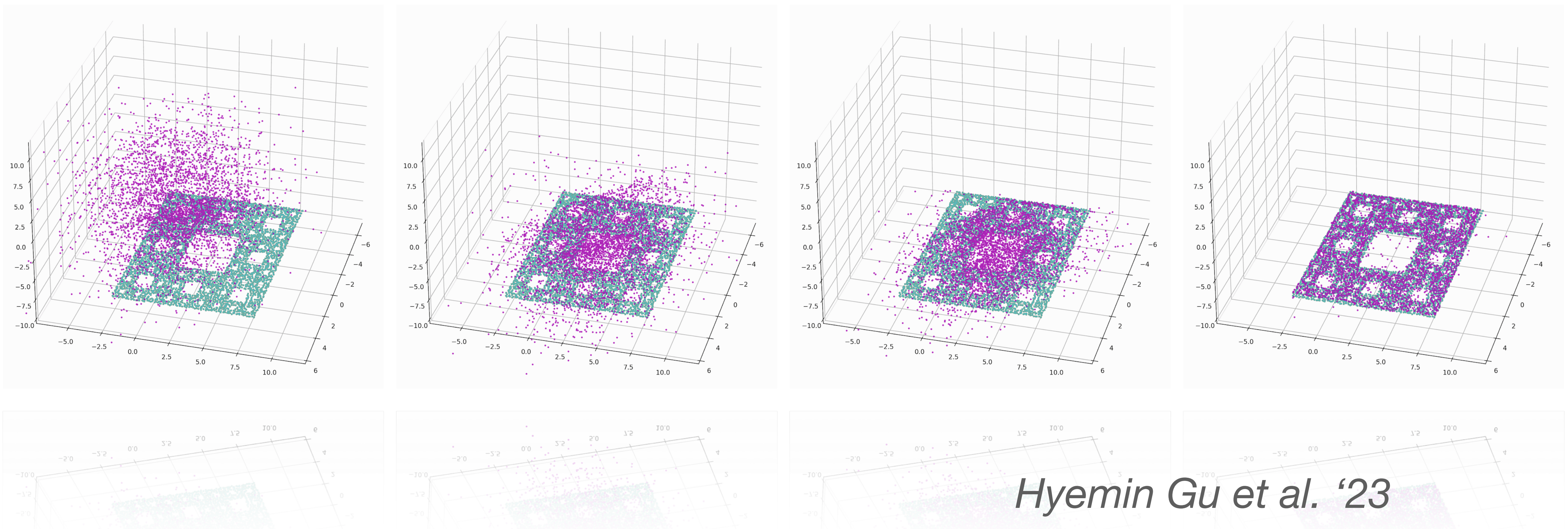
Ho et al. '20



Wasserstein Gradient flows

JKO '98

Santambrogio '15



Hyemin Gu et al. '23

Continuous-time normalizing flows

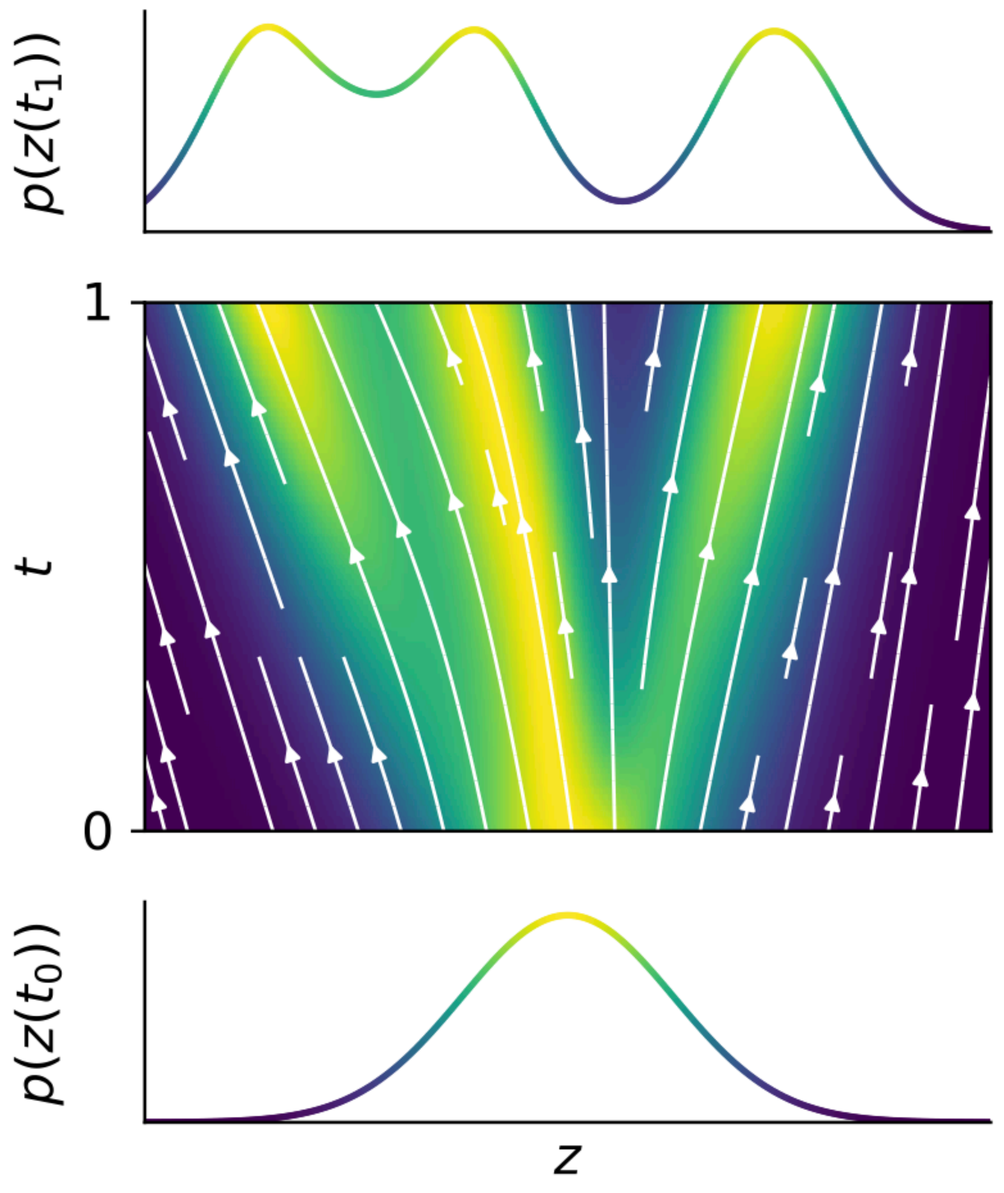
Target: $\pi(x)$

Reference: $\rho_{ref}(x) = \mathcal{N}(0,\mathbf{I})$

Find velocity field: $v = v_\theta$

$$\frac{dx}{dt} = v_\theta(x(t))$$

$$x(0) \sim \pi(x),\ x(T) \sim \rho_{ref}(x)$$



Grathwohl et al. '18

$$\min_\theta D_{KL}(\pi \| f_{\theta\sharp}\rho_0)$$

The 'usual' divergence

$$\iff \min_\theta \left\{ -\mathbb{E}_\pi \left[\log \rho_{ref}(x(0)) + \int_T^0 \nabla \cdot v_\theta(x(s), s)\, ds \right] : x(s) = x + \int_T^s v_\theta(x(s'), s')\, ds', x \sim \pi \right\}$$

Highlights in this talk

1. **Mean-field games** as a mathematical framework for **generative flows:**

- optimal control of **particle** dynamics + cost functions/distances to target π
- backward Hamilton-Jacobi **+** forward Transport PDE

See the posters by Hyemin Gu and Ben Zhang

2. **Model-form UQ + PDE regularity theory for generative flows**:

- Score-based, diffusion generative models are **robust**

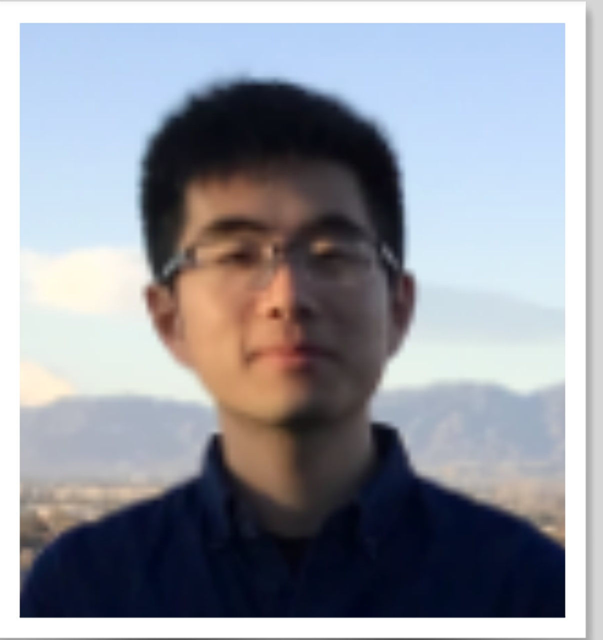
3. **Structure-informed learning:**

- Equivariance provably enhances generative algorithms

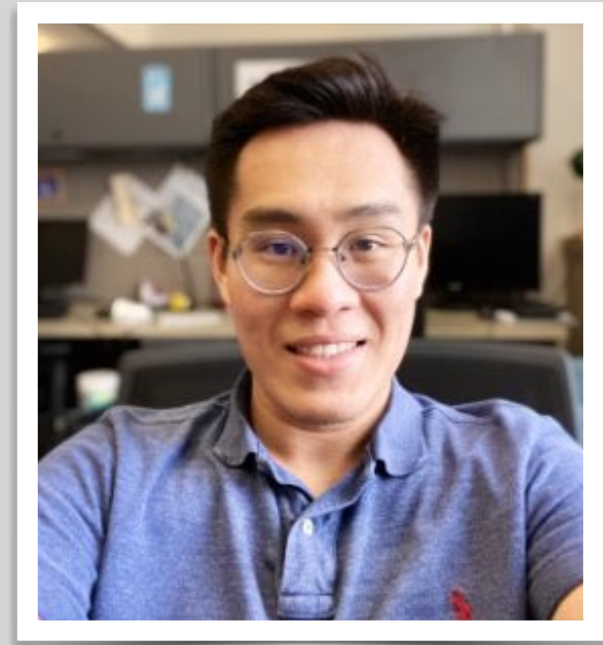
See poster by Ziyu Chen

L. Rey-Bellet

Hyemin Gu,
UMass Amherst

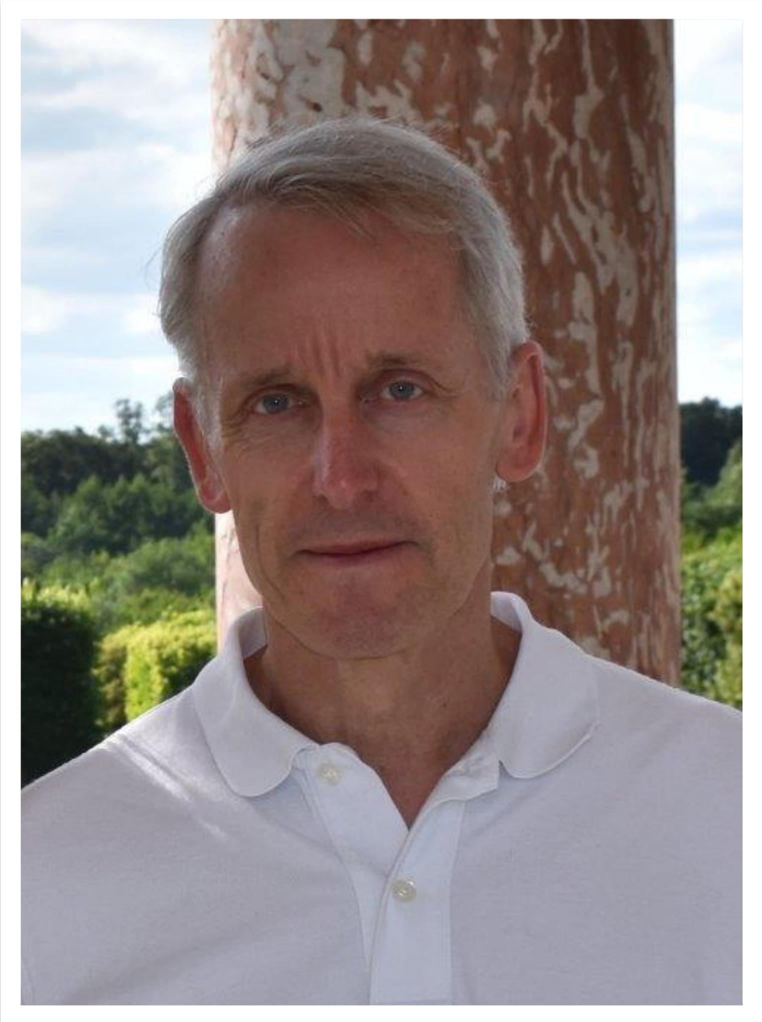


Ziyu Chen,
UMass Amherst



Benjamin Zhang, UMass
Amherst → Brown

N. Mimikos-
Stamatopoulos,
Université Côte d'Azur

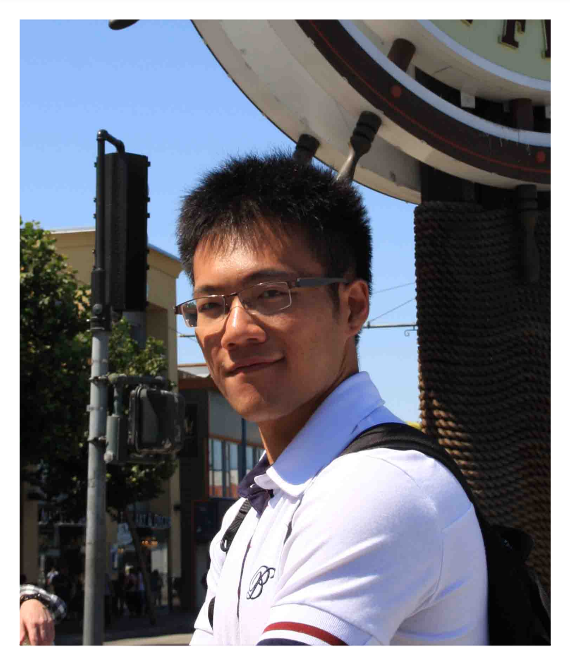


P. Dupuis

S. Osher, UCLA

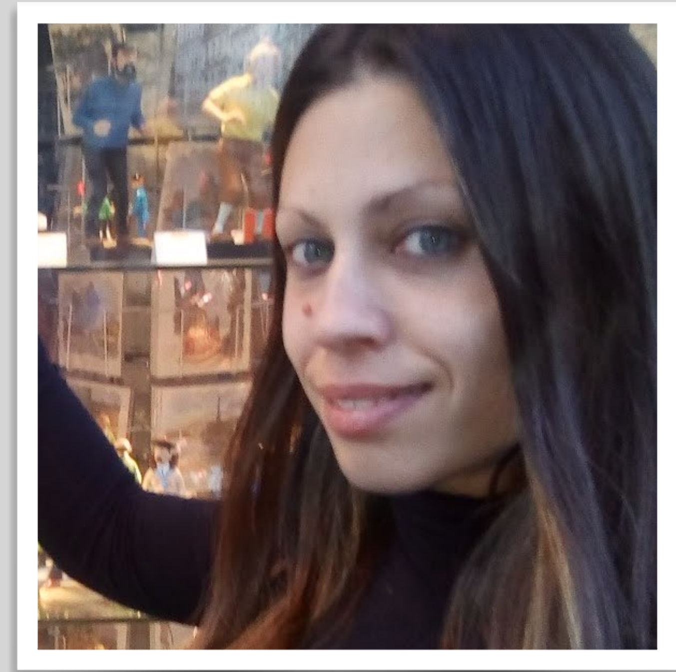
W. Li, U of SC

S. Liu, UC Riverside



W. Zhu,
GaTech

Jeremiah Birrell,
Texas State

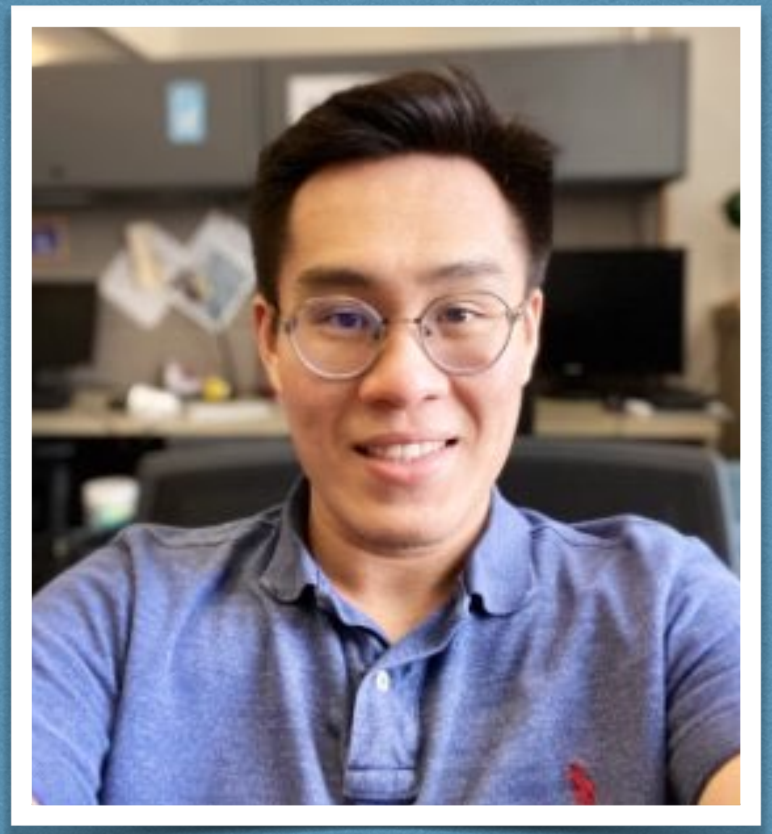


P. Birmpa,
Heriot-Watt, UK

Y. Pantazis,
FORTH, Greece

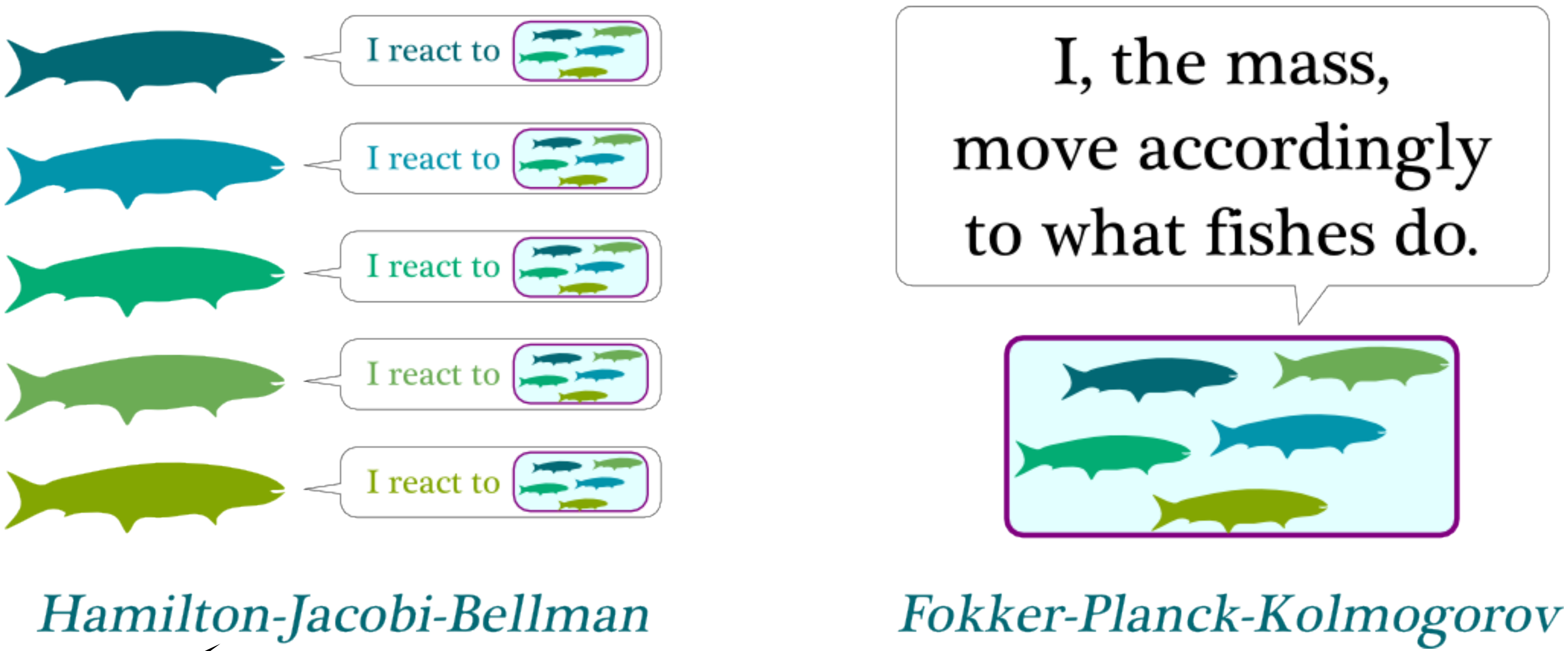
A Mean-field games laboratory for generative modeling:

Neural ODE flows & diffusion-based generative algorithms as MFG



Benjamin Zhang, Brown

Mean-field games



http://www.science4all.org/article/mean-field-games/

Optimal control

Transport PDE

Motivation

Mean-field games as a unifying mathematical framework

- **Explaining** — Understanding generative models in relation to each other
- **Enhancing** — MFGs inform exploitable mathematical structure
- **Inventing** — A laboratory for experimenting with new models

---

- **Normalizing flows** as solutions of MFGs
- **Score-based generative models** as solutions of MFGs

Mean-field games

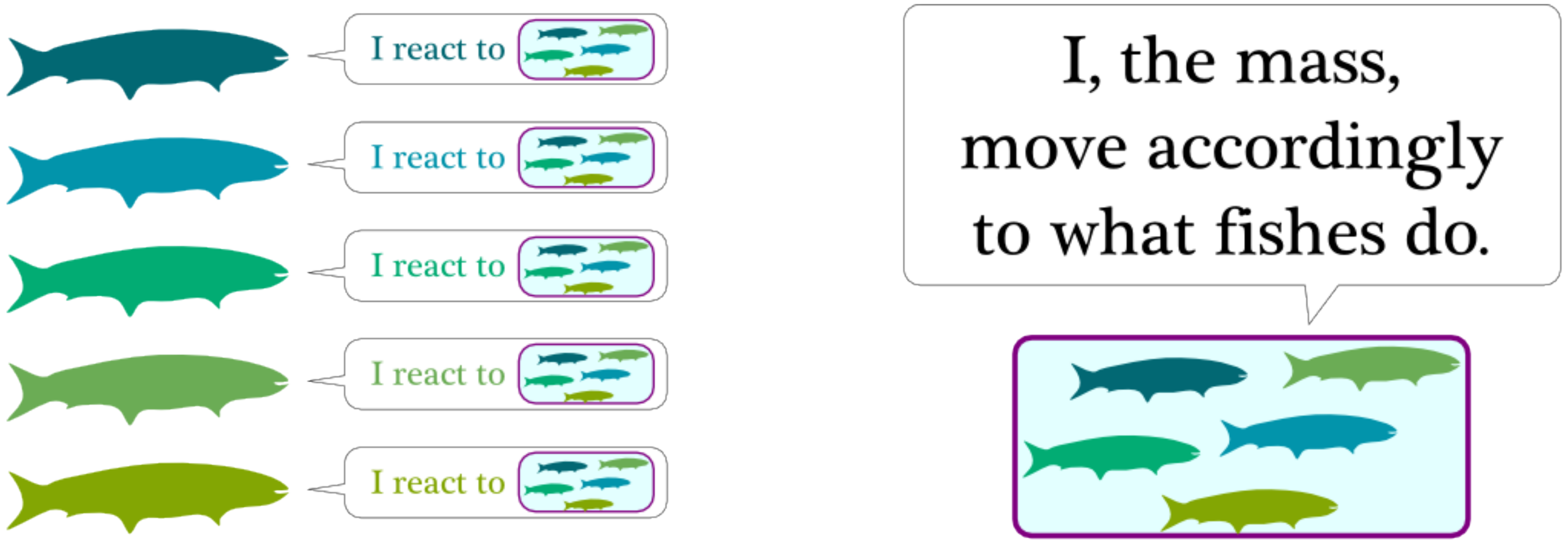
$$dx(t) = v(x(t), t)\, ds + \sigma(x(t), t)\, dW(t)$$
$$x(0) = x_0$$

Agent dynamics

$$\partial_t \rho + \nabla \cdot (v\rho) = \frac{\sigma^2}{2}\Delta\rho$$
$$\rho(x,0) = \rho_0(x)$$

Agent cost function

$$\min_{v,\rho} J_{x,t}(v,\rho) = \min_{v,\rho} \mathbb{E}\left[\underbrace{M(x(T), \rho(x(T), T))}_{\text{Terminal cost}} + \int_t^T \underbrace{I(x(s), \rho(x(s), s))}_{\text{Interaction cost}} + \underbrace{L(x(s), v(x(s), s))}_{\text{Running cost}} ds\right]$$



Potential Mean-field games and optimality conditions

$$\inf_{v,\rho} \left\{ \underbrace{\mathcal{M}(\rho(\,\cdot\,,T))}_{\text{Terminal cost}} + \int_0^T \underbrace{\mathcal{I}(\rho(\,\cdot\,,t))}_{\text{Interaction cost}} dt + \int_0^T \int_{\mathbb{R}^d} \underbrace{L(x,v)}_{\text{Running cost}} \rho(x,t)\,dx\,dt \right\}$$

$$\text{s.t. } \partial_t \rho + \nabla \cdot (v\rho) = \frac{\sigma^2}{2}\Delta\rho, \qquad \rho(x,0) = \rho_0(x)$$

[Lasry & Lions '07]

Optimality conditions characterize solution!

Hamiltonian

$$H(x,p) = \sup_v -p^\top v - L(x,v)$$

Optimal velocity field

$$\boxed{v^*(x,t) = -\nabla_p H(x,\nabla U)}$$

$$\partial_t \rho - \nabla \cdot (\nabla_p H(x,\nabla U)\rho) = \frac{\sigma^2}{2}\Delta\rho \qquad \text{Fokker-Planck}$$

$$-\partial_t U + H(x,\nabla U) - \frac{\sigma^2}{2}\Delta U = \frac{\delta \mathcal{I}}{\delta \rho}(x,\rho(x,t)) \qquad \text{Hamilton-Jacobi-Bellman}$$

$$U(x,T) = \frac{\delta \mathcal{M}}{\delta \rho}(x,\rho(\,\cdot\,,T)), \qquad \rho(x,0) = \rho_0(x)$$

Why mean-field games for generative modeling?

Training flow-based models looks like solving MFGs

$$\min_v \overbrace{\mathcal{D}(\pi, \rho_v(\,\cdot\,, T))}^{\text{Loss}} + \mathbb{E}\left[\overbrace{\int_0^T L(x, v)\, dt}^{\text{Transport cost}}\right]$$

Examples: KL Divergence, Cross-entropy, f-divergences, Jensen-Shannon

Example: Optimal transport cost
See [Onken et al. '21]
Or L = 0

Generative model	s.t. $dx(t) = v(x(t), t)\, dt + \sigma\, dW(t)$
Reference distribution	$x(0) \sim \rho_0$
Target distribution	π

Generative models as solutions to MFGs

Generative modeling

Optimization problem

$$\min_v \mathscr{D}(\pi, \rho_v(\cdot, T)) + \mathbb{E}\left[\int_0^T L(x,v)\, dt\right]$$

$$dx(t) = v(x(t), t)\, dt + \sigma\, dW(t)$$

$$x(0) \sim \rho_0$$

Generative model

$$dx(t) = v^*(x(t), t)\, dt + \sigma\, dW(t)$$

$$x(0) \sim \mathcal{N}(0, \mathbf{I})$$

$$x(T) \sim \pi$$

Generative models as solutions to MFGs

Mean-field game

$$\inf_{v,\rho} \left\{ \mathcal{M}(\rho(\cdot,T)) + \int_0^T \mathcal{I}(\rho(\cdot,t))\,dt + \int_0^T \int_{\mathbb{R}^d} L(x,v)\rho(x,t)\,dx\,dt \right\}$$

$$\partial_t \rho + \nabla\cdot(v\rho) = \frac{\sigma^2}{2}\Delta\rho \qquad \rho(x,0) = \rho_0(x)$$

Generative modeling

Optimization problem

$$\min_v \mathscr{D}(\pi, \rho_v(\,\cdot\,,T)) + \mathbb{E}\left[\int_0^T L(x,v)\,dt\right]$$

$$dx(t) = v(x(t),t)\,dt + \sigma\,dW(t)$$

$$x(0) \sim \rho_0$$

Generative model

$$dx(t) = v^*(x(t),t)\,dt + \sigma\,dW(t)$$

$$x(0) \sim \mathcal{N}(0,\mathbf{I})$$

$$x(T) \sim \pi$$

Generative model structure

$$v^*(x,t) = \arg\sup_v \left[-v^\top \nabla U - L(x,v)\right] = -\nabla_p H(x,\nabla U)$$

- Common **backward-forward** structure
- **Backward** eq. determines **optimal velocity field**
- **Forward** eq. determines **generation**
- Applies to **all flow and diffusion-based** models

Optimality conditions & well-posedness

Hamiltonian

$$H(x,p) = \sup_v -p^\top v - L(x,v)$$

Forward in time: **Fokker-Planck**

$$\partial_t \rho - \nabla\cdot(\nabla_p H(x,\nabla U)\rho) = \frac{\sigma^2}{2}\Delta\rho$$

$$\rho(x,0) = \rho_0(x)$$

Generator

Coupled PDEs

Backward in time: **Hamilton-Jacobi-Bellman**

$$-\partial_t U + H(x,\nabla U) - \frac{\sigma^2}{2}\Delta U = \frac{\delta \mathcal{I}}{\delta\rho}$$

$$U(x,T) = \frac{\delta\mathcal{M}}{\delta\rho}(\rho(\cdot,T))$$

Identifies velocity field

Optimal velocity field

$$v^*(x,t) = \arg\sup_v \left[-v^\top \nabla U - L(x,v)\right] = -\nabla_p H(x,\nabla U)$$

Exhibit A: *Continuous* normalizing flows

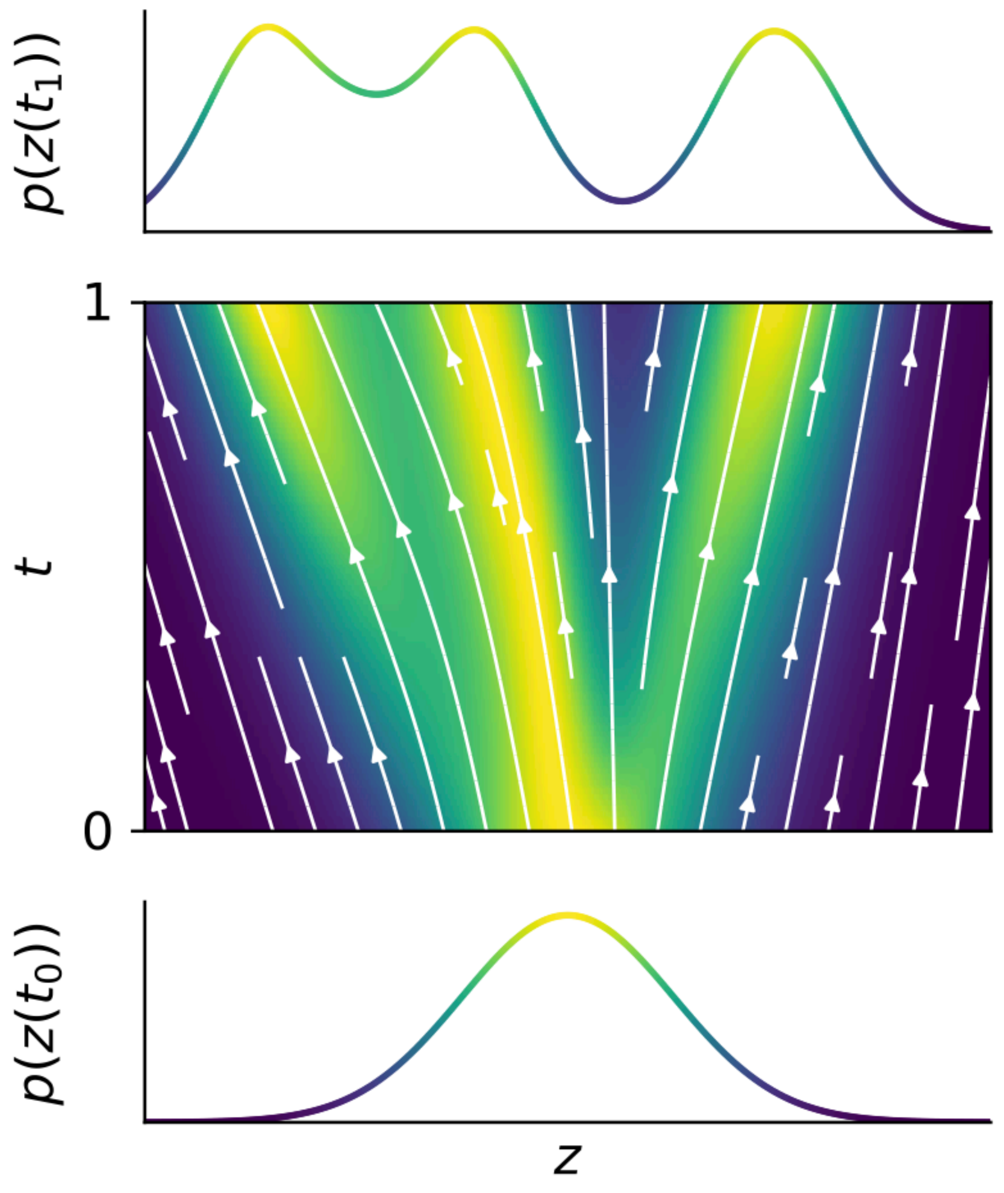
Target: $\pi(x)$

Reference: $\rho_{ref}(x) = \mathcal{N}(0, \mathbf{I})$

Find velocity field: $v = v_\theta$

$$\frac{dx}{dt} = v_\theta(x(t))$$

$$x(0) \sim \pi(x),\ x(T) \sim \rho_{ref}(x)$$



Grathwohl et al. '18

$$\min_\theta D_{KL}(\pi \| f_{\theta\sharp}\rho_0)$$

The 'usual' divergence

$$\iff \min_\theta \left\{ -\mathbb{E}_\pi \left[\log \rho_{ref}(x(0)) + \int_T^0 \nabla \cdot v_\theta(x(s), s)\, ds \right] : x(s) = x + \int_T^s v_\theta(x(s'), s')\, ds', x \sim \pi \right\}$$

CNFs trained with KL divergence are ill-posed: discretization-dependent
Sensitive to parametrization!

Continuous normalizing flows are ill-posed

Well-noted that CNFs are ill-posed. Study CNFs as MFG

$$\inf_{v,\rho} \left\{ \overbrace{\mathcal{M}(\rho(\,\cdot\,,T))}^{\text{Terminal cost}} + \int_0^T \overbrace{\mathcal{I}(\rho(\,\cdot\,,t))}^{\text{Interaction cost}} dt + \int_0^T \int_{\mathbb{R}^d} \overbrace{L(x,v)}^{\text{Running cost}} \rho(x,t)\,dx\,dt \right\}$$

s.t. $\partial_t \rho + \nabla \cdot (v\rho) = \dfrac{\sigma^2}{2}\Delta\rho, \qquad \rho(x,0) = \pi(x)$

$$\inf_{v,\rho} \left\{ \mathcal{D}_{KL}(\rho(\,\cdot\,,T) \| \overbrace{\rho_{ref}}^{\text{Reference distribution}}) \right\}$$

s.t. $\partial_t \rho + \nabla \cdot (v\rho) = 0, \qquad \rho(x,0) = \pi(x)$

Hamiltonian is degenerate!

$$\begin{aligned} H(x,p) &= \sup_v -p^\top v - L(x,v) \\ &= \sup_v -p^\top v \\ &= \infty \text{ if } p \neq 0 \end{aligned}$$

$H(x,p) = 0$ if $p = 0$

Continuous normalizing flow as an MFG

Additional constraints yield well-posedness & math structure!

Option 1: Bound the set of feasible velocities

$$\inf_{v,\rho} \left\{ \mathscr{D}_{KL}(\rho(\cdot, T) \| \rho_{ref}) : \|v\| \leq c \right\}$$

s.t. $\partial_t \rho + \nabla \cdot (v\rho) = 0, \qquad \rho(x,0) = \pi(x)$

$$\partial_t \rho - c\nabla \cdot \left(\rho \frac{\nabla U}{\|\nabla U\|} \right) = 0$$

Fokker-Planck

$$-\partial_t U + c\|\nabla U\| = 0$$

HJB: A **level set** equation!

$$U(x,T) = 1 + \log \frac{\rho(x,T)}{\rho_{ref}(x)}, \qquad \rho(x,0) = \pi(x)$$

Hamiltonian

$$H(x,p) = \sup_{\|v\|<c} -p^\top v$$

$$= c\|p\|$$

Optimal velocity field

$$v^*(x,t) = -c\frac{\nabla U(x,t)}{\|\nabla U(x,t)\|}$$

Continuous normalizing flow as an MFG

Additional regularizations yield well-posedness and structure

Option 2: Optimal transport cost *[Onken et al. '21]*

$$\inf_{v,\rho} \left\{ \mathcal{D}_{KL}(\rho(\,\cdot\,,T)\|\rho_{ref}) + \int_0^T \int_{\mathbb{R}^d} \frac{1}{2}\|v(x,t)\|^2 \rho(x,t)\,dx\,dt \right\}$$

s.t. $\partial_t \rho + \nabla \cdot (v\rho) = 0, \qquad \rho(x,0) = \pi(x)$

$$\partial_t \rho - \nabla \cdot \left(\rho \nabla U\right) = 0$$

Fokker-Planck

$$-\partial_t U + \frac{1}{2}\|\nabla U\|^2 = 0$$

HJB

$$U(x,T) = 1 + \log \frac{\rho(x,T)}{\rho_{ref}(x)}, \quad \rho(x,0) = \pi(x)$$

Hamiltonian

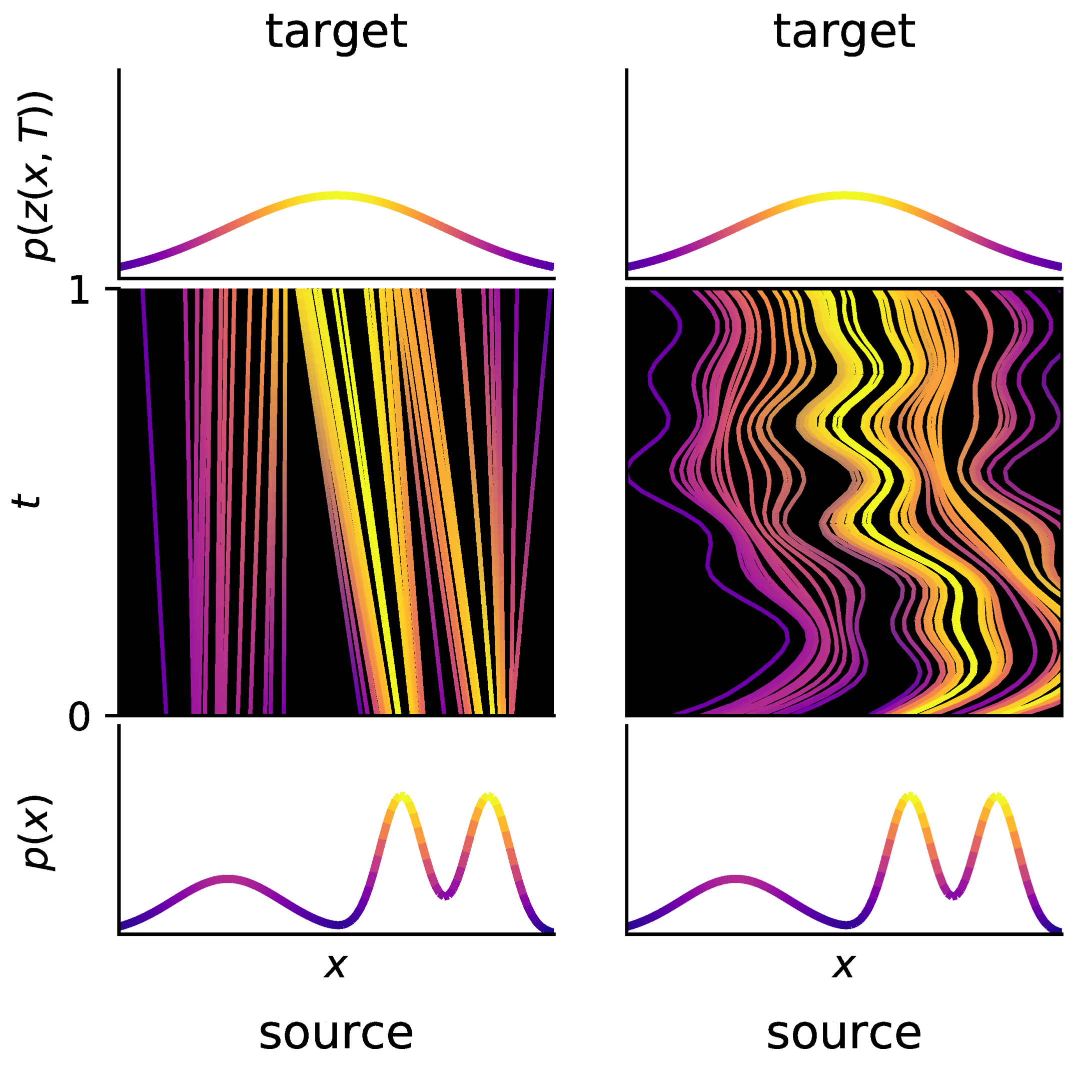
$$H(x,p) = \sup_v -p^\top v - \frac{1}{2}\|v\|^2 = \frac{1}{2}\|p\|^2$$

Optimal velocity field

$$v^*(x,t) = -\nabla U(x,t)$$

Mathematical structure of CNFs

- **Mean-field games provide well-posedness and structure to normalizing flows**
- **Well-posedness of NF training tied to well-posedness of Hamilton-Jacobi**
- **Empirically observed and explained via an Optimal Transport argument, Finlay et al, 2021:**



(a) Optimal transport map (b) generic flow

Optimal transport regularization

Hamilton-Jacobi equation

$$-\partial_t U + \frac{1}{2}\|\nabla U\|^2 = 0$$

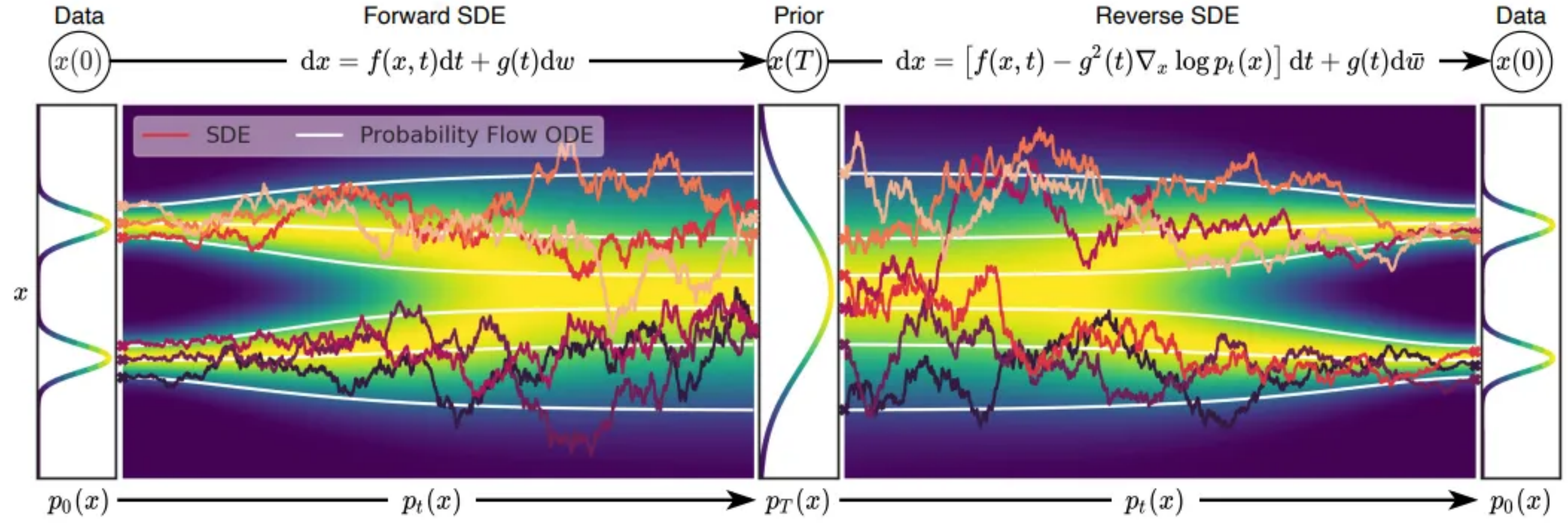
$$U(x,T) = 1 + \log \frac{\rho(x,T)}{\rho_{ref}(x)}$$

Optimal velocity field

$$v^*(x,t) = -\nabla U(x,t)$$

See talk by L. Rey-Bellet , poster by Hyemin Gu on a complete analysis & experiments using Wasserstein Proximals and MFG

Score-based generative modeling with SDEs



Song et al. '20

Two SDEs

Noising process

$$dY(s) = -f(Y(s), T-s)ds + \sigma(T-s)dW(s)$$

$$Y(0) \sim \pi$$

$$Y(s) \sim \eta(\,\cdot\,, s)$$

Denoising process

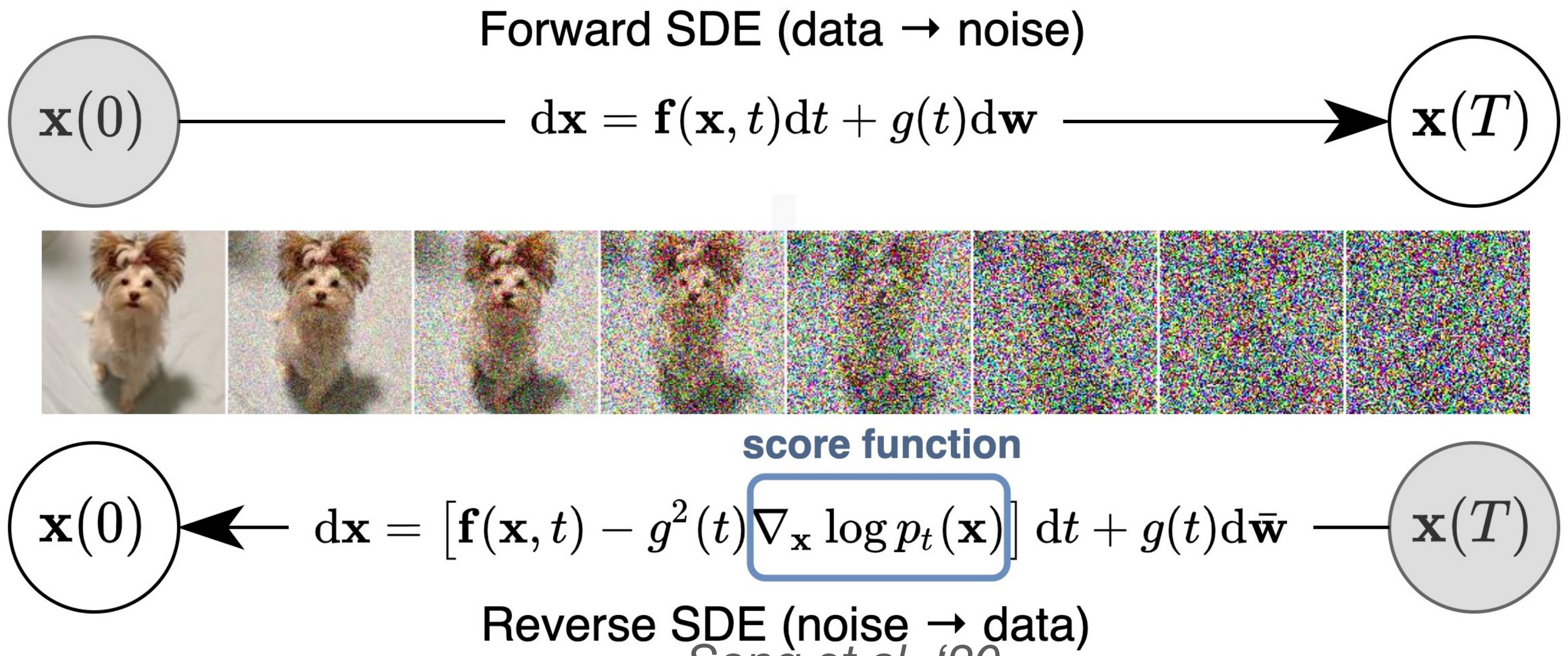
$$dX(t) = \left[f(X(t),t) + \sigma(t)^2 \nabla \log \eta(x,t)\right] dt + \sigma(t)dW(t)$$

$$X(0) \sim \eta(\,\cdot\,, T)$$

Score-matching

$$\min_\theta C_{ESM}(\theta) = \min_\theta \int_0^T \int_{\mathbb{R}^d} \frac{\sigma(T-s)^2}{2} \|s_\theta(y,s) - \nabla \log \eta(y,s)\|^2 \eta(y,s)\, dy\, ds$$

$$\min_\theta C_{ISM}(\theta) = \min_\theta \int_0^T \int_{\mathbb{R}^d} \sigma(T-s)^2 \left[\frac{1}{2}\|s_\theta(y,s)\|^2 + \nabla \cdot s_\theta(y,s)\right] \eta(y,s)\, dy\, ds$$



Song et al. '20

Exhibit B: SGM as an MFG

$$\inf_{v,\rho}\left\{\overbrace{\mathscr{M}(\rho(\cdot,T))}^{\text{Terminal cost}}+\int_0^T\overbrace{\mathscr{I}(\rho(\cdot,t))}^{\text{Interaction cost}}dt+\int_0^T\int_{\mathbb{R}^d}\overbrace{L(x,v)}^{\text{Running cost}}\rho(x,t)\,dx\,dt\right\}$$

s.t. $\partial_t\rho+\nabla\cdot(v\rho)=\frac{\sigma^2}{2}\Delta\rho,\qquad \rho(x,0)=\rho_0(x)$

$$\inf_{v,\rho}\left\{\underbrace{-\int\rho(x,T)\log\pi(x)dx}_{\text{Cross Entropy}}+\int_0^T\int_{\mathbb{R}^d}\left(\frac{1}{2}\|v\|^2-\nabla\cdot f\right)\rho(x,t)\,dx\,dt\right\}$$

s.t. $\partial_t\rho+\nabla\cdot((f+\sigma v)\rho)=\frac{\sigma^2}{2}\Delta\rho,\qquad \rho(x,0)=\eta(x,T)$

Cross Entropy

$$CE(\pi,\rho)=-\mathbb{E}_\rho[\log\pi]$$

$$=-\mathbb{E}_\rho[\log\rho]+\mathbb{E}_\rho\left[\log\frac{\pi}{\rho}\right]$$

= Entropy + KL Divergence

Wasserstein Proximal of Cross Entropy when f=0

SGM as an MFG: optimality conditions

$$\partial_t \rho + \nabla \cdot (\rho(f - \sigma^2 \nabla U)) = \frac{\sigma^2}{2}\Delta\rho$$

Controlled Fokker-Planck

$$-\partial_t U - f^\top \nabla U + \frac{1}{2}\|\sigma \nabla U\|^2 + \nabla \cdot f = \frac{\sigma^2}{2}\Delta U$$

Hamilton-Jacobi-Bellman

$$U(x, T) = -\log \pi(x), \quad \rho(x,0) = \eta(x, T)$$

Cole-Hopf

$$U(x,t) = -\log \eta(x, T-t)$$

$$\partial_t \rho + \nabla \cdot (\rho(f + \sigma^2 \nabla \log \eta)) = \frac{\sigma^2}{2}\Delta\rho$$

Controlled Fokker-Planck

$$\partial_s \eta + \nabla \cdot (-f\eta) = \frac{\sigma^2}{2}\Delta\eta$$

Uncontrolled Fokker-Planck

$$\eta(y,0) = \pi(y), \quad \rho(x,0) = \eta(x, T)$$

See also optimal control perspective [Berner et al. 2022]

SGM as an MFG: Noising SDE is a HJB

$$\partial_t \rho + \nabla \cdot ((f + \sigma^2 \nabla \log \eta)\rho) = \frac{\sigma^2}{2}\Delta \rho$$

Controlled Fokker-Planck

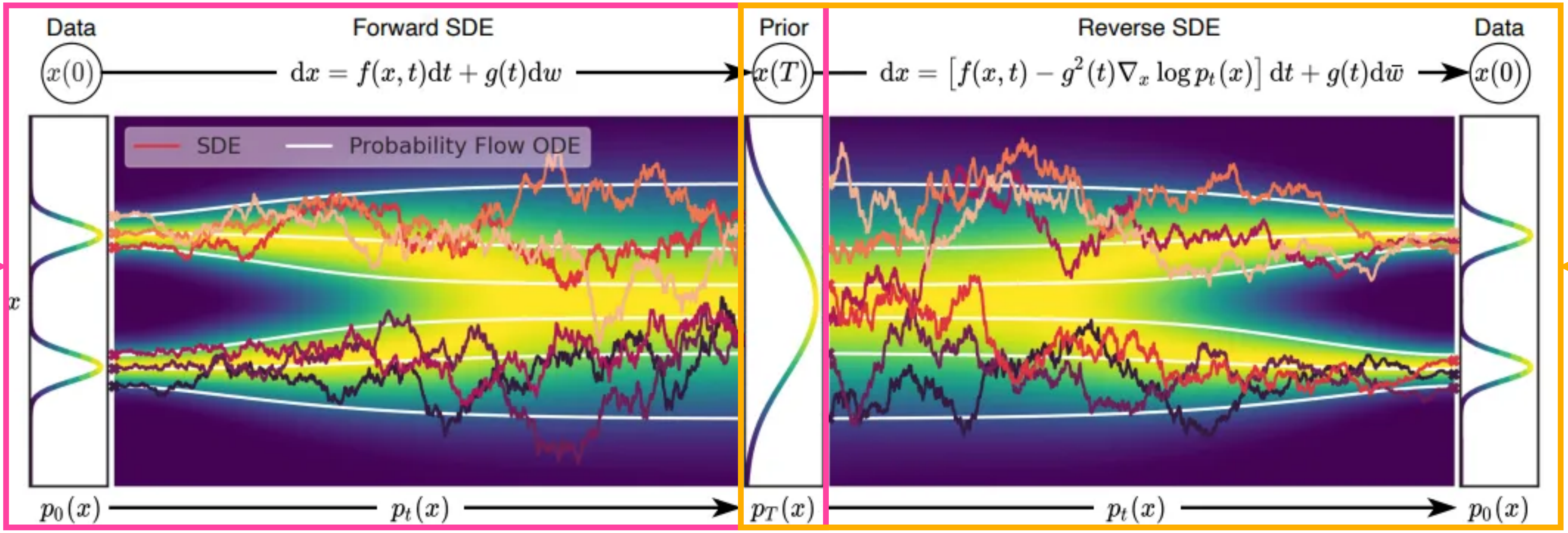
$$\partial_s \eta + \nabla \cdot (-f\eta) = \frac{\sigma^2}{2}\Delta \eta$$

Uncontrolled Fokker-Planck

(was HJB)

$$\eta(y,0) = \pi(y), \quad \rho(x,0) = \eta(x,T)$$

HJB equation hiding in plain sight



Fokker-Planck

Continuous normalizing flows vs score-based generative models

Less than meets the eye

$$\inf_{v,\rho}\left\{\underbrace{\mathscr{M}(\rho(\,\cdot\,,T))}_{\text{Terminal cost}}+\int_0^T\underbrace{\mathscr{I}(\rho(\,\cdot\,,t))}_{\text{Interaction cost}}dt+\int_0^T\int_{\mathbb{R}^d}\underbrace{L(x,v)}_{\text{Running cost}}\rho(x,t)\,dx\,dt\right\}$$

$$\text{s.t. } \partial_t\rho+\nabla\cdot(v\rho)=\frac{\sigma^2}{2}\Delta\rho,\qquad \rho(x,0)=\rho_0(x)$$

	$\mathscr{M}(\rho)$	$\mathscr{I}(\rho)$	$L(x,v)$	**Dynamics**
OT-Flow (Alternate formulation)	$\mathscr{D}_{KL}(\pi\|\rho)$	0	$\frac{1}{2}\|v\|^2$	$dx = v\,dt$
SGM via SDEs	$-\mathbb{E}_\rho\left[\log\pi\right]$	0	$\frac{1}{2}\|v\|^2-\nabla\cdot f$	$dx=(f+\sigma v)\,dt+\sigma dW$

SGM vs. Normalizing Flows: an optimality conditions comparison

Score-based Generative Model

$$\inf_{v,\rho}\left\{-\int \rho(x,T)\log\pi(x)dx + \int_0^T\int_{\mathbb{R}^d}\left(\frac{1}{2}\|v\|^2 - \nabla\cdot f\right)\rho(x,t)\,dx\,dt\right\}$$

s.t. $\partial_t\rho + \nabla\cdot((f+\sigma v)\rho) = \frac{\sigma^2}{2}\Delta\rho, \qquad \rho(x,0) = \eta(x,T)$

OT Normalizing flow (alternate form)

$$\inf_{v,\rho}\left\{\mathscr{D}_{KL}(\pi\|\rho(\,\cdot\,,T)) + \int_0^T\int_{\mathbb{R}^d}\frac{1}{2}\|v(x,t)\|^2\rho(x,t)\,dx\,dt\right\}$$

s.t. $\partial_t\rho + \nabla\cdot(v\rho) = 0, \qquad \rho(x,0) = \rho_0(x)$

Fokker-Planck/ transport PDE

$$\partial_t\rho + \nabla\cdot((f-\sigma^2\nabla U)\rho) = \frac{\sigma^2}{2}\Delta\rho$$

$$\partial_t\rho - \nabla\cdot\left(\rho\nabla U\right) = 0$$

Hamilton-Jacobi-Bellman

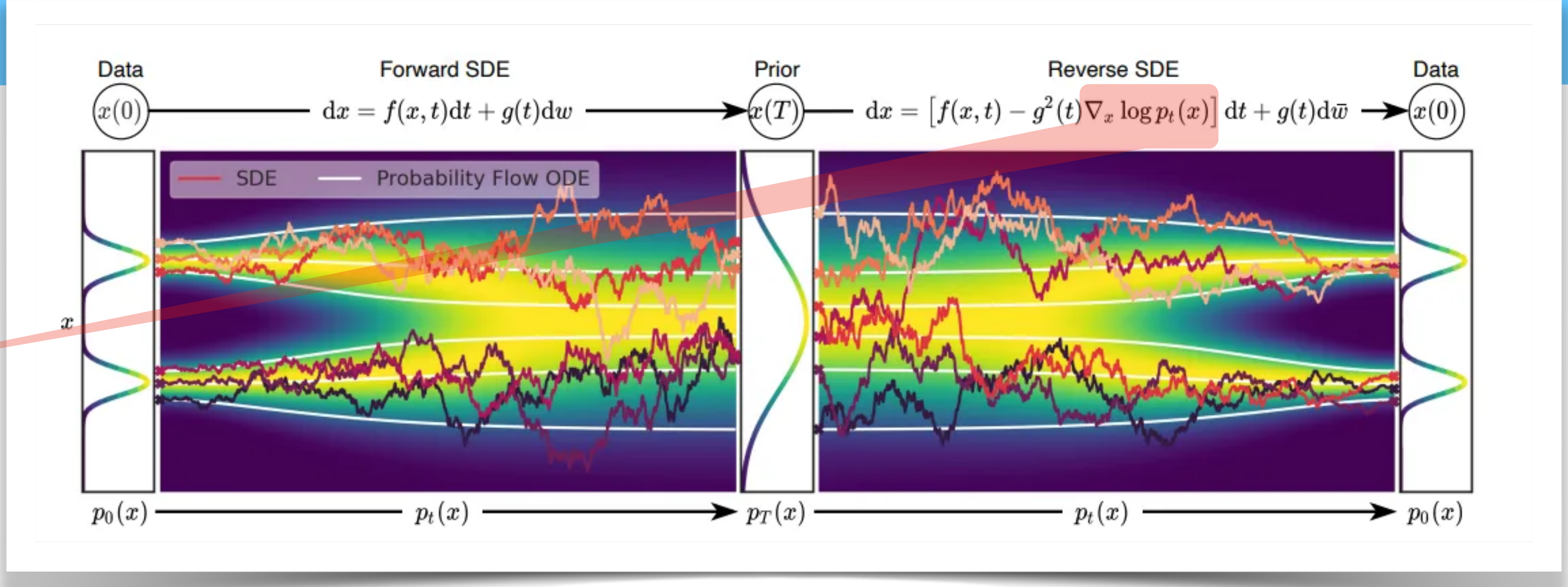
$$-\partial_t U - f^\top\nabla U + \frac{1}{2}\|\sigma\nabla U\|^2 + \nabla\cdot f = \frac{\sigma^2}{2}\Delta U$$

$$-\partial_t U + \frac{1}{2}\|\nabla U\|^2 = 0$$

initial/terminal conditions

$$U(x,T) = -\log\pi(x), \qquad \rho(x,0) = e^{-U(x,0)}$$

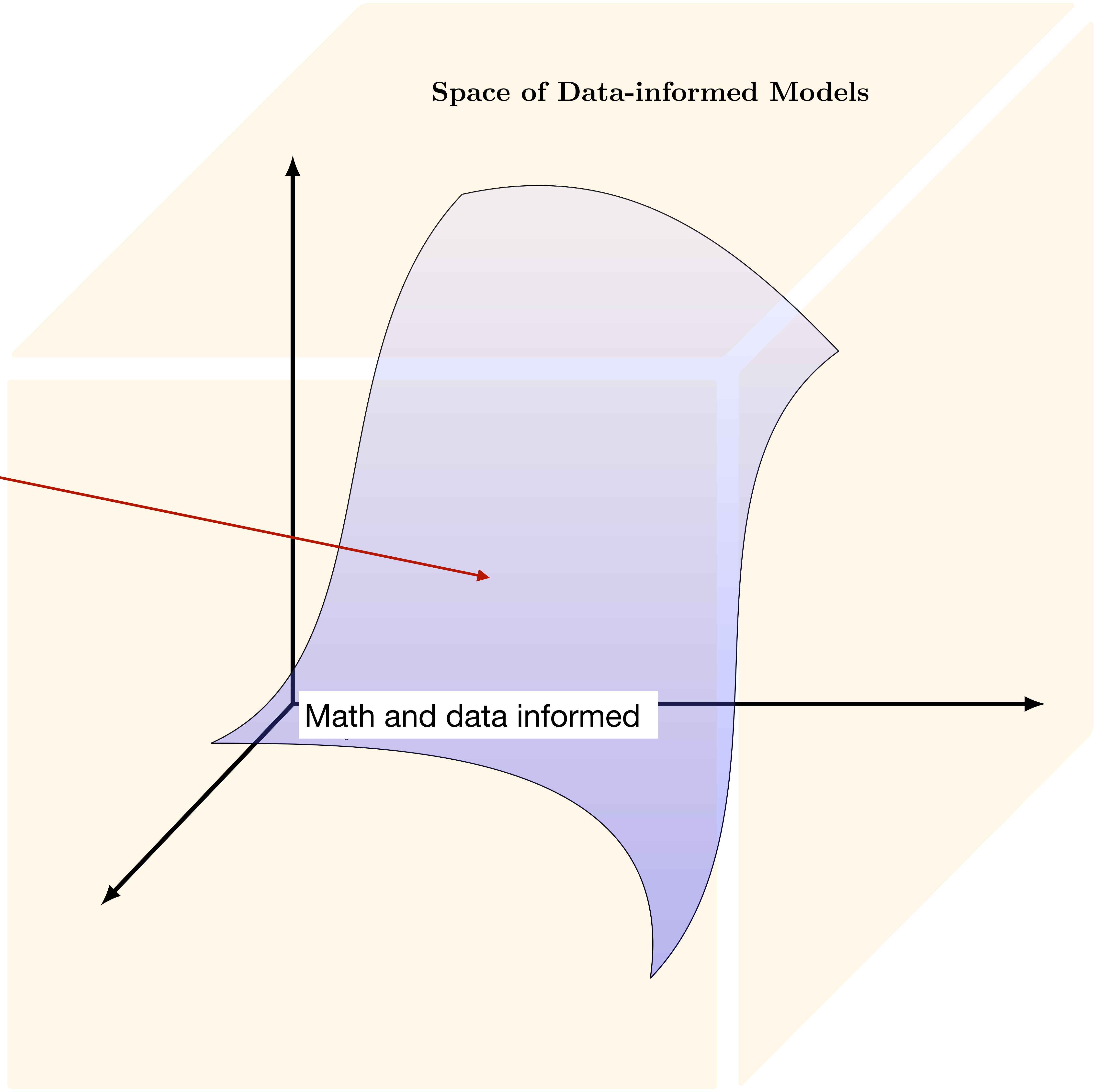
$$U(x,T) = -\frac{\pi(x)}{\rho(x,T)}, \qquad \rho(x,0) = \rho_0(x)$$



In SGM: HJB decouples from FP due to terminal condition and can be solved first, to provide the optimal velocity field for the FP (see reverse SDE)

MFG-informed generative models

MFG formulation describes
math structure
We learn in a restricted, more relevant space



Provably yields:
- Better generative models
- Better training objectives
- Better regularizers

See posters by **Hyemin Gu, Ben Zhang**

A modular mean-field games laboratory

	Mean-field game (11)					
Model	$\mathcal{M}(\rho)$	$\mathcal{I}(\rho)$	$L(x,v)$	**Dynamics**	$H(x,p)$	**Optimal** v^*
Continuous normalizing flow	$\mathcal{D}_{KL}(\pi\Vert\rho)$	0	0	$dx = v\,dt$	$\sup_{v\in K} -p^\top v$	$-\nabla_p H(x,\nabla U)$
Score-based generative modeling	$-\mathbb{E}_\rho\left[\log \pi\right]$	0	$\frac{\lvert v\rvert^2}{2} - \nabla\cdot f$	$dx = (f+\sigma v)\,dt$ $+\sigma\,dW_t$	$-f\cdot p + \frac{\sigma^2}{2}\lvert p\rvert^2$ $+\nabla\cdot f$	$\sigma\nabla\log\eta_{T-t}$
Score-based probability flow	$-\frac{1}{2}\mathbb{E}_\pi\left[\log\pi\right]$	$\mathbb{E}_\rho\left[\frac{\lvert\sigma\nabla\log\rho\rvert^2}{8}\right]$	$\frac{\lvert v\rvert^2}{2} - \frac{\nabla\cdot f}{2}$	$dx = (f+\sigma v)\,dt$	$-f\cdot p + \frac{\sigma^2}{2}\lvert p\rvert^2$ $+\nabla\cdot f$	$\frac{\sigma}{2}\nabla\log\eta_{T-t}$
Wasserstein gradient flow (WGF) ($\epsilon\to 0$)	$\mathcal{F}(\rho)e^{-T/\epsilon}$	$\frac{e^{-t/\epsilon}}{\epsilon}\mathcal{F}(\rho)$	$\frac{e^{-t/\epsilon}}{2}\lvert v\rvert^2$	$dx = v\,dt$	$\frac{1}{2}e^{t/\epsilon}\lvert p\rvert^2$	$-\nabla\frac{\delta\mathcal{F}}{\delta\rho}$
OT-Flow	$\mathcal{D}_{KL}(\pi\Vert\rho)$	0	$\frac{1}{2}\lvert v\rvert^2$	$dx = v\,dt$	$\frac{1}{2}\lvert p\rvert^2$	$-\nabla U$
Boltzmann generator	$\lambda\mathcal{D}_{KL}(\pi\Vert\rho)$ $+(1-\lambda)\mathcal{D}_{KL}(\rho\Vert\pi)$	0	0	$dx = v\,dt$	$\sup_{v\in K} -p^\top v$	$-\nabla_p H(x,\nabla U)$
Schrödinger bridge	$\rho=\pi$	0	$\frac{1}{2}\lvert v\rvert^2$	$dx = \sigma v\,dt + \sigma\,dW_t$	$\frac{1}{2}\lvert p\rvert^2$	$-\sigma\nabla U$
Generalized Schrödinger bridge	$\rho=\pi$	$\mathcal{I}(x,\rho)$	$\frac{1}{2}\lvert v\rvert^2$	$dx = \sigma v\,dt + \sigma\,dW_t$	$\frac{1}{2}\lvert p\rvert^2$	$-\sigma\nabla U$
HJB-regularized SGM	$-\mathbb{E}_\rho[\log\pi]$	0	$\frac{\lvert v\rvert^2}{2} - \nabla\cdot f$	$dx = (f+\sigma v)\,dt$ $+\sigma\,dW_t$	$-f\cdot p + \frac{\sigma^2}{2}\lvert p\rvert^2$ $+\nabla\cdot f$	$\sigma\nabla\log\eta_{T-t}$
Stochastic OT normalizing flow	$\mathcal{D}_{KL}(\pi\Vert\rho)$	0	$\frac{1}{2}\lvert v\rvert^2$	$dx = \sigma v\,dt + \sigma\,dW_t$	$\frac{1}{2}\lvert p\rvert^2$	$-\sigma\nabla U$
OT-Boltzmann generator	$\lambda\mathcal{D}_{KL}(\pi\Vert\rho)$ $+(1-\lambda)\mathcal{D}_{KL}(\rho\Vert\pi)$	0	$\frac{1}{2}\lvert v\rvert^2$	$dx = v\,dt$	$\frac{1}{2}\lvert p\rvert^2$	$-\nabla U$
Generalized OT-Flow (Relaxed WGF $\epsilon>0$)	$\mathcal{F}(\rho)e^{-T/\epsilon}$	$\frac{e^{-t/\epsilon}}{\epsilon}\mathcal{F}(\rho)$	$\frac{e^{-t/\epsilon}}{2}\lvert v\rvert^2$	$dx = v\,dt$	$\frac{1}{2}e^{t/\epsilon}\lvert p\rvert^2$	$-\nabla U$
Build your own generative model	Choose your	own cost	functions and	dynamics here	?	?

See full chart: arXiv:2304.1353!

Experiment with your own algorithm here...

Successful generative flows & diffusions are mean-field games

Generative modeling benefits from PDE analysis

Common **backward-forward** mathematical structure

Backward equation determines optimal velocity field

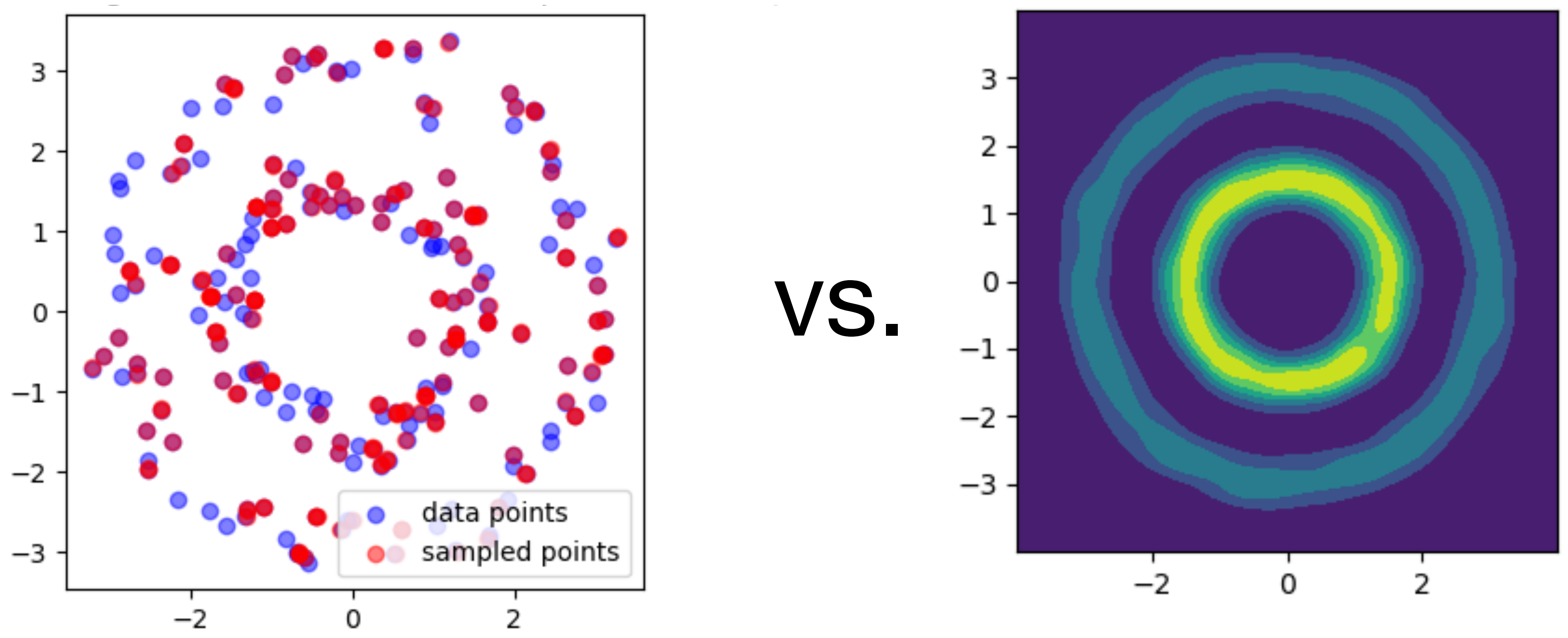
Forward equation determines generation

Applies to **all flow and diffusion-based models**

Other related topics & recent extensions

- **Score probability flow** as an MFG — Fisher information as interaction
- Learn **robustly** distributions on manifolds via **MFG & Wasserstein proximals, see poster by Hyemin Gu**
- **SGM MFG approximates Wasserstein proximal operator, poster by Ben Zhang**

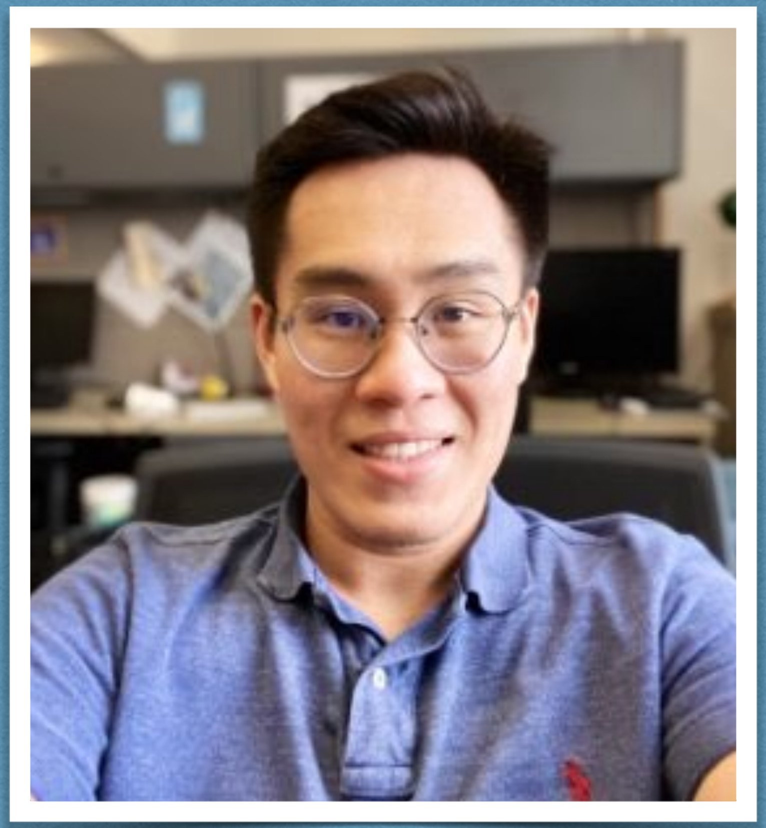
Wasserstein proximal operators describe SGMs and resolve memorization



Ben Zhang, Siting Liu, Wuchen Li, M.K., & Stan Osher

Score-based generative models are provably robust: a UQ perspective

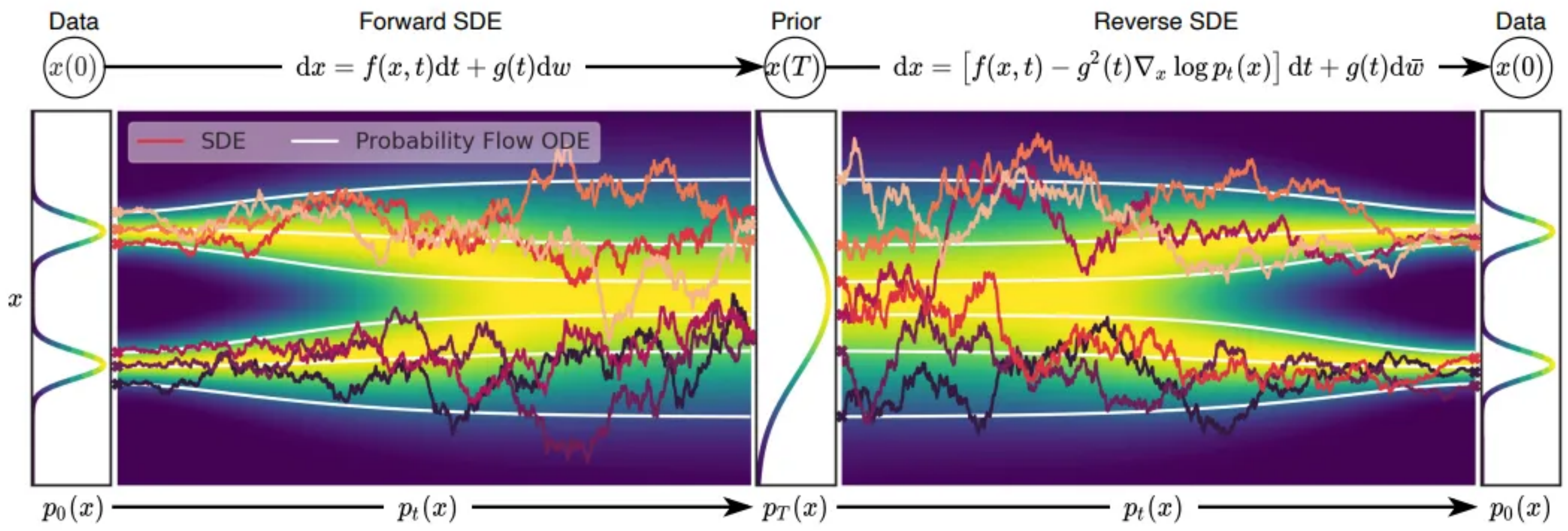
Nikiforos Mimikos-Stamatopoulos, UChicago → Université Côte d'Azur



Benjamin Zhang, Brown

Mimikos-Stamotopoulos, N., Zhang, B. J., & Katsoulakis, M. A. (2024). arXiv preprint arXiv:2405.15754

Score-based generative modeling with SDEs



Song et al. '20

Two SDEs

Noising process

$$dY(s) = -f(Y(s), T-s)ds + \sigma(T-s)dW(s)$$
$$Y(0) \sim \pi$$
$$Y(s) \sim \eta(\,\cdot\,, s)$$

Denoising process

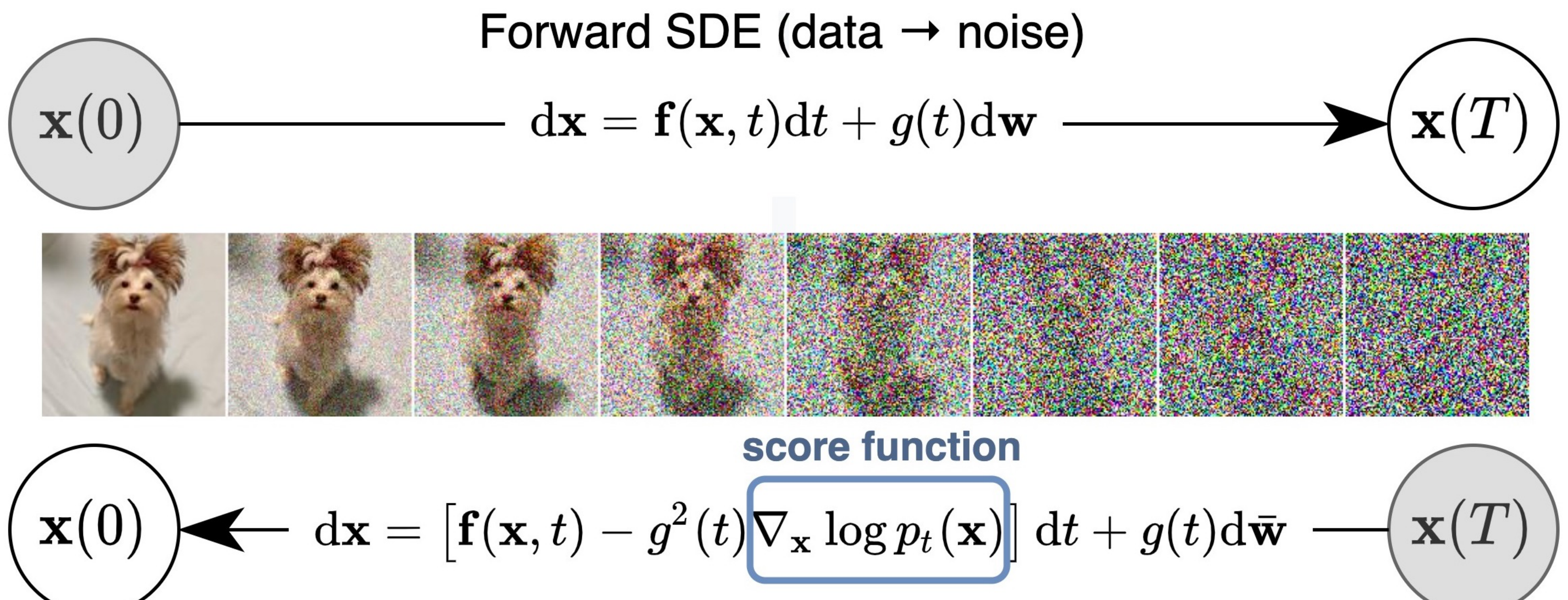
$$dX(t) = \left[f(X(t), t) + \sigma(t)^2 \nabla \log \eta(x, t)\right] dt + \sigma(t)dW(t)$$
$$X(0) \sim \eta(\,\cdot\,, T)$$

Different score-matching objectives

$$\min_\theta C_{ESM}(\theta) = \min_\theta \int_0^T \int_{\mathbb{R}^d} \frac{\sigma(T-s)^2}{2} \|s_\theta(y,s) - \nabla \log \eta(y,s)\|^2 \eta(y,s)\, dy\, ds$$

$$\min_\theta C_{ISM}(\theta) = \min_\theta \int_0^T \int_{\mathbb{R}^d} \sigma(T-s)^2 \left[\frac{1}{2}\|s_\theta(y,s)\|^2 + \nabla \cdot s_\theta(y,s)\right] \eta(y,s)\, dy\, ds$$

$$\min_\theta C_{DSM}(\theta) = \min_\theta \int_0^T \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \sigma(T-s) \|s_\theta(y,s) - \nabla \log \eta(y, s\,|\,y')\|^2 \eta(y, s\,|\,y') \pi(y')dy\, dy'\, ds$$



Song et al. '20

Errors of SGM

K. Chowdhary and **P. Dupuis**, *Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification*, ESAIM: M^2NA (2013)

- Finite sample error e_1
- Choice of score-matching objective e_2
- Score function approximation e_3
- Reference measure e_4
- Early stopping e_5
- Discretization error e_6

Research question:

How well does generative distribution $m_g(T)$ approximate data distribution π?

$$\text{"}\mathbf{d}(m_g(T), \pi) \le \mathscr{F}(e_1, e_2, e_3, e_4, e_5)\text{"}$$

Integral probability metric (IPM): $\mathbf{d}(\nu_1, \nu_2) = \sup_{\psi \in \mathscr{X}} \int \psi(x)\, d(\nu_1 - \nu_2)$

E.g., Wasserstein-1, $\mathscr{X} = \{\psi : \Omega \to \mathbb{R}, \|\nabla \psi\|_\infty \le 1\}$

Model form uncertainty quantification for SGMs

Wasserstein Uncertainty Propagation (WUP) theorem (partial)

Two SDEs with drifts b^1 and b^2 on domain $\Omega = R\mathbb{T}^d$

$$\partial_t m^1 - \nabla \cdot (m^1 b^1) = \Delta m^1, m^1(0) = m_1 \qquad \partial_t m^2 - \nabla \cdot (m^2 b^2) = \Delta m^2, m^2(0) = m_2$$

If $L^2(m^2)$ error is bounded

Model form error

$$\|b^2 - b^1\|^2_{L^2(m^2)} = \int_0^T \int_\Omega \|b^2(x,t) - b^1(x,t)\|^2 m^2(t,x)\, dx\, dt \leq \varepsilon^2$$

Then the Wasserstein distance between distributions $m^1(T)$ and $m^2(T)$ is bounded

$$\mathbf{d}_1(m^2(T), m^1(T)) \leq CR^{3/2}\left(1 + \sqrt{\|\nabla b^1\|_\infty}\right)(\mathbf{d}_1(m_1, m_2) + \varepsilon)$$

Direct bound for Wasserstein-1 without appealing to KL divergence!

Robustness under explicit score matching

Application of WUP to SGM with explicit score matching

Data distribution $\pi \in \mathscr{P}(\Omega)$ on domain $\Omega = R\mathbb{T}^d$

Two SDEs: True drift $\nabla \log \eta^\pi$ and approximate drift: $\mathbf{b}_\theta = \mathbf{s}_\theta(T - t, x)$

$$\partial_t m_g - \nabla \cdot (m_g \mathbf{b}_\theta) = \Delta m_g, m_g(0) = \frac{1}{vol(R\mathbb{T}^d)}$$ Denoising process

Model form error

If ESM error is e_{nn} : $\displaystyle\int_0^T \int_{\mathbb{R}^d} \|\mathbf{s}_\theta(s, y) - \nabla \log \eta(y, s)\|^2 \eta(y, s)\, dy\, ds < e_{nn}$, then

$$\mathbf{d}_1(\pi, m_g(T)) \leq CR^{3/2}\left(1 + \sqrt{\|\nabla \mathbf{s}_\theta\|_\infty}\right)\left(Re^{-\frac{\omega T}{R^2}}\mathbf{d}_1\left(\pi, \frac{1}{vol(R\mathbb{T}^d)}\right) + \sqrt{e_{nn}}\right)$$

Choice of reference measure

ESM error

Direct bound for Wasserstein-1 without appealing to KL divergence!

Robustness under denoising score matching

Main result

Assume score $\mathbf{s}_\theta$ learned via DSM with early stopping, which provides density lower bound

$$\pi^\epsilon > \delta, \hat{\pi}^{N,\epsilon} > \delta$$

Model form error

$$C^N_{DSM}(\theta) = \int_\epsilon^T \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \sigma(T-s) \| s_\theta(y,s) - \nabla \log \eta(y,s\,|\,y') \|^2 \eta(y,s\,|\,y') \hat{\pi}^N(y') dy\, dy'\, ds < e_{nn}$$

Early stopping

Choice of reference measure

Then,

$$\mathbf{d}_1(\pi, m_g(T)) \lesssim \sqrt{\epsilon} + R^{3/2}\left(1 + \sqrt{\|\nabla \mathbf{s}_\theta\|_\infty}\right)\left(Re^{-\frac{\omega T}{R^2}} \mathbf{d}_1\left(\pi, \frac{1}{vol(R\mathbb{T}^d)}\right) + \sqrt{e'_{nn}} \right),$$

where

Choice of score matching objective

$$\sqrt{e'_{nn}} \lesssim \sqrt{e_{nn}} + \sqrt{\left(1 + \frac{|\log(\delta)|}{\sqrt{\epsilon}} + T\|\mathbf{s}_\theta\|^2_{C^2([0,T]\times\Omega)}\right) \mathbf{d}_1(\pi^N, \pi)}.$$

DSM error

Finite sample error

Direct bound for IPMs without appealing to KL divergence!

DSM to ESM bridge

Regularity theory of HJB PDEs enables UQ in SGMs

Main ideas of Wasserstein Uncertainty Propagation proof

Step 1: Kolmogorov backward equation determines suitable test functions

Step 2: Integral probability metrics bounds depend on gradient estimates

Step 3: Bernstein estimates from HJB equations provide gradient estimates

- Regularizing test functions allows us to bound stronger TV norm with a weaker Wasserstein-1 norm.

Regularity theory of HJB PDEs enables UQ in SGMs

Main ideas of Wasserstein Uncertainty Propagation proof

Step 1: Kolmogorov backward equation (KBE) determines suitable test functions

$$\partial_t m^1 - \nabla \cdot (m^1 b^1) = \Delta m^1, m^1(0) = m_1 \qquad \partial_t m^2 - \nabla \cdot (m^2 b^2) = \Delta m^2, m^2(0) = m_2$$

$\lambda = m^1 - m^2$ satisfies

Difference of measures evolution

$$\partial_t \lambda - \Delta \lambda - \nabla \cdot (\lambda b^1 + m^2(b^1 - b^2)) = 0 \text{ in } (0,T) \times \Omega, \quad \lambda(0) = m_2 - m_1 \text{ in } \Omega\,.$$

Integrate against a test function $\phi(t,x)$ that satisfies KBE with terminal condition $\psi \in \mathcal{X}$

$$-\partial_t \phi - \Delta \phi + b^1 \cdot \nabla \phi = 0 \text{ in } [0,T) \times \Omega, \quad \phi(T,x) = \psi(x) \text{ in } \Omega$$

Regularity theory of HJB PDEs enables UQ in SGMs

Main ideas of Wasserstein Uncertainty Propagation proof

Step 2: Integral probability metrics bounds depend on gradient estimates

$$\begin{cases} \partial_t \lambda - \Delta\lambda - \nabla\cdot(\lambda b^1 + m^2(b^1 - b^2)) = 0 \text{ in } (0,T)\times\Omega, \quad \lambda(0) = m_2 - m_1 \text{ in } \Omega\,. \\ -\partial_t\phi - \Delta\phi + b^1\cdot\nabla\phi = 0 \text{ in } [0,T)\times\Omega, \quad \phi(T,x) = \psi(x) \text{ in } \Omega \end{cases}$$

Backward-forward structure once again!

Integrate first equation against the second, then integrate by parts. Then,

$$\mathbf{d}(m^1(T), m^2(T)) \le \sup_{\psi\in\mathcal{X}} \left| \int_\Omega \lambda(0,x)\phi(0,x)dx \right| + \sup_{\psi\in\mathcal{X}} \left| \int_0^T \int_\Omega m^2 \boxed{\nabla\phi} \cdot (b^2 - b^1)dxdt \right| .$$

Bounds needs gradient estimates! For Wasserstein-1, $\sup_{\psi\in\mathcal{X}} \left| \int_\Omega \lambda(0,x)\phi(0,x)dx \right| \le \mathbf{d}_1(m_1, m_2)\boxed{\|\nabla\phi(0,x)\|_\infty}$

Regularity theory of HJB PDEs enables UQ in SGMs

Main ideas of Wasserstein Uncertainty Propagation proof

Step 3: Bernstein estimates from HJB theory provide gradient estimates

$$\mathbf{d}(m^1(T), m^2(T)) \leq \sup_{\psi \in \mathscr{X}} \left| \int_\Omega \lambda(0,x)\phi(0,x)dx \right| + \sup_{\psi \in \mathscr{X}} \left| \int_0^T \int_\Omega m^2 \boxed{\nabla \phi} \cdot (b^2 - b^1) dx dt \right| .$$

$$-\partial_t \phi - \Delta \phi + b^1 \cdot \nabla \phi = 0 \text{ in } [0,T) \times \Omega, \quad \phi(T,x) = \psi(x) \text{ in } \Omega$$

- Derive a PDE for $z = \frac{1}{2} \|\nabla \log \phi\|^2$
- Apply the maximum principle to obtain a bound on $z(t,x) \leq C\|\log \psi\|_\infty + c\|\nabla \log \psi\|_\infty$
- Derive a bound for $\nabla \phi(t,x)$, bound the IPM.

Conclusion and ongoing work

UQ and PDE regularity theory contributes to analysis, robustness of generative AI algorithms

- Making the bounds explicitly computable: provide *a posteriori* estimates on the quality of a generative model

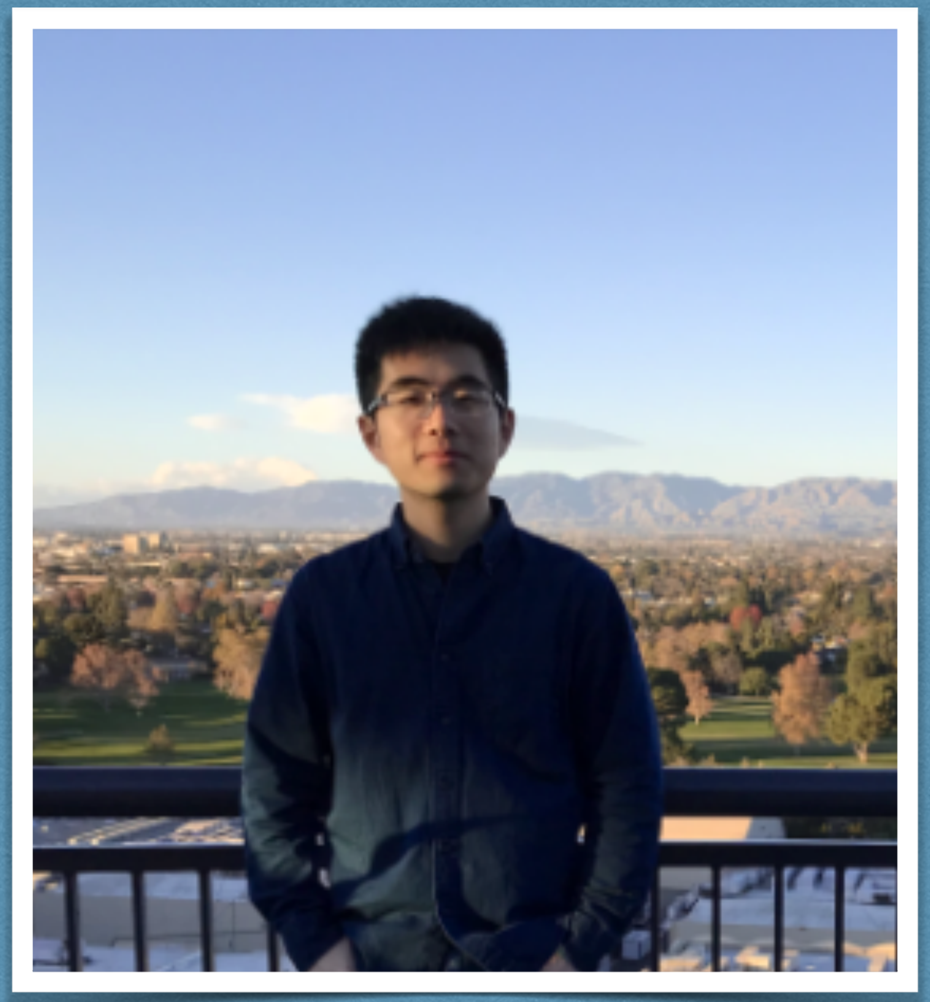
- Most useful for guarantees in likelihood-free inference settings & use IPM:

$$\left| \mathbb{E}_{\pi} h - \mathbb{E}_{m_g(T)} h \right| \leq \mathbf{d}(m_g(T), \pi) \leq \mathscr{F}(e_1, e_2, e_3, e_4, e_5) \, .$$

- Extensions to other generative flows with similar UQ issues (learning a drift) e.g., normalizing flows

Structure-informed generative modeling

Jeremiah Birrell,
Texas State, see poster

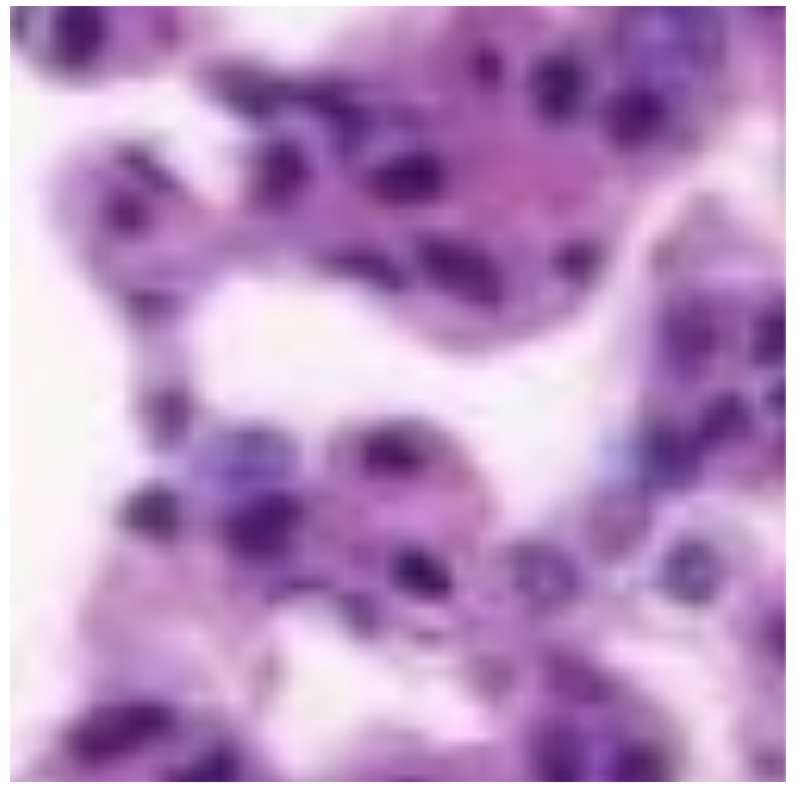


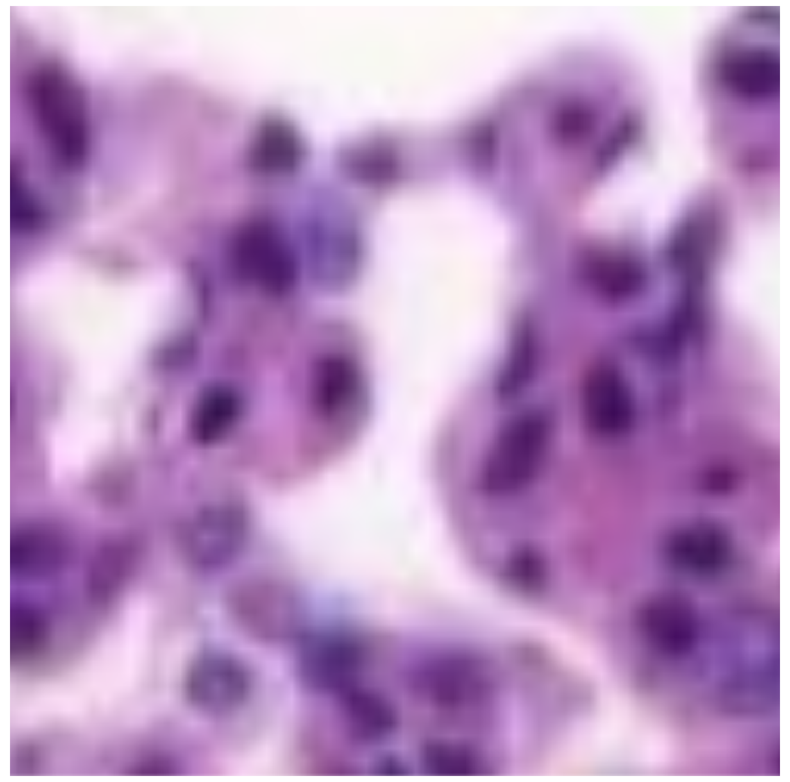
Ziyu Chen, UMass
Amherst

Structured-informed learning: target data & distribution π

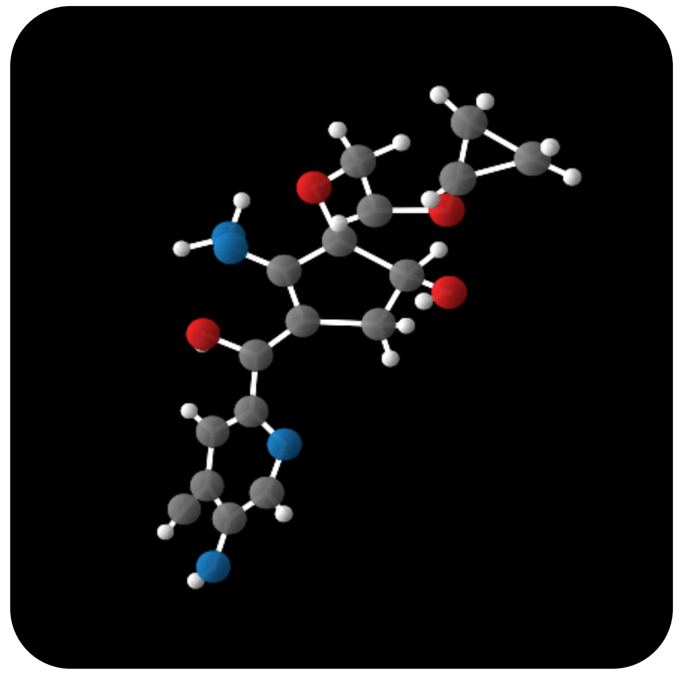
equivariance⟹equiprobable

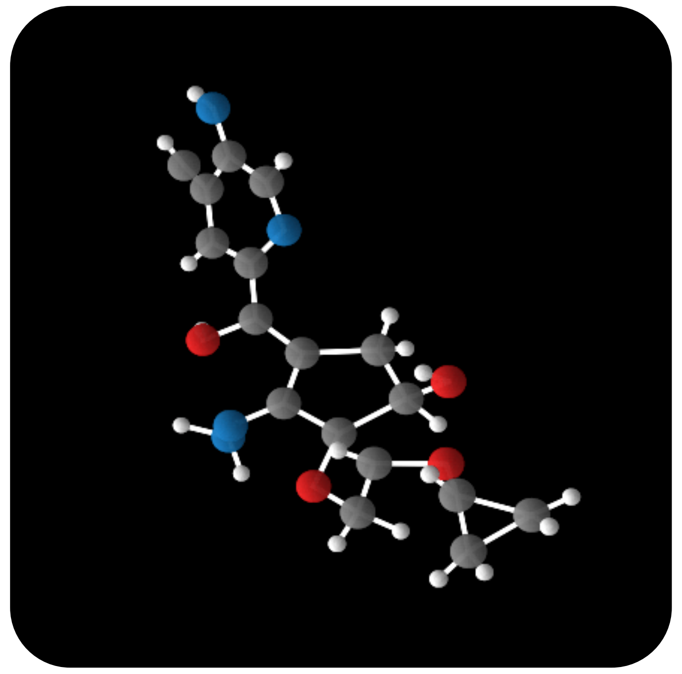
π





π

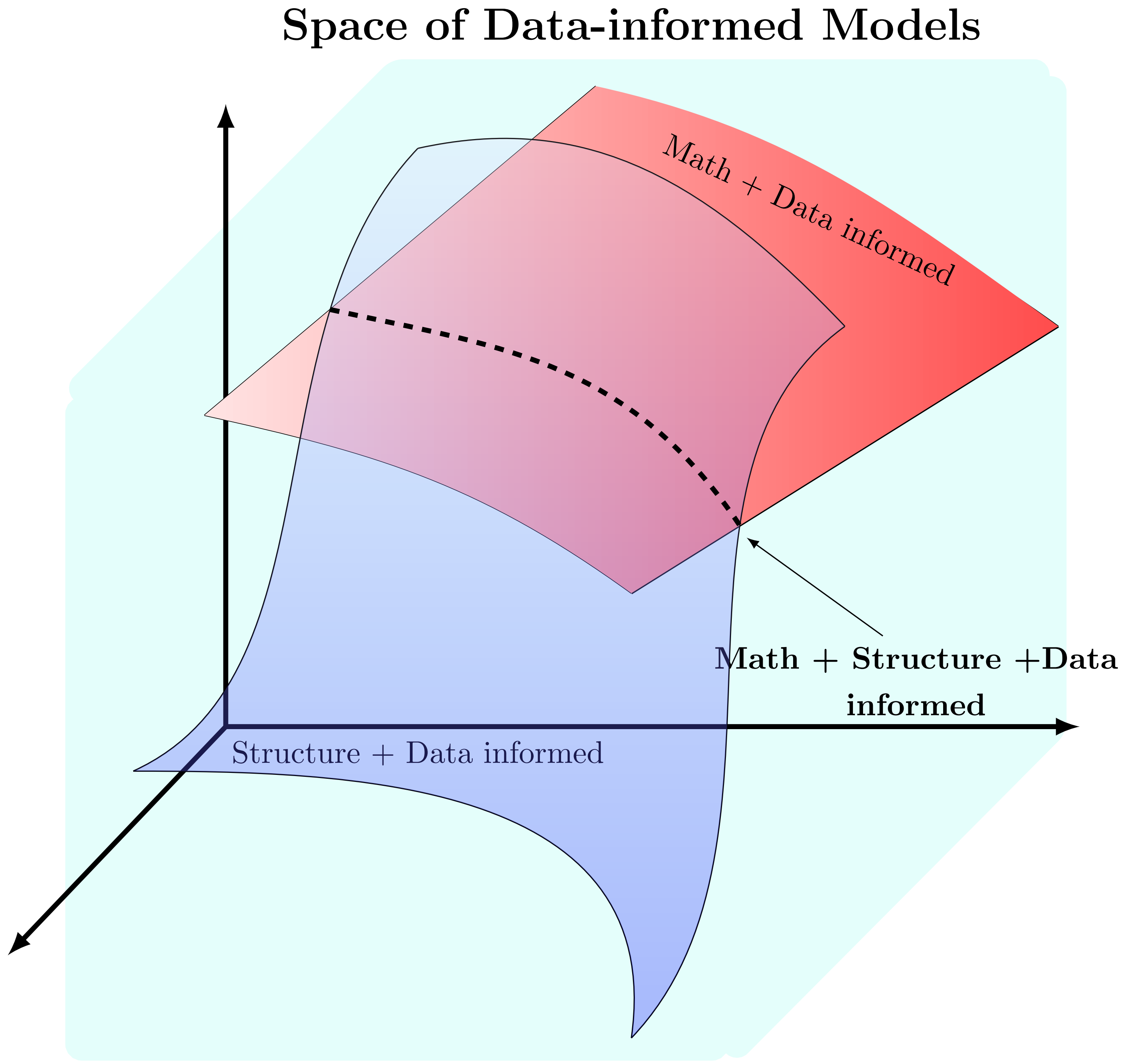




Learning from data, math- & physics-informed structures

We learn in (an even more) restricted & relevant space

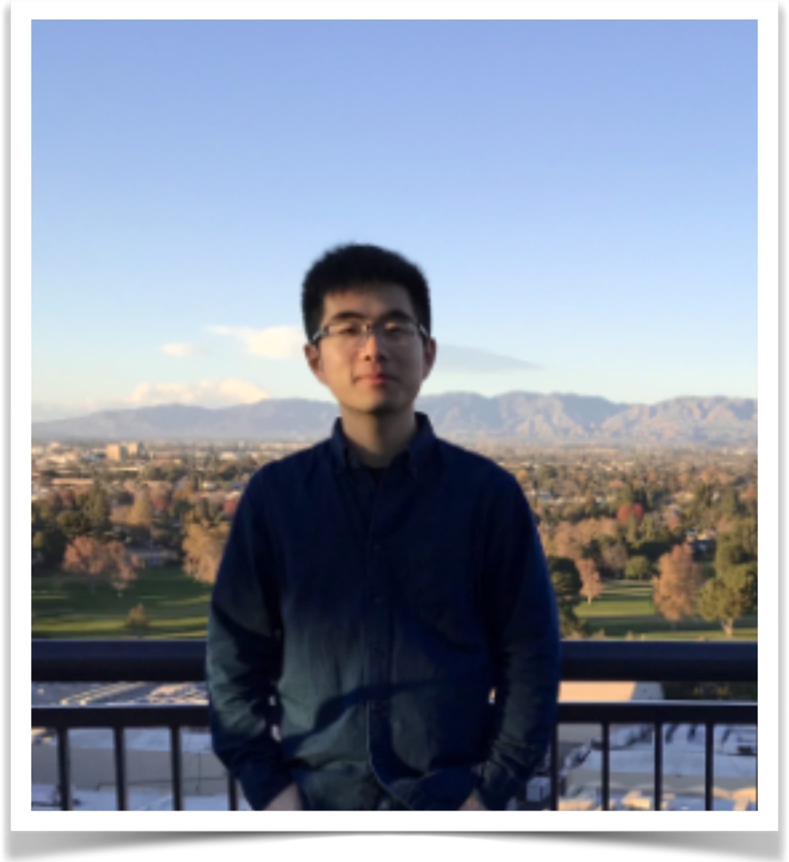
- How to build **embedded structure** into generative models for data-efficient distribution learning?
- Can we use structure/ physics to **learn faster**?
- **Quantify** the gains in performance

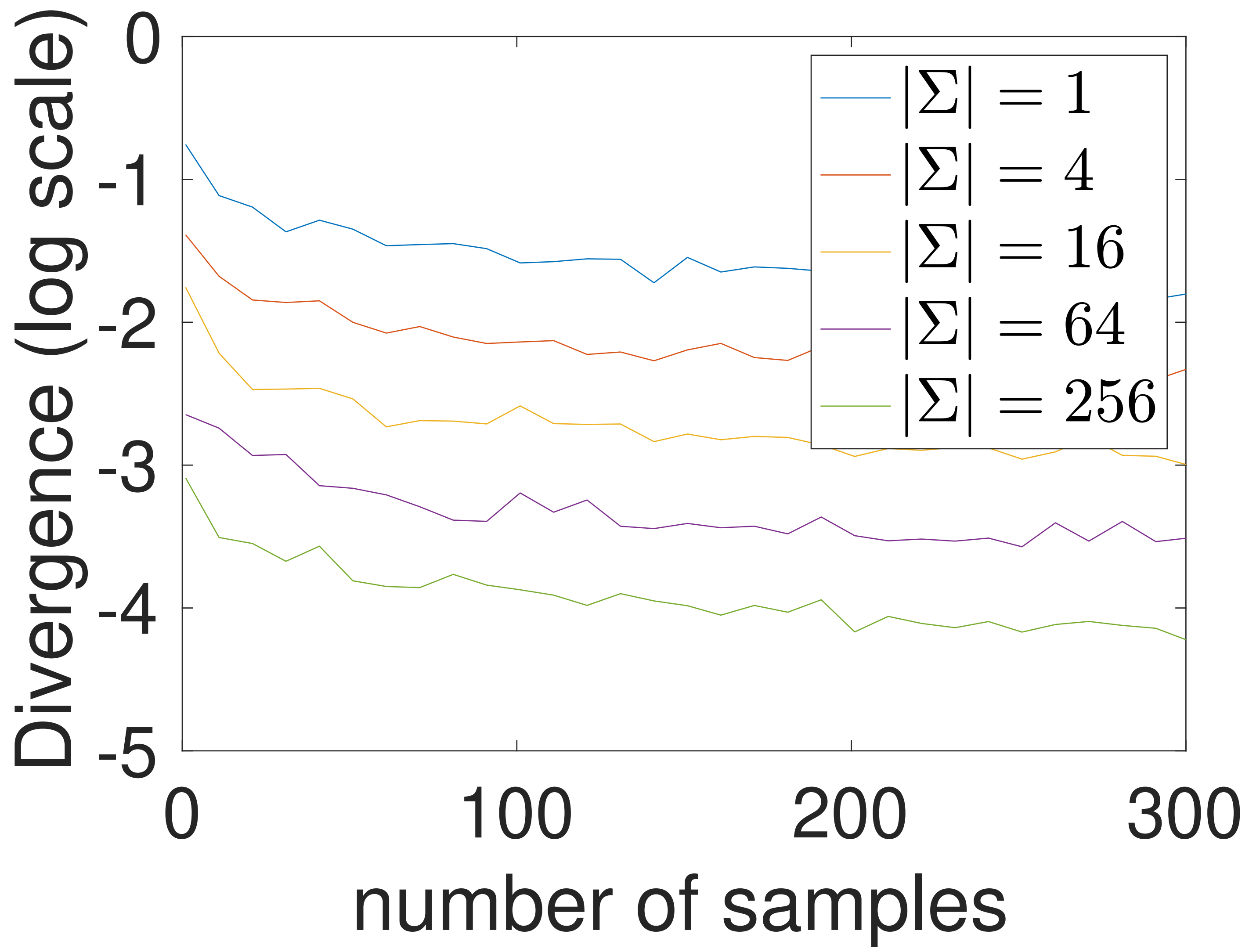


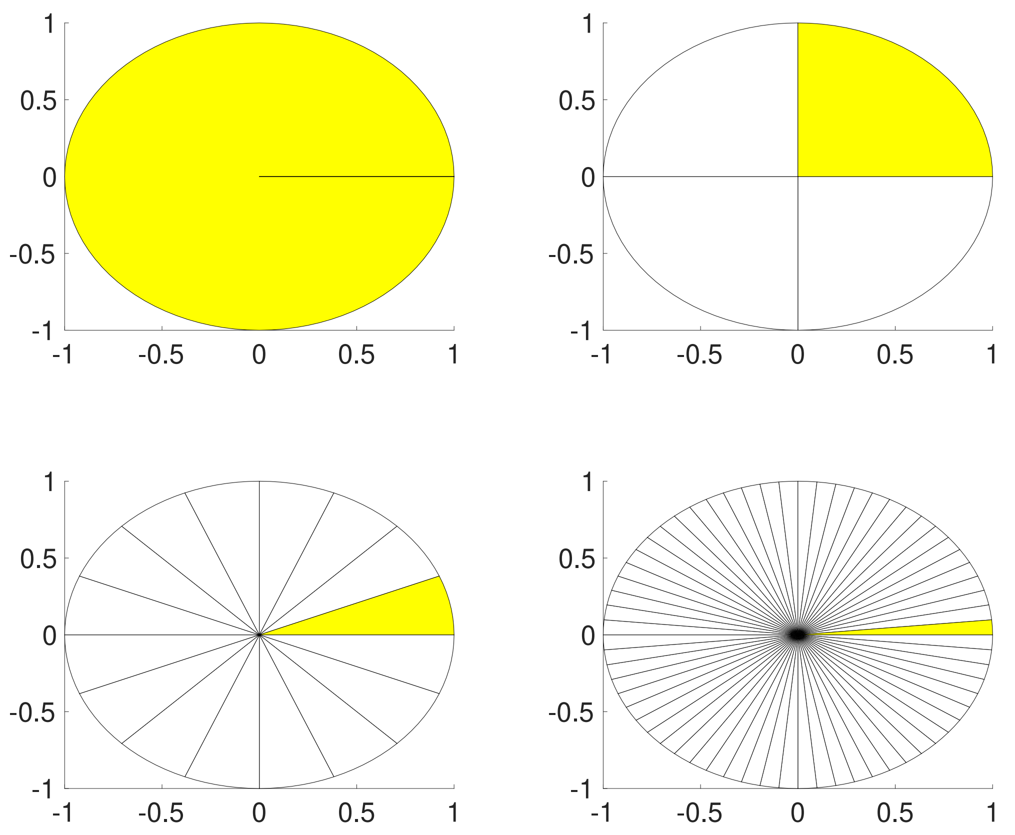
Sample Complexity of Probability Divergences under Symmetry: Quantify the gains in data needed for the same performance

P. Dupuis and Y. Mao, *Formulation and properties of a divergence used to compare probability measures without absolute continuity*, **ESAIM: COCV, ** 2022
Birrell, **P. Dupuis**, M. A. Katsoulakis, Y. Pantazis, L. Rey-Bellet, *(f, Γ) -Divergences: Interpolating between f-Divergences and IPMs*, **Journal of Machine Learning Research**, 2022

$$|D^{\Gamma}_{f_\alpha}(P\|\pi) - D^{\Gamma_\Sigma}_{f_\alpha}(P_m\|\pi_n)| \le C_1\left(\frac{1}{|\Sigma|m}\right)^{\frac{1}{d+s}} + C_2\left(\frac{1}{|\Sigma|n}\right)^{\frac{1}{d+s}}$$







reduction in # of needed data due to equivariance

see poster by Ziyu Chen

Related papers

- J. Birrell, M.A. Katsoulakis, L. Rey-Bellet, W. Zhu. *Structure-preserving GANs.* **ICML 2022**
- Z. Chen, M. A. Katsoulakis, L. Rey-Bellet, W. Zhu, *Sample Complexity of Probability Divergences under Group Symmetry,* **ICML 2023**
- Z. Chen, M. A. Katsoulakis, L. Rey-Bellet, W. Zhu, *Statistical Guarantees of Group-Invariant GANs, submitted,* (2024).
- Z. Chen, M.A. Katsoulakis, B. Zhang, ***Sample complexity and equivariant score matching of SGM with group symmetry***, *in preparation.*

A. Robust Generative Modeling: Mean-field Games, Hamilton-Jacobi, Proximals & UQ

1. B. J. Zhang, M. A. Katsoulakis, *A Mean-Field Games **laboratory** for generative modeling*, ***Arxiv,*** (2023).
2. H. Gu, M. A. Katsoulakis, L. Rey-Bellet, B. J. Zhang, *Combining Wasserstein-1 and Wasserstein-2 proximals: **robust manifold learning** via well-posed generative flows*, ***Arxiv,*** (2024).
3. N. Mimikos-Stamatopoulos, B. J. Zhang, M. A. Katsoulakis, ***Score-**based generative models are provably robust: a **UQ perspective***, ***Arxiv,*** (2024).
4. B. J. Zhang, S. Liu, W. Li, M. A. Katsoulakis, S. Osher, Wasserstein **proximal** operators describe score-based generative models and resolve **memorization**, ***Arxiv,*** (2024).

References

B. Structure-informed Learning

1. Z. Chen, M. A. Katsoulakis, L. Rey-Bellet, W. Zhu, *Statistical **Guarantees** of Group-Invariant GANs,* ***Arxiv,*** (2024).
2. Z. Chen, M. A. Katsoulakis, L. Rey-Bellet, W. Zhu, ***Sample Complexity** of Probability Divergences under Group Symmetry,* **ICML 2023**
3. J. Birrell, M.A. Katsoulakis, L. Rey-Bellet, W. Zhu, *Structure-preserving GANs.* **ICML 2022**
4. J. Birrell, M.A. Katsoulakis, L. Rey-Bellet, B. J. Zhang, W. Zhu, ***Nonlinear denoising** score matching for enhanced learning of structured distributions*, ***Arxiv,*** (2024).

C. Optimal Transport Proximals for Machine Learning

1. J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, L. Rey-Bellet, *Function-space **regularized** Rényi divergences*, International Conference on Learning Representations, **ICLR 2023**
2. H. Gu, P. Birmpa, Y. Pantazis, M. A. Katsoulakis, and L. Rey-Bellet, ***Lipschitz-regularized** gradient flows and generative particle algorithms for high-dimensional **scarce** data,* **SIAM Data Science,** to appear, (2024).
3. J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, L. Rey-Bellet, *(f, Γ) -Divergences: **Interpolating** between f-Divergences and Integral Probability Metrics*, **Journal of Machine Learning Research & NeurIPS**, (2022)
4. Z. Chen, H. Gu, M. A. Katsoulakis, L. Rey-Bellet, W. Zhu, *Learning **heavy-tailed** distributions with Wasserstein-**proximal**-regularized α-divergences*, ***Arxiv,*** (2024)

Extra Slides

Wasserstein gradient flows

Gradient flows on the space of probability distributions

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho \nabla \frac{\delta \mathscr{F}}{\delta \rho} \right)$$

Describes a path on the space of probability distributions

$$\mathscr{F}(\rho) = \mathscr{D}_{KL}(\rho \| \pi) = \mathbb{E}_{\rho} \left[\log \frac{\rho}{\pi} \right]$$

Overdamped Langevin

$$\mathscr{F}(\rho) = \mathscr{D}_{\alpha}(\rho \| \pi) = \mathbb{E}_{\pi} \left[f_{\alpha} \left(\frac{\rho}{\pi} \right) \right]$$

Porous medium equation

$$f_{\alpha}(x) = \frac{x^{\alpha}}{\alpha(\alpha - 1)}$$

Wasserstein gradient flows as solutions to MFGs

$$\inf_{v,\rho}\left\{\mathcal{F}(\rho(\cdot\,,T)) + \int_0^T \frac{e^{-t/\epsilon}}{\epsilon}\mathcal{F}(\rho(\cdot\,,t))\,dt + \int_0^T\int_{\mathbb{R}^d}\frac{e^{-t/\epsilon}}{2}\|v\|^2\rho(x,t)\,dx\,dt\right\}$$

s.t. $\partial_t\rho + \nabla\cdot(v\rho) = 0, \qquad \rho(x,0) = \rho_0(x)$

$$-\epsilon\frac{\partial U}{\partial t} + U + \frac{\epsilon}{2}|\nabla U|^2 = \frac{\delta\mathcal{F}}{\delta\rho}$$

$$\frac{\partial\rho}{\partial t} - \nabla\cdot\left(\rho\nabla U\right) = 0.$$

$$U(x,T) = \frac{\delta\mathcal{F}}{\delta\rho}(\rho(\cdot\,,T)), \rho(x,0) = \rho_0(x)$$

$\epsilon \to 0$

Relaxation limit enforces a “local equilibrium” :

$$U \approx \frac{\delta\mathcal{F}}{\delta\rho}$$

$$U = \frac{\delta\mathcal{F}}{\delta\rho}$$

$$\frac{\partial\rho}{\partial t} - \nabla\cdot\left(\rho\nabla U\right) = 0$$

$$\rho(x,0) = \rho_0(x)$$

MFGs reveal an interpolation between Wasserstein gradient flows & normalizing flows

$$\inf_{v,\rho} \left\{ \mathcal{F}(\rho(\cdot, T)) + \int_0^T \frac{e^{-t/\epsilon}}{\epsilon} \mathcal{F}(\rho(\cdot, t))\, dt + \int_0^T \int_{\mathbb{R}^d} \frac{e^{-t/\epsilon}}{2} \|v\|^2 \rho(x, t)\, dx\, dt \right\}$$

s.t. $\partial_t \rho + \nabla \cdot (v\rho) = 0, \qquad \rho(x,0) = \rho_0(x)$

$\epsilon \to 0$

$$U = \frac{\delta \mathcal{F}}{\delta \rho}$$

$$\frac{\partial \rho}{\partial t} - \nabla \cdot \left(\rho \nabla U\right) = 0$$

$$\rho(x,0) = \rho_0(x)$$

Wasserstein gradient flows

$0 < \epsilon < \infty$

?

$\epsilon \to \infty$

$$\inf_{v,\rho} \left\{ \mathcal{F}(\rho(\cdot, T)) + \int_0^T \int_{\mathbb{R}^d} \frac{1}{2} \|v\|^2 \rho(x, t)\, dx\, dt \right\}$$

s.t. $\partial_t \rho + \nabla \cdot (v\rho) = 0, \qquad \rho(x,0) = \rho_0(x)$

Optimal transport normalizing flow