Title: Measure Transport Perspectives on Sampling, Generative Modeling, and Beyond

Speakers: Michael Albergo

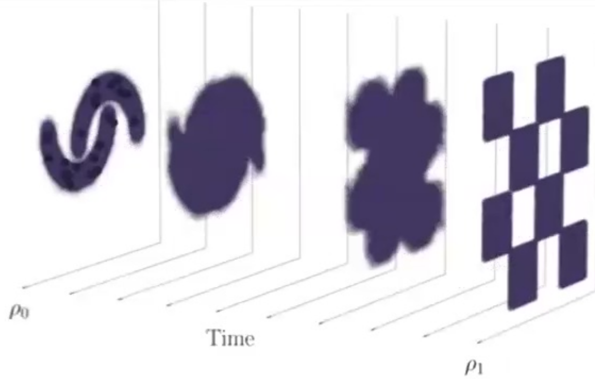
Series: Machine Learning Initiative

Date: April 12, 2024 - 2:30 PM

URL: https://pirsa.org/24040087

Abstract: Both the social and natural world are replete with complex structure that often has a probabilistic interpretation. In the former, we may seek to model, for example, the distribution of natural images or language, for which there are copious amounts of real world data. In the latter, we are given the probabilistic rule describing a physical process, but no procedure for generating samples under it necessary to perform simulation. In this talk, I will discuss a generative modeling paradigm based on maps between probability distributions that is applicable to both of these circumstances. I will describe a means for learning these maps in the context of problems in statistical physics, how to impose symmetries on them to facilitate learning, and how to use the resultant generative models in a statistically unbiased fashion. I will then describe a paradigm that unifies flow-based and diffusion based generative models by recasting generative modeling as a problem of regression. I will demonstrate the efficacy of doing this in computer vision problems and end with some future challenges and applications.

Zoom link



Measure transport perspectives on sampling, generative modeling, and beyond

Perimeter Institute Machine Learning Initiative Seminar

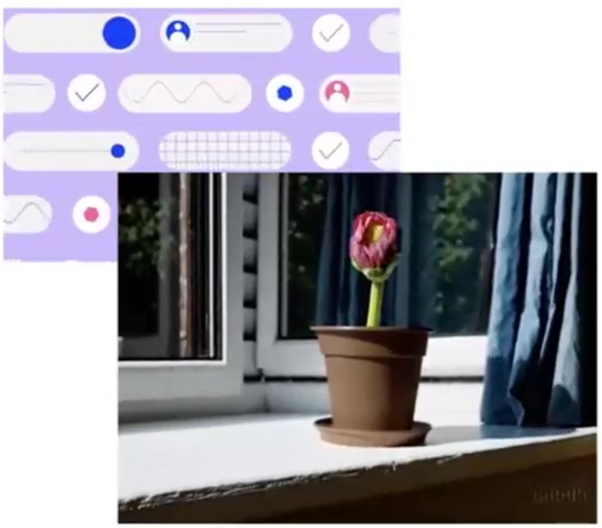
Michael Albergo April 12, 2024


Thanks to collaborators and mentors!!!!!!!!!!
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
Universität Bern
AEC Albert Einstein Center for Fundamental Physics
G. Kanwar
University of Maryland
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
April 11, 2024
2
Michael Samuel Albe...

Complexity all around

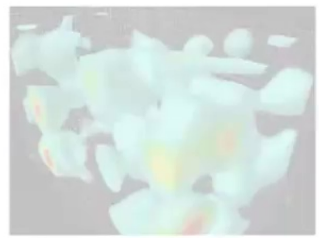
The social and natural worlds are replete with complex structure that often has a probabilistic interpretation

Social: abundance of data

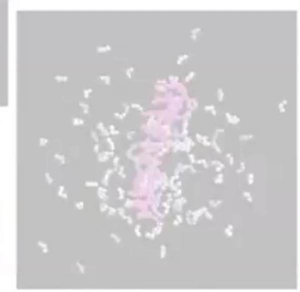


Sora (2024): "A flower growing out on the windowsill"

Natural: limited data, but theory



Quantum Theory



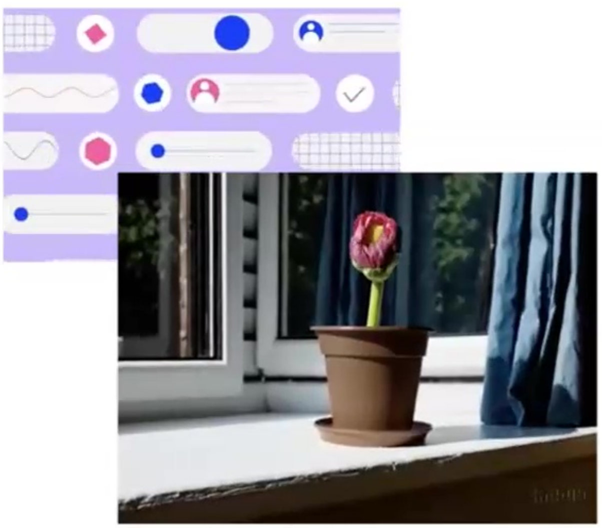
Molecular conformation

Forecasting

Complexity all around

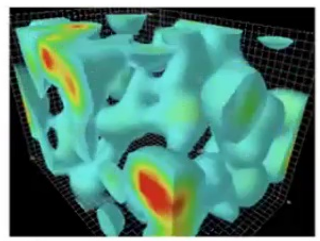
The social and natural worlds are replete with complex structure that often has a probabilistic interpretation

Social: abundance of data

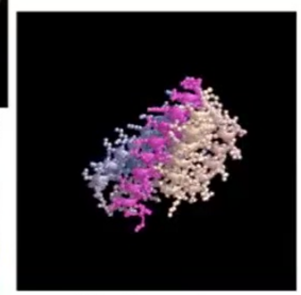


Sora (2024): "A flower growing out on the windowsill"

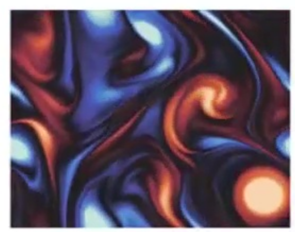
Natural: limited data, but theory



Quantum Theory

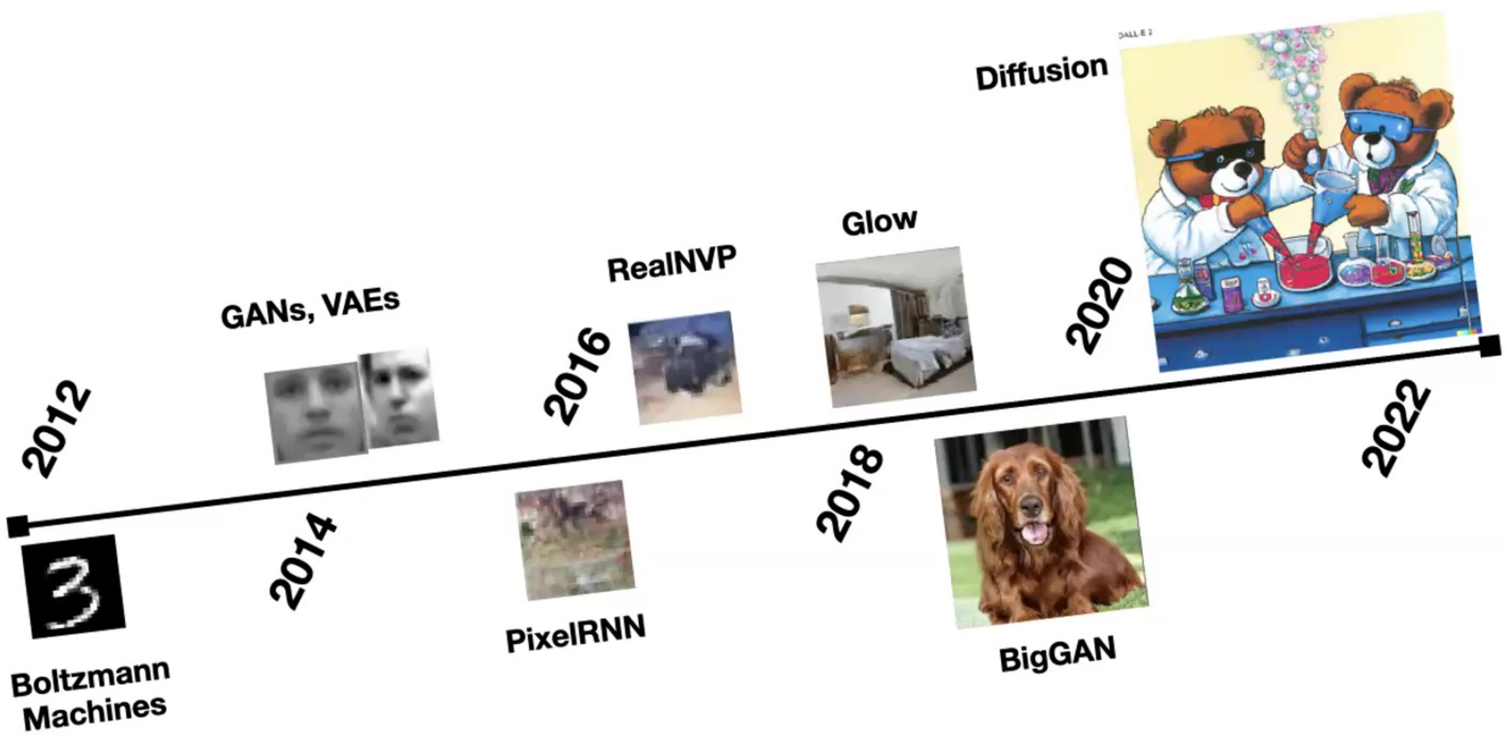


Molecular conformation



Forecasting

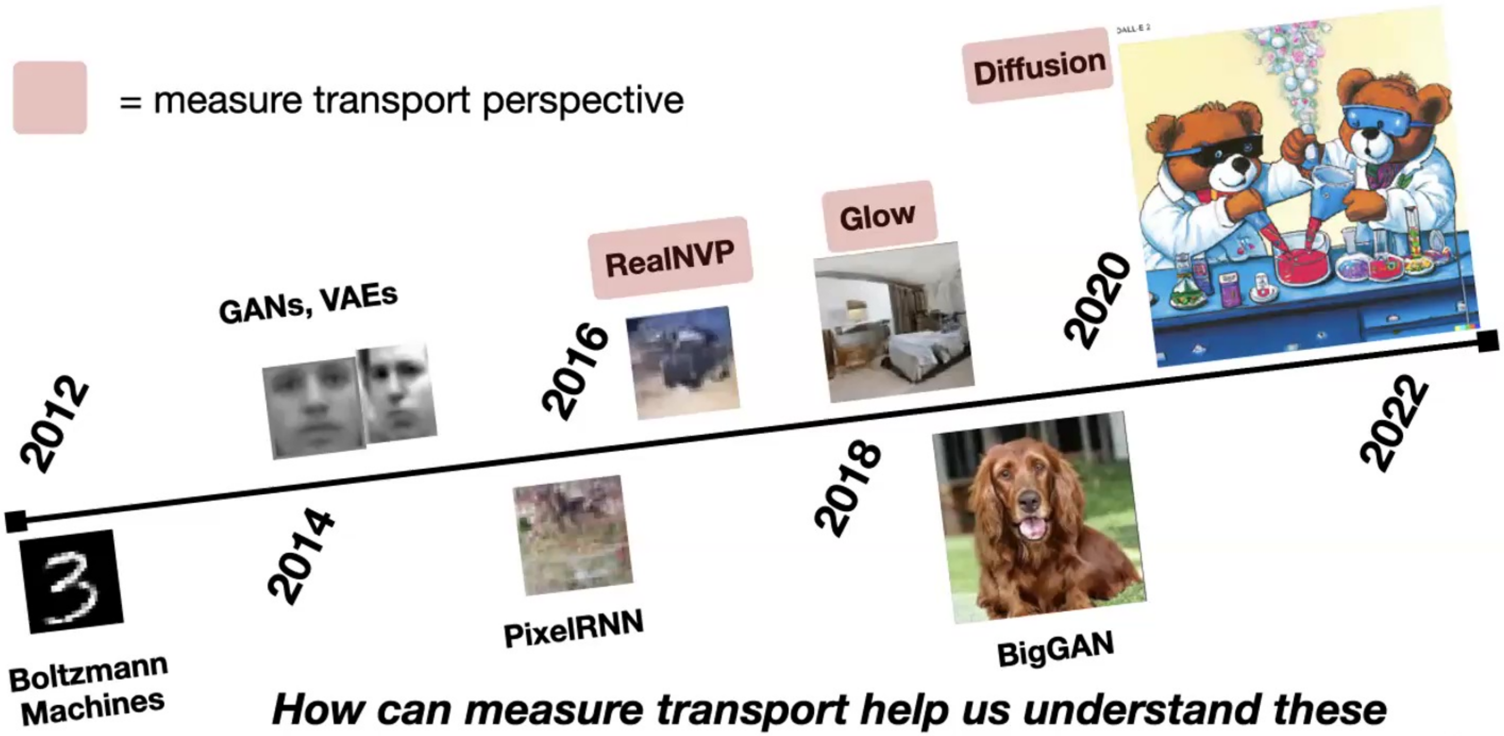
Ascendancy of generative modeling



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



How can measure transport help us understand these successes and build more performant, understandable tools?

Agenda

Part 1: New algorithms for dynamical measure transport

- **Problem Challenge**
- **Stochastic Interpolants**
- **Applications**

Artificial

Part 2: Statistically robust ML for statistical physics

- Learning without data
- Non-Euclidean data
- Symmetries

Natural

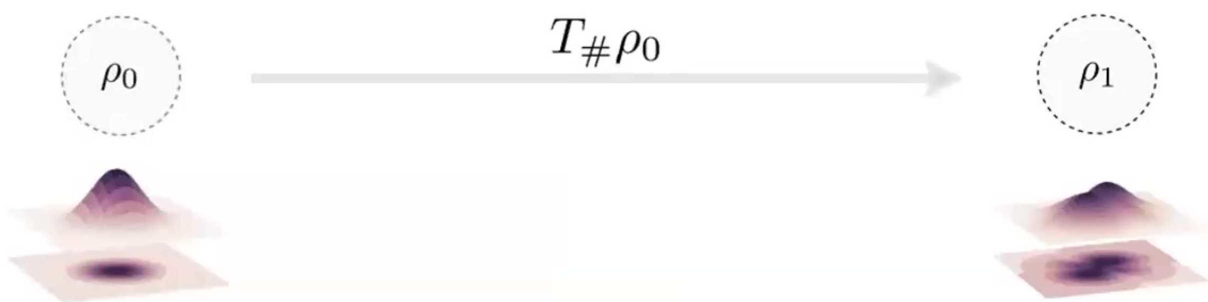
Problem Setup

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

The transport framework

- Take a simple *base density* ρ_0 (e.g. Gaussian) and;
- Build a (reversible) map $T : \Omega \to \Omega$ such that the *pushforward of* ρ_0 *by* T *is* ρ_1: $T\sharp\rho_0 = \rho_1$

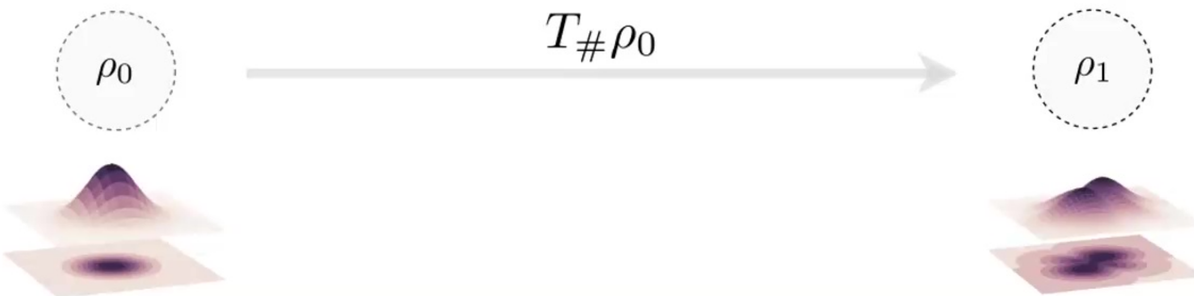


Likelihood under $\rho(1)$ given by: $\rho 1(x_1) = \rho_0(T^{-1}(x)) \det[\nabla T^{-1}(x)]$

Problem Setup

The transport framework

- Build a (reversible) map $T : \Omega \to \Omega$ such that the *pushforward of* $\rho(0)$ *by* T *is* $\rho(1)$: $T\sharp\rho(0) = \rho(1)$



Likelihood: $\rho_1(x) = \rho_0(T^{-1}(x)) \det[\nabla T^{-1}(x)]$

For parametric $\hat{T}(x)$ to be useful

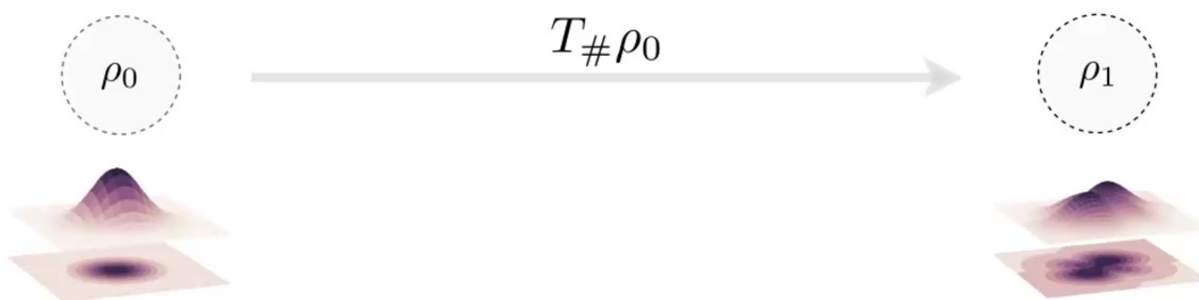
- $\det[\nabla \hat{T}^{-1}(x)]$ to be **tractable**
- $\hat{T}(x)$ **maximally unconstrained**

Tradeoff!

Problem Setup

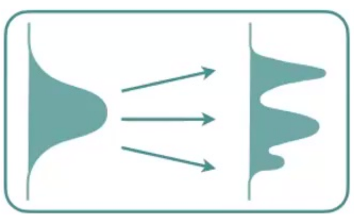
The transport framework

- Build a (reversible) map $T : \Omega \to \Omega$ such that the *pushforward of $\rho(0)$ by T is $\rho(1)$:* $T\sharp\rho(0) = \rho(1)$



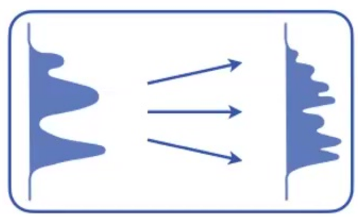
Likelihood: $\rho_1(x) = \rho_0(T^{-1}(x)) \det[\nabla T^{-1}(x)]$

Generative modeling



Ex. Image generation
Ex. Statistical physics

Domain Adaptation



Ex. Translation

Forecasting

Ex. Climate/weather
Ex. Dynamical systems

Problem Setup

The transport framework

- Build a (reversible) map $T : \Omega \to \Omega$ such that the *pushforward of* $\rho(0)$ *by* T *is* $\rho(1)$: $T\sharp\rho(0) = \rho(1)$

$$\rho_0 \xrightarrow{T_{\#}\rho_0} \rho_1$$

How do we harness measure transport for these various tasks in probabilistic modeling? How do we learn these maps?

Ex. Image generation
Ex. Statistical physics

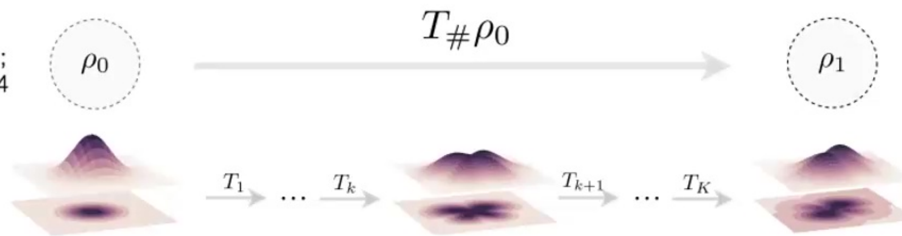
Ex. Translation

Ex. Climate/weather
Ex. Dynamical systems

Brief history on transport realizations

Series of discrete transforms

$\det[\nabla T^{-1}(x)]$ tractable, but too constrained?



T_k learned sequentially

Chen & Gopinath, NeurIPS 13 (2000);
Tabak & V.-E., Commun. Math. Sci. 8: 217-233 (2010);
Tabak & Turner, Comm. Pure App. Math LXVI, 145-164 (2013).

T_k structured invertible NNs

NICE: Dinh *et al.* arXiv:1410.8516 (2014);
Real NVP: Dinh *et al.* arXiv:1605.08803 (2016)
Rezende *et al.*, arXiv:1505.05770 (2015);
Papamakarios *et al.* arXiv:1912.02762 (2019); …

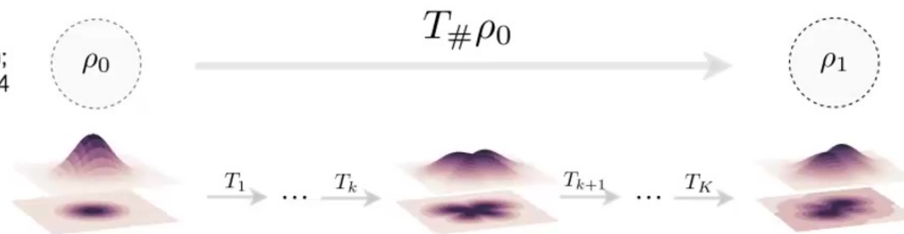
Brief history on transport realizations

Series of discrete transforms

$\det[\nabla T^{-1}(x)]$ tractable, but too constrained?

T_k learned sequentially

Chen & Gopinath, NeurIPS 13 (2000);
Tabak & V.-E., Commun. Math. Sci. 8: 217-233 (2010);
Tabak & Turner, Comm. Pure App. Math LXVI, 145-164 (2013).



T_k structured invertible NNs

NICE: Dinh *et al.* arXiv:1410.8516 (2014);
Real NVP: Dinh *et al.* arXiv:1605.08803 (2016)
Rezende *et al.*, arXiv:1505.05770 (2015);
Papamakarios *et al.* arXiv:1912.02762 (2019); …

$k \to \infty$

T solution of *continuous time flow*

FFJORD: Grathwohl *et al.* arXiv:1810.01367 (2018)

- $\det[\nabla T^{-1}(x)] \to \mathrm{Tr}[\frac{\partial b_t}{\partial x(t)}]$
- estimable via Skilling-Hutchinsion $O(D)$
- integrable with Neural ODEs

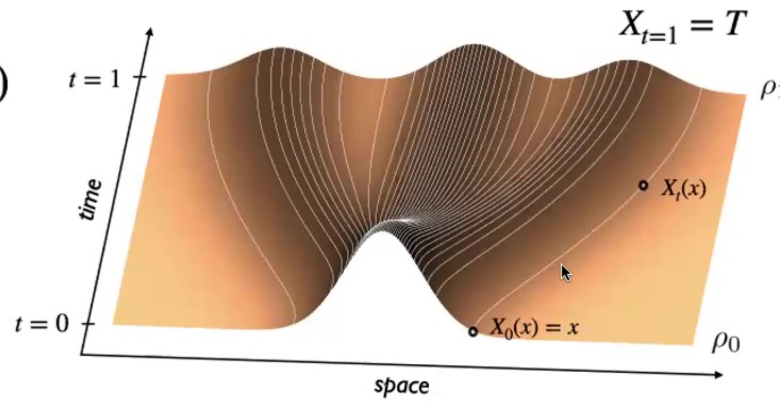
April 11, 2024 12

The continuous time picture

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

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

$$\dot{X}_t(x) = b(t, X_t(x))$$

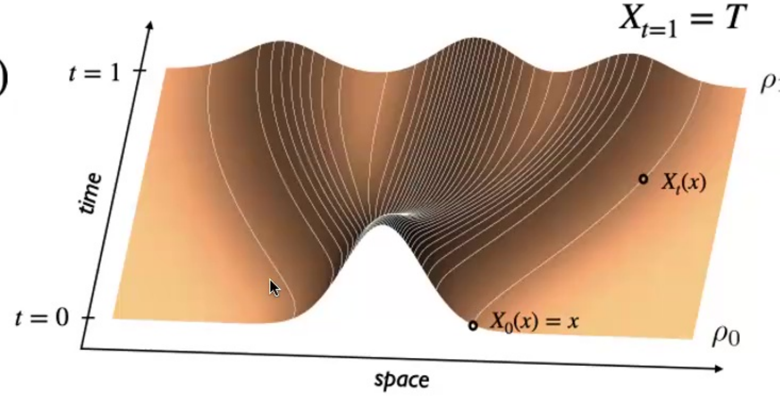


The continuous time picture

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

$$\partial_t \rho(t, x) + \nabla \cdot \big(b(t, x)\rho(t, x)\big) = 0, \quad \rho(t = 0, \cdot) = \rho_0$$

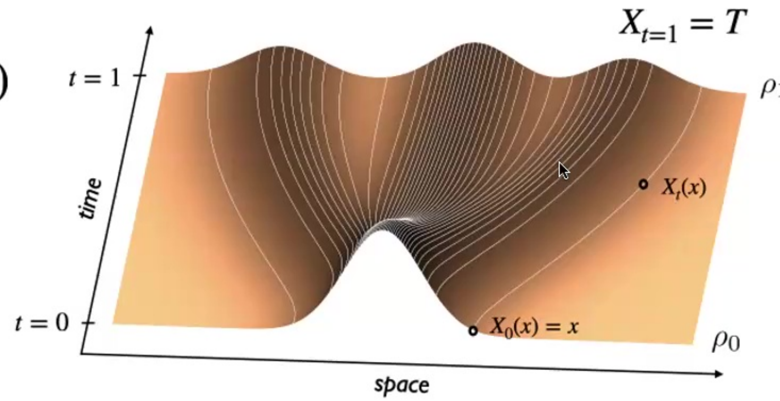
If $\rho(t)$ solves TE, **then** $\rho(t = 1, \cdot) = \rho_1$

The continuous time picture

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 $$\partial_t \rho(t,x) + \nabla \cdot \big(b(t,x)\rho(t,x)\big) = 0, \quad \rho(t=0,\cdot) = \rho_0$$

If $\rho(t)$ solves TE, **then** $\rho(t=1,\cdot) = \rho_1$

Benamou-Brenier theory says that $b(t,x)$ exists (assuming Lipschitz)

How to find a sufficient $b(t,x)$ to map ρ_0 to ρ_1?

Solving for $b(t, x)$ solves the transport

Is there a simple paradigm for learning $b(t, x)$?

Dream scenario: figure out a way to perform regression on the velocity field

$$\min_{\hat{b}} \int_{t=0}^{t=1} |b(t,x) - \hat{b}(t,x)|^2 \rho(t,x) dx dt$$

Problems:

- Don't have a fixed $b(t, x)$ to regress on
- Don't have a $\rho(t, x)$ to sample from!

How can we work exactly on $t \in [0,1]$ with arbitrary ρ_0 and ρ_1, build a connection between them, and get the velocity $b(t, x)$ directly?

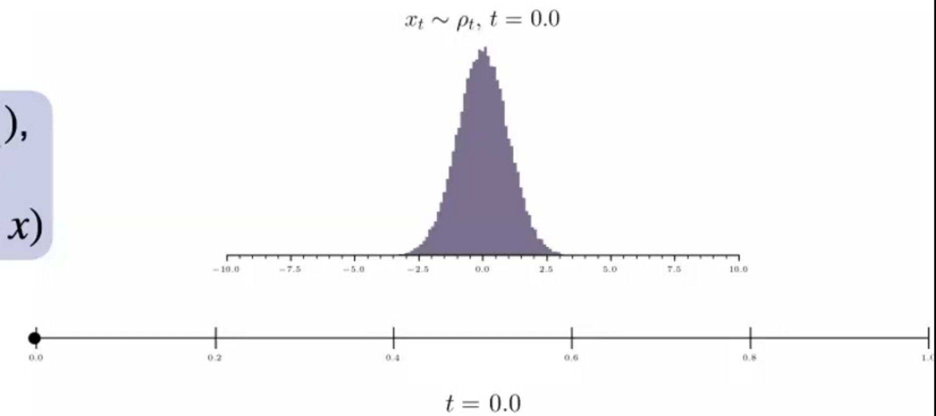
Stochastic Interpolants

MSA & Vanden-Eijnden arXiv:2209.15571 (2022);

Interpolant Function $x(t, x_0, x_1)$

- A function of x_0, x_1, and time t with b.c.'s: $x_{t=0} = x_0$ and $x_{t=1} = x_1$
- Example: $x(t, x_0, x_1) = (1 - t)x_0 + tx_1$

If x_0, x_1 drawn from some $\rho(x_0, x_1)$, then $x(t, x_0, x_1)$ is a **stochastic process which samples** $x_t \sim \rho(t, x)$



Interpolant Density

$$\rho(t, x) = \mathbb{E}_{\rho(x_0, x_1)}\left[\delta\big(x - x(t, x_0, x_1)\big)\right]$$

What fixes $\rho(t, x)$?

1. Choice of **coupling**: how to sample x_0, x_1

   simple example: $\rho(x_0, x_1) = \rho_0(x_0)\rho_1(x_1)$
2. Choice of **interpolant** $x(t, x_0, x_1)$:

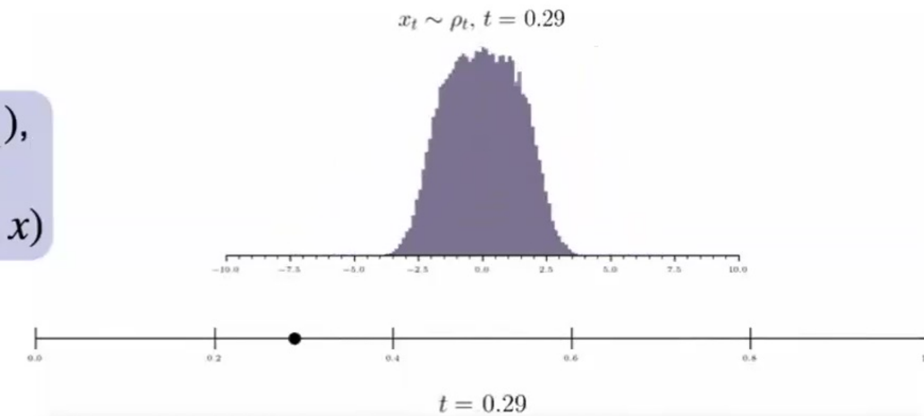
Stochastic Interpolants

MSA & Vanden-Eijnden arXiv:2209.15571 (2022);

Interpolant Function $x(t, x_0, x_1)$

- A function of x_0, x_1, and time t with b.c.'s: $x_{t=0} = x_0$ and $x_{t=1} = x_1$
- Example: $x(t, x_0, x_1) = (1-t)x_0 + tx_1$

If x_0, x_1 drawn from some $\rho(x_0, x_1)$, then $x(t, x_0, x_1)$ is a **stochastic process which samples** $x_t \sim \rho(t, x)$



Interpolant Density

$$\rho(t,x) = \mathbb{E}_{\rho(x_0,x_1)}\left[\delta\big(x - x(t,x_0,x_1)\big)\right]$$

1. Choice of **coupling**: how to sample x_0, x_1
   simple example: $\rho(x_0, x_1) = \rho_0(x_0)\rho_1(x_1)$
2. Choice of **interpolant** $x(t, x_0, x_1)$:

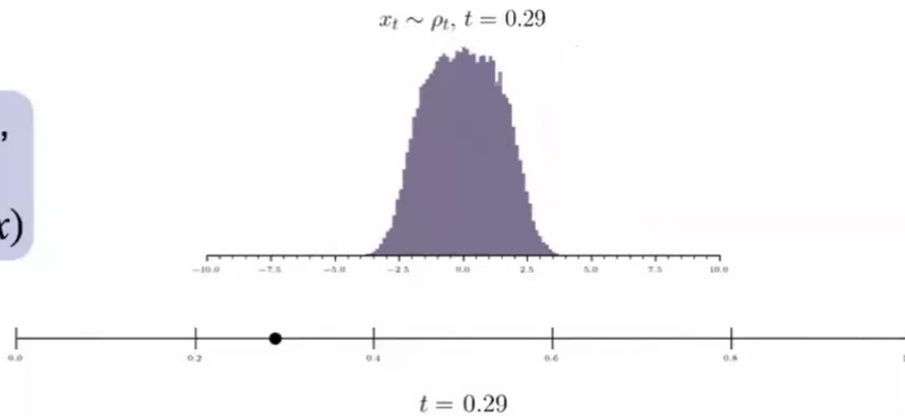
Stochastic Interpolants

Interpolant Function $x(t, x_0, x_1)$

MSA & Vanden-Eijnden arXiv:2209.15571 (2022);

- A function of x_0, x_1, and time t with b.c.'s: $x_{t=0} = x_0$ and $x_{t=1} = x_1$
- Example: $x(t, x_0, x_1) = (1 - t)x_0 + tx_1$

If x_0, x_1 drawn from some $\rho(x_0, x_1)$, then $x(t, x_0, x_1)$ is a **stochastic process which samples** $x_t \sim \rho(t, x)$



Interpolant Density

$$\rho(t,x) = \mathbb{E}_{\rho(x_0,x_1)}\left[\delta\big(x - x(t,x_0,x_1)\big)\right]$$

Can sample $\rho(t,x)$!

$$\min_{\hat{b}} \int_{t=0}^{t=1} |b(t,x) - \hat{b}(t,x)|^2 \rho(t,x) dx dt$$

Stochastic Interpolants: what is $b(t,x)$?

Interpolant Function $x(t, x_0, x_1)$

- Example: $x(t, x_0, x_1) = (1-t)x_0 + tx_1$
- when $x_0, x_1 \sim \rho(x_0, x_1)$, $x_t \sim \rho(t)$

$$\min_{\hat{b}} \int_{t=0}^{t=1} |\, b(t,x) - \hat{b}(t,x)\,|^2 \rho(t,x) dx dt$$

We have samples $x_t \sim \rho(t,x)$ via the interpolant, but what is $b(t,x)$?

Definition

The $\rho(t,\cdot)$ of x_t satisfies a transport equation

$$\partial_t \rho + \nabla \cdot \big(b(t,x)\rho\big) = 0, \quad \rho(t=0,\cdot) = \rho_0$$

and $b(t,x)$ is given as the conditional expectation

$$b(t,x) = \mathbb{E}[\partial_t x(t) \,|\, x_t = x]$$

prove with characteristic function, sketch in backup slides.

Stochastic Interpolants: Simple Objective

$$\min_{\hat{b}} \int_{t=0}^{t=1} |\hat{b}(t,x) - b(t,x)|^2 \rho(t,x) dx dt$$

plug in definition of $b(t,x)$

$$\min_{\hat{b}} \int_{t=0}^{t=1} \int_{\mathbb{R}^d} |\mathbb{E}[\partial_t x(t) \,|\, x_t = x] - \hat{b}(t,x)|^2 \rho(t,x) dx dt$$

$$\int_{\mathbb{R}^d} \mathbb{E}[\partial_t x(t) \,|\, x_t = x] \rho(t,x) = \mathbb{E}_{\rho(x_0,x_1)}[\partial_t x(t)]$$

Note: definition of conditional expectation

Prop.

$b(t,x)$ is the minimizer of

$$L[\hat{b}] = \int_0^1 \mathbb{E}_{\rho(x_0,x_1)} \left[|\hat{b}(t,x(t)) - \partial_t x(t)|^2 \right] dt$$

using shorthand $x(t) = x(t, x_0, x_1)$

April 11, 2024 20

Stochastic Interpolants: Generative Model

Flow matching

***MSA** & Vanden-Eijnden arXiv:2209.15571 (2022);*
Liu et al. arXiv:2209.03003 (2022);
Lipman et al. arXiv:2210.02747 (2022)

Prop.

$b(t,x)$ is the minimizer of

$$L[\hat{b}] = \int_0^1 \mathbb{E}_{\rho(x_0,x_1)}\left[\,|\,\hat{b}(t,x(t)) - \partial_t x(t)\,|^2\right] dt$$

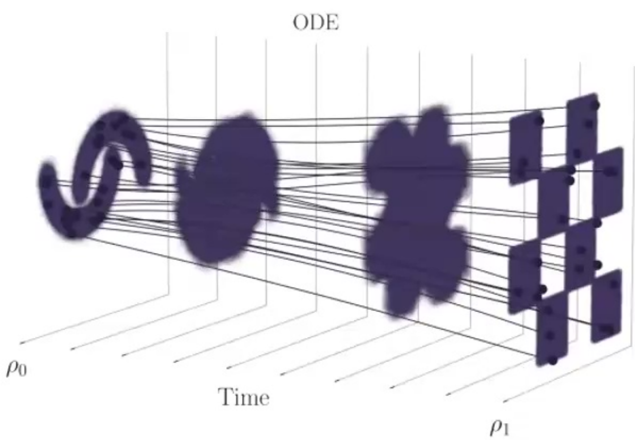
using shorthand $x(t) = x(t, x_0, x_1)$

- Loss is directly estimable over ρ_0, ρ_1
- Generative model connects *any* two densities
- Likelihood and sampling available via fast ODE integrators
- Loss bounds Wasserstein-2 between $\rho(1,x)$ and ρ_1 (Gronwall)

Generative model $\dot{X}_t(x) = b\big(t, X_t(x)\big)$

Deterministic vs stochastic transport

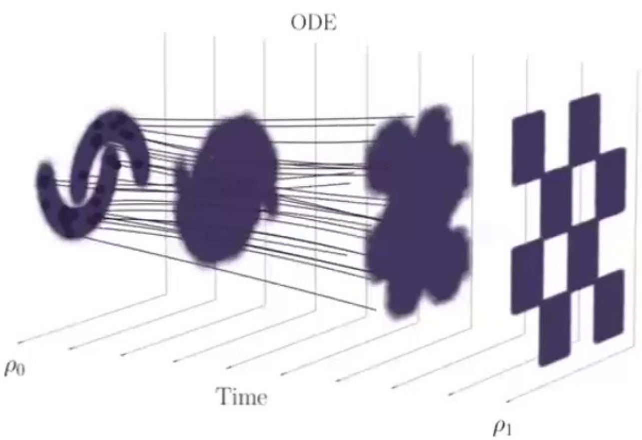
Example learned flow map



Deterministic

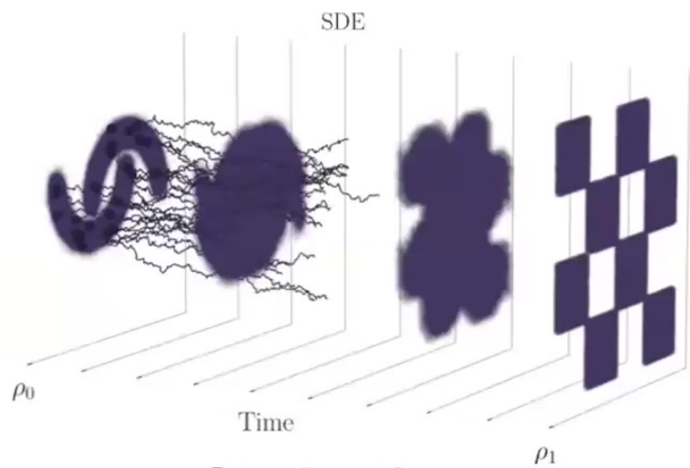
Deterministic vs stochastic transport

Example learned flow map



Deterministic

What about diffusion?



Stochastic

A simple set of criteria fulfills this

The interpolant score $s(t, x)$

***MSA**, Boffi, Vanden-Eijnden arXiv:2303.08797 (2023)*

Introduce Gaussianity into the interpolant

$$x(t) = I(t, x_0, x_1) + \gamma(t)z$$

where $z \sim \mathrm{N}(0,1)$
and $\gamma(0) = \gamma(1) = 0$
e.g. $\gamma(t) = \sqrt{t(1-t)}$

The interpolant score $s(t,x)$

***MSA**, Boffi, Vanden-Eijnden arXiv:2303.08797 (2023)*

Introduce Gaussianity into the interpolant

$$x(t) = I(t, x_0, x_1) + \gamma(t) z$$

where $z \sim \mathrm{N}(0,1)$
and $\gamma(0) = \gamma(1) = 0$
e.g. $\gamma(t) = \sqrt{t(1-t)}$

Proposition:

$\rho(t,x)$ satisfies a transport equation as before, with $b(t,x)$ of the form

$$b(t,x) = \mathbb{E}\left[\partial_t I(t, x_0, x_1) + \partial_t \gamma(t) z \,|\, x(t) = x\right]$$

Moreover, the score of $\rho(t,x)$ is given by

$$s(t,x) = -\gamma(t)^{-1} \mathbb{E}\left[z \,|\, x(t) = x\right]$$

which minimizes

$$L[\hat{s}] = \int \mathbb{E}\left[\frac{1}{2} |\hat{s}(t,x_t)|^2 + \gamma(t)^{-1} z \cdot \hat{s}(t,x_t)\right] dt$$

Unifying flow-based and diffusion-based generative models

MSA *& Vanden-Eijnden arXiv:2209.15571 (2022)*
MSA *& Boffi, Vanden-Eijnden arXiv:2303.08797 (2023)*

Transport equation

$$\partial_t \rho + \nabla \cdot (b\rho) = 0$$

ODE

$$\frac{d}{dt} X_t = b\left(t, X_t\right)$$

Learn $\hat{b}$

Fokker-Planck Equations

$$\partial_t \rho + \nabla \cdot (b^{\mathrm{F/B}} \rho) = \epsilon \Delta \rho$$

where $b^{\mathrm{F/B}} = b \pm \epsilon s$

SDE

$$dX_t^{\mathrm{F/B}} = b_{\mathrm{F/B}}\left(t, X_t^{\mathrm{F}}\right) dt + \sqrt{2\epsilon}\, dW_t^{\mathrm{F/B}}$$

Learn $\hat{b}_{\mathrm{F/B}}$

Are there fundamental differences between stochastic deterministic generative models?

Bounding the KL between ρ and $\hat{\rho}$

***MSA**, Boffi, Vanden-Eijnden arXiv:2303.08797 (2023);*

If $\hat{\rho}$ the density pushed by *estimated* deterministic dynamics $\hat{b}$, then

$$\partial_t \hat{\rho} + \nabla \cdot (\hat{b}\hat{\rho}) = 0$$

$$\mathrm{KL}(\rho(1) \| \hat{\rho}(1)) = \int_0^1 \int_{\mathbb{R}^d} (\nabla \log \hat{\rho} - \nabla \log \rho) \cdot (\hat{b} - b)\rho \, dx \, dt$$

matching b's does not bound KL, Fisher is uncontrolled by small error in $\hat{b} - b$

Bounding the KL between ρ and $\hat{\rho}$

***MSA**, Boffi, Vanden-Eijnden arXiv:2303.08797 (2023);*

If $\hat{\rho}$ the density pushed by *estimated* deterministic dynamics $\hat{b}$, then

$$\partial_t \hat{\rho} + \nabla \cdot (\hat{b}\hat{\rho}) = 0$$

$$\mathrm{KL}(\rho(1)\|\hat{\rho}(1)) = \int_0^1 \int_{\mathbb{R}^d} (\nabla \log \hat{\rho} - \nabla \log \rho) \cdot (\hat{b} - b)\rho \, dx \, dt$$

matching b's does not bound KL, Fisher is uncontrolled by small error in $\hat{b} - b$

If $\hat{\rho}$ the density pushed by *estimated* stochastic dynamics $\hat{b}_{\mathrm{F}} = \hat{b} + \epsilon s$, then

$$\partial_t \hat{\rho} + \nabla \cdot (b^{\mathrm{F}}\hat{\rho}) = \epsilon \Delta \hat{\rho}$$

$$\mathrm{KL}(\rho(1)\|\hat{\rho}(1)) \leq \frac{1}{4\epsilon} \int_0^1 \int_{\mathbb{R}^d} \left| \hat{b}_{\mathrm{F}} - b_{\mathrm{F}} \right|^2 \rho \, dxdt$$

$\hat{b}_{\mathrm{F}} - b_{\mathrm{F}}$ does control KL divergence

ODE vs SDE, numerical experiments

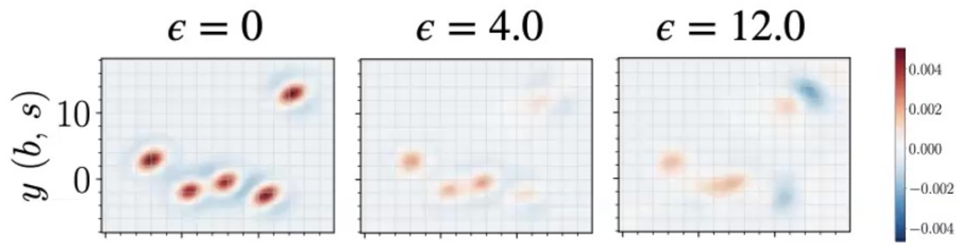
What does this mean practically?

Theory says

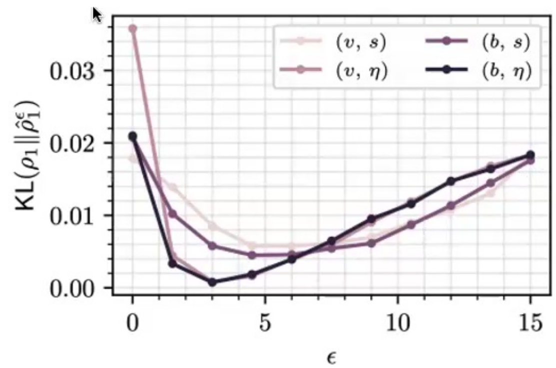
$$\epsilon^* = \left(\frac{L_b[\hat{b}] - \min_{\hat{b}} L_b[\hat{b}]}{L_s[\hat{s}] - \min_{\hat{s}} L_s[\hat{s}]} \right)^{1/2}$$

128 dimensional Gaussian Mixtures

$\hat{\rho}_1(x, y) - \rho_1(x, y)$: Error in kernel density estimate of 2D cross section



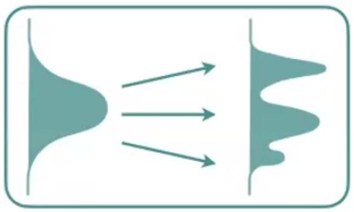
KL for learned $\hat{b}, \hat{s}$ minimal around $\epsilon \approx 5.0$, then increases



*SDE dominance not necessarily generalize to images

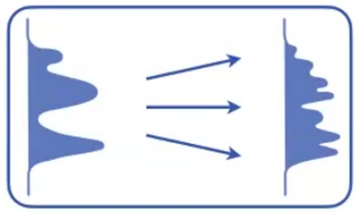
Context and Applications

Generative modeling



Ex. Image generation
Ex. Statistical physics

Domain Adaptation



Ex. Translation
Ex. Superresolution

Forecasting

Ex. Climate/weather
Ex. Dynamical systems

We will use the ***design flexibility of the interpolant*** *and the* ***coupling between*** x_0, x_1 *to approach various problems*

27

Context: Relation to Score-Based Diffusion (SBDM)

Song et al. arXiv:2011.13456 (2021)
Sohl-Dickstein et al arXiv:1503.03585 (2021)
Hyvärinen JMLR **6** (2005)
Vincent, Neural Comp. **23**, 1661 (2011)

SBDM introduces a noising process

$$dX_t = -X\,dt + \sqrt{2}\,dW_t \longrightarrow$$

$$dX_t^B = -X_t\,dt + \nabla \log \rho(t, X_t)\,dt + \sqrt{2}\,dW_t \longleftarrow$$

Generative model

$$\hat{s}(t,x) \approx \nabla \log \rho(t,x)$$

When recasted as an interpolant:

$$x(t) = x_0 e^{-t} + \sqrt{1 - e^{-2t}}z, \quad x_0 \sim \rho_0, \quad z \sim \mathrm{N}(0, Id), \quad t \in [0, \infty)$$

These coefficients are fixed by the noising process above

Only maps to a Gaussian and does so in infinite time

SBDM is but one possible interpolant!

Example: Interpolants for image generation

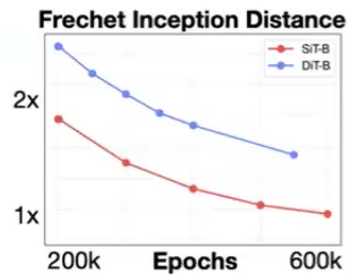
MSA *& EVE arXiv:2209.15571 (2022);*
NM, MG, ***MSA,*** *NB, EVE, SX arXiv:2401.08740 (2024)*

Freedom to choose α, β in:

$$x(t) = \alpha(t)x_0 + \beta(t)x_1$$

to reduce transport cost:

$$C[b] = \int_0^1 \mathbb{E}[\,|\,b(t,x)\,|^2\,]dt$$



Model	Params(M)	Training Steps	FID ↓
DiT-S	33	400K	68.4
SiT-S	33	400K	**57.6**
DiT-B	130	400K	43.5
SiT-B	130	400K	**33.5**
DiT-L	458	400K	23.3
SiT-L	458	400K	**18.8**
DiT-XL	675	400K	19.5
SiT-XL	675	400K	**17.2**
DiT-XL	675	7M	9.6
SiT-XL	675	7M	**8.6**
DiT-XL (cfg=1.5)	675	7M	2.27
SiT-XL (cfg=1.5)	675	7M	**2.06**

Freedom to choose $\epsilon(t)$ in:

$$dX_t^{\mathrm{F}} = b_{\mathrm{F}}\,dt + \sqrt{2\epsilon(t)}\,dW_t^{\mathrm{F}}$$

to tighten bounds on:

$$\mathrm{D}_{KL}(\hat{\rho}_1\,||\,\rho_1)$$

Systematic improvements to methods underlying, e.g. Sora (OpenAI, 2024)

Designing different ***couplings***

One is free to construct a variety of couplings, following the rules!

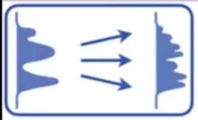
For any coupling (x_0, x_1) and any conditioning set ξ, the joint must marginalize

$$\int_{\mathbb{R}^d} \rho\left(x_0, x_1 \mid \xi\right) dx_1 = \rho_0\left(x_0 \mid \xi\right), \quad \int_{\mathbb{R}^d} \rho\left(x_0, x_1 \mid \xi\right) dx_0 = \rho_1\left(x_1 \mid \xi\right).$$

Recent example in literature: minibatch OT (**Tong** et al (2023), **Pooladian** et al 2023)

Rather than use an approximate algorithm for constructing such couplings, there are *many* that we have natural access to

Example: Data-dependent coupling



***MSA**, MG, NB, RR, EVE arXiv:2310.03725 (2023)*
***MSA**, NB, ML, EVE arXiv:2310.03695 (2023)*

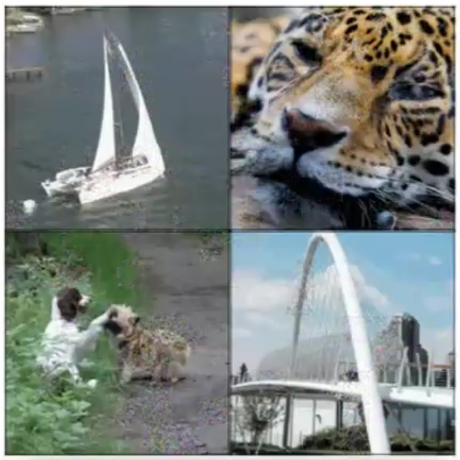
What if one x_0 is coupled to another x_1?

$$\rho(x_0, x_1) = \rho_1(x_1)\rho_0(x_0 \mid x_1)$$

In-painting

x_0 a masked image

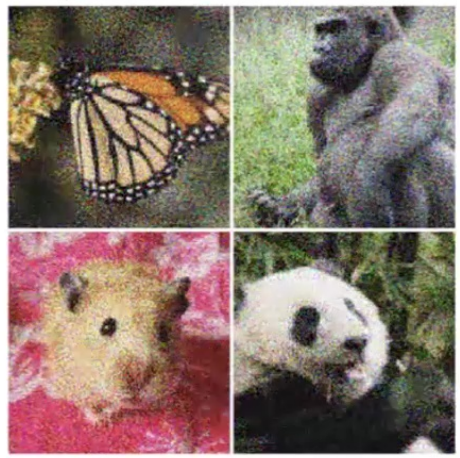
$b(t, x)$ invariant in unmasked areas



Super-resolution

x_0 a low-res image

x_0 now *proximal* to its target



Frechet Inception Distance

Model	Train	Valid
Improved DDPM (Nichol & Dhariwal, 2021)	12.26	–
SR3 (Saharia et al., 2022)	11.30	5.20
ADM (Dhariwal & Nichol, 2021)	7.49	3.10
Cascaded Diffusion (Ho et al., 2022a)	4.88	4.63
I^2SB (Liu et al., 2023a)	–	2.70
Dependent Coupling (**Ours**)	**2.13**	**2.05**

More efficient and better performance across tasks

MSA*, *Goldstein*, Boffi, Ranganath, Vanden-Eijnden 2310.03725 (2023)*

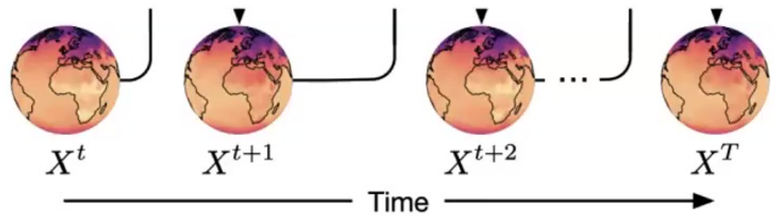
Designing different ***new types of maps***

Use interpolant blueprint to learn coefficients of new types of generative models

Example: are there processes that allow me to predict ensembles of future events given just one condition?

Weather

Dynamical systems



What specific processes can we construct to meet these goals?

32

Parameterizing the Föllmer process

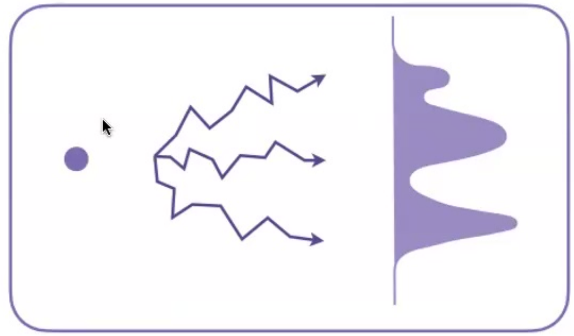
Interpolant

$$x(t) = \alpha(t)x_0 + \beta(t)x_1 + \sigma(t)\sqrt{t}z$$

$(x_0, x_1) \sim \rho(x_0, x_1)$, $\quad z \sim \mathrm{N}(0, \mathrm{I}_d)$ with $z \perp (x_0, x_1)$

Reference Dynamics

$$r(t) = \dot{\alpha}(t)x_0 + \dot{\beta}(t)x_1 + \dot{\sigma}(t)\sqrt{t}z$$



$x_0 \sim \rho_0$ — **"Weather at time s"**

$x_1 \sim \rho_c(x_1 \mid x_0)$ — **"Ensemble of weather predictions for time $s + \Delta t$"**

Learning $b(t, x, x_0) = \mathbb{E}[r(t) \mid x(t) = x, x_0]$ solves the SDE

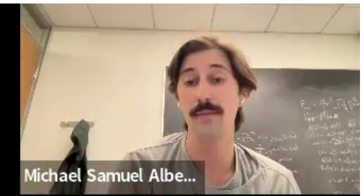
$$dX_t = b(t, X_t, x_0)dt + \sigma(t)dW_t$$

such that:

$$\mathit{Law}(X_s) = \mathit{Law}(x_t \mid x_0), \textit{ with } X_{s=1} \sim \rho_c(x_1 \mid x_0)$$

Example: Probabilistic forecasting

YC, MG, MH, **MSA**, NB, EVE arXiv:2402.XXXXX (2024)



Interpolants for ensembles of future events

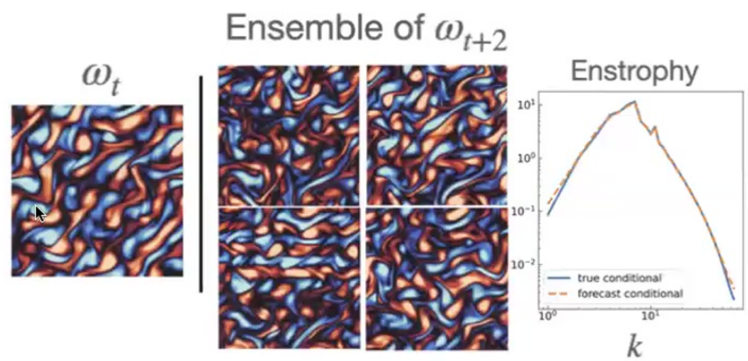
$$\rho(x_0, x_1) = \rho_0(x_0)\rho_1(x_1 \mid x_0)$$

Navier Stokes

Evolution of the vorticity ω

Map ω_t to distribution $\rho(\omega_{t+\tau} \mid \omega_t)$

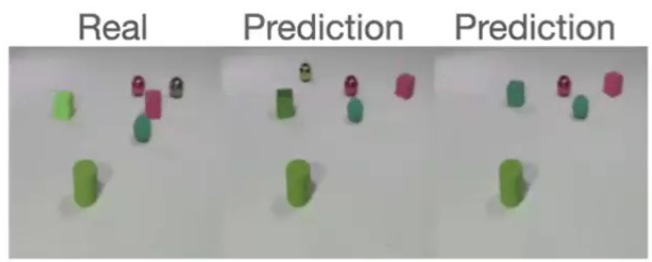
Choose NS w/ random forcing that has invariant measure



Video completion

Map x_t to distribution $\rho(x_{t+1} \mid x_{t-\tau:t})$

Roll out subsequent frames



Multimarginal Interpolants

The learning paradigm behind interpolants (and diffusions!) can be independent of interpolation schedule (e.g. noise schedule)

Generic 2-marginal interpolant, with *interpolation coordinates* $\alpha = [\alpha_0, \alpha_1]$

$$x(t) = \alpha_0(t)x_0 + \alpha_1(t)x_1$$

gives velocity field

$$b(t,x) = \mathbb{E}\left[\dot{x}(t) \mid x(t) = x\right] = \dot{\alpha}_0(t)\mathbb{E}\left[x_0 \mid x(t) = x\right] + \dot{\alpha}_1(t)\mathbb{E}\left[x_1 \mid x(t) = x\right]$$

call $g_0(t,x)$

call $g_1(t,x)$

All you need to learn are conditional expectations, learned on an interval.
To use it in an ODE, you can choose a time parameterization after

ODE vs SDE, numerical experiments

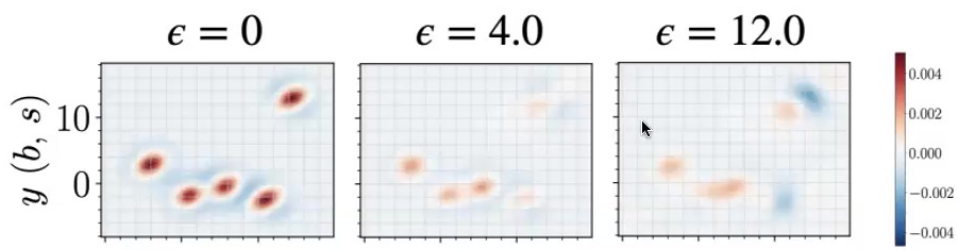
What does this mean practically?

Theory says

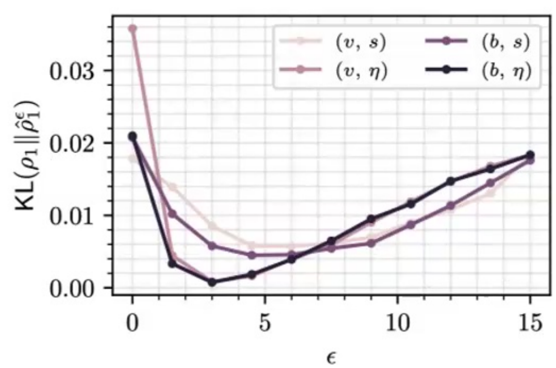
$$\epsilon^* = \left(\frac{L_b[\hat{b}] - \min_{\hat{b}} L_b[\hat{b}]}{L_s[\hat{s}] - \min_{\hat{s}} L_s[\hat{s}]} \right)^{1/2}$$

128 dimensional Gaussian Mixtures

$\hat{\rho}_1(x, y) - \rho_1(x, y)$**: Error in kernel density estimate of 2D cross section**



KL for learned $\hat{b}, \hat{s}$ minimal around $\epsilon \approx 5.0$, then increases



*SDE dominance not necessarily generalize to images

April 11, 2024 26

Multimarginal Interpolants

The learning paradigm behind interpolants (and diffusions!) can be independent of interpolation schedule (e.g. noise schedule)

Generic 2-marginal interpolant, with *interpolation coordinates* $\alpha = [\alpha_0, \alpha_1]$

$$x(t) = \alpha_0(t)x_0 + \alpha_1(t)x_1$$

gives velocity field

$$b(t,x) = \mathbb{E}\left[\dot{x}(t) \mid x(t) = x\right] = \dot{\alpha}_0(t)\mathbb{E}\left[x_0 \mid x(t) = x\right] + \dot{\alpha}_1(t)\mathbb{E}\left[x_1 \mid x(t) = x\right]$$

call $g_0(t,x)$ call $g_1(t,x)$

All you need to learn are conditional expectations, learned on an interval.
To use it in an ODE, you can choose a time parameterization after

Multimarginal Interpolant

***MSA**, Boffi, Lindsey, Vanden-Eijnden, arXiv:2310.03695 (2023)*

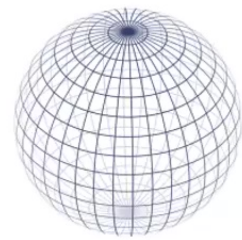
But also note !

Nothing is stopping you from interpolating between **more** densities

$$x(\alpha) = \sum_{k=0}^{K} \alpha_k x_k$$

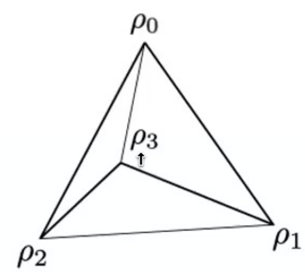
The minimal conditions are

$$\sum_k \alpha_k^2 > 0, \qquad \sum_k \alpha_k^2 < C^2$$



α a coordinate vector on the surface of, or within an $K-$sphere of radius C

Slightly more pragmatic (?) condition: $\sum_k \alpha_k = 1$



α a coordinate vector on the $K-$simplex

Multimarginal Interpolants

The learning paradigm behind interpolants (and diffusions!) can be independent of interpolation schedule (e.g. noise schedule)

Generic 2-marginal interpolant, with *interpolation coordinates* $\alpha = [\alpha_0, \alpha_1]$

$$x(t) = \alpha_0(t)x_0 + \alpha_1(t)x_1$$

gives velocity field

$$b(t,x) = \mathbb{E}\left[\dot{x}(t) \mid x(t) = x\right] = \dot{\alpha}_0(t)\mathbb{E}\left[x_0 \mid x(t) = x\right] + \dot{\alpha}_1(t)\mathbb{E}\left[x_1 \mid x(t) = x\right]$$

call $g_0(t,x)$ call $g_1(t,x)$

All you need to learn are conditional expectations, learned on an interval.
To use it in an ODE, you can choose a time parameterization after

35

Multimarginal Interpolant

__MSA__, Boffi, Lindsey, Vanden-Eijnden, arXiv:2310.03695 (2023)

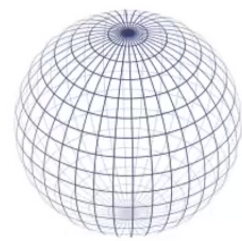
But also note !

Nothing is stopping you from interpolating between **more** densities

$$x(\alpha) = \sum_{k=0}^{K} \alpha_k x_k$$

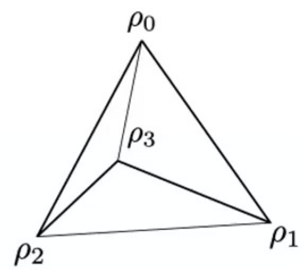
The minimal conditions are

$$\sum_k \alpha_k^2 > 0, \qquad \sum_k \alpha_k^2 < C^2$$



α a coordinate vector on the surface of, or within an $K-$sphere of radius C

Slightly more pragmatic (?) condition: $\sum_k \alpha_k = 1$



α a coordinate vector on the $K-$ simplex

Multimarginal Interpolant

__MSA__, Boffi, Lindsey, Vanden-Eijnden, arXiv:2310.03695 (2023)

Definition

The barycentric interpolant $x(\alpha)$ with $\alpha = (\alpha_0, \ldots, \alpha_K) \in \Delta^K$ is the stochastic process

$$x(\alpha) = \sum_{k=0}^{K} \alpha_k x_k$$

where $(x_1, \ldots, x_K)$ are drawn from $\rho(x_1, \ldots, x_K)$ and we set $x_0 \sim N(0, Id_d)$ drawn independently (x_0 needed if you want score function).

Generalized continuity equation

The probability distribution of $x(\alpha)$ has a density $\rho(\alpha, x)$ which satisfies $K + 1$ continuity equations

$$\partial \alpha_k \rho(\alpha) \nabla_x \cdot (g_k(\alpha, x)\rho(\alpha)) = 0$$

where $g_k(\alpha, x)$ is the conditional expectation $\mathbb{E}[x_k \,|\, x(\alpha) = x]$

Multimarginal Interpolant

***MSA**, Boffi, Lindsey, Vanden-Eijnden, arXiv:2310.03695 (2023)*

Once you have learned the g_k, you can any path on the simplex as a generative model from *any* ρ_i to *any* ρ_j

- Just choose a parameterization of $\alpha(t)$ for $t \in [0,1]$ that starts and ends at one of the marginal densities, e.g.

$$\alpha(t=0) = [1,0,\ldots,0] \text{ and } \alpha(t=1) = [0,\ldots,1,\ldots,0]$$

Velocity field	$b(t,x) = \sum_{k=0}^{K} \dot{\alpha}_k(t) g_k(\alpha(t), x)$
Probability flow ODE	$\dot{X}_t = b(t, X_t)$

April 11, 2024 38

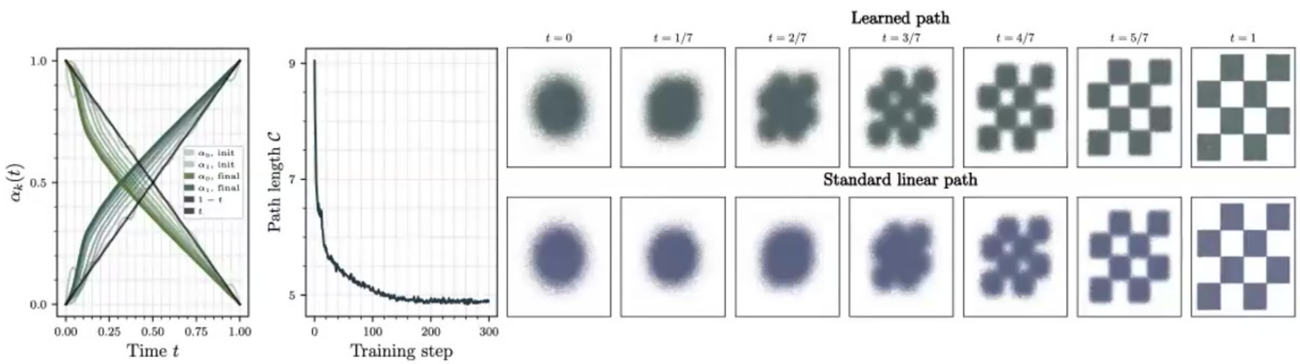
A geometric algorithm for selecting a performant α

Much effort has gone into choosing an appropriate noise schedule for diffusions. The multimarginal picture gives a straightforward algorithm

$$C_\alpha(\hat{\alpha}) = \min_{\hat{\alpha}} \int_0^1 \mathbb{E}\left[\left|\sum_{k=0}^{K} \hat{\alpha}_k(t) g_k(\hat{\alpha}(t), x(\hat{\alpha}(t)))\right|^2\right] dt$$

Riemannian geometric "path length" depends on $\alpha(t)$

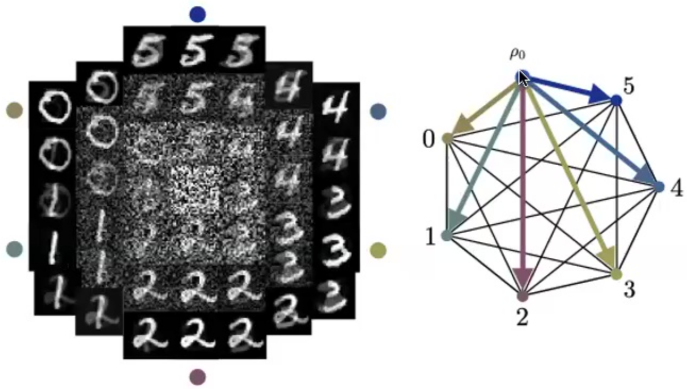
- **reduce transport cost over restricted class to learn better $b(t,x)$.**
- **Extremely simple optimization of $\hat{\alpha}$**



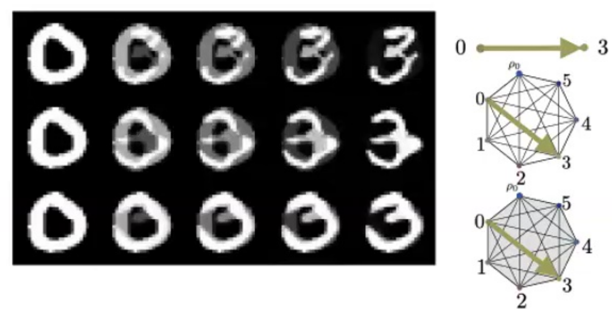
Applications, different paths on simplex

6 classes of MNIST digits as marginals

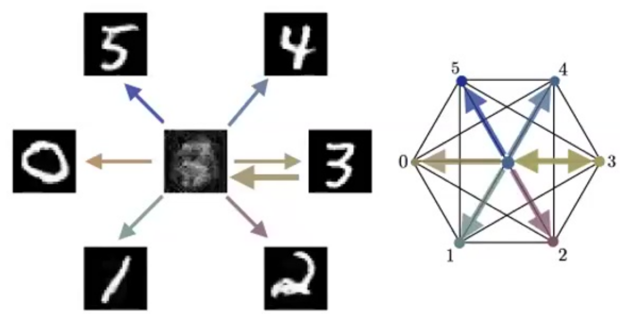
Generate from same initial condition to anywhere on 6-simplex



More natural style-transfer by learning on whole simplex



Sample from the empirical barycenter



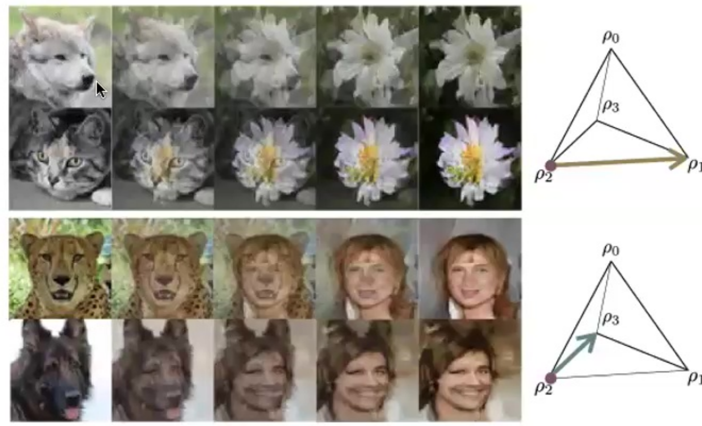
April 11, 2024

40

More Applications

Simultaneous access to marginal sampling

Natural style transfer



Part 1 summary

Laying out some tools to work with dynamical measure transport and generative modeling

Approaching some of these topics from an applied maths perspective can give some better control on performance and methods