Understand score-based generative models via lens of Wasserstein proximal operators

Siting Liu
University of California, Los Angeles

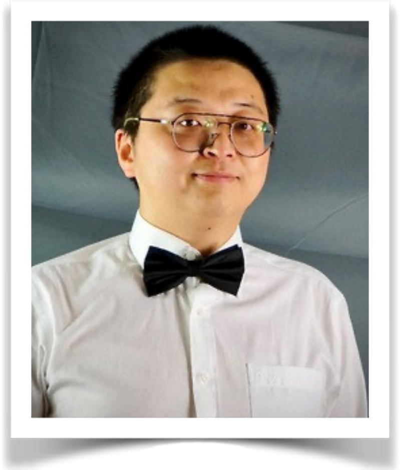
Woman in Optimal Transport

April 18th

With B. Zhang, M. Katsoulakis (University of Massachusetts, Amherst), W. Li (University of South Carolina), S. Osher (University of California, Los Angeles)

B. Zhang

M. Katsoulakis



W. Li

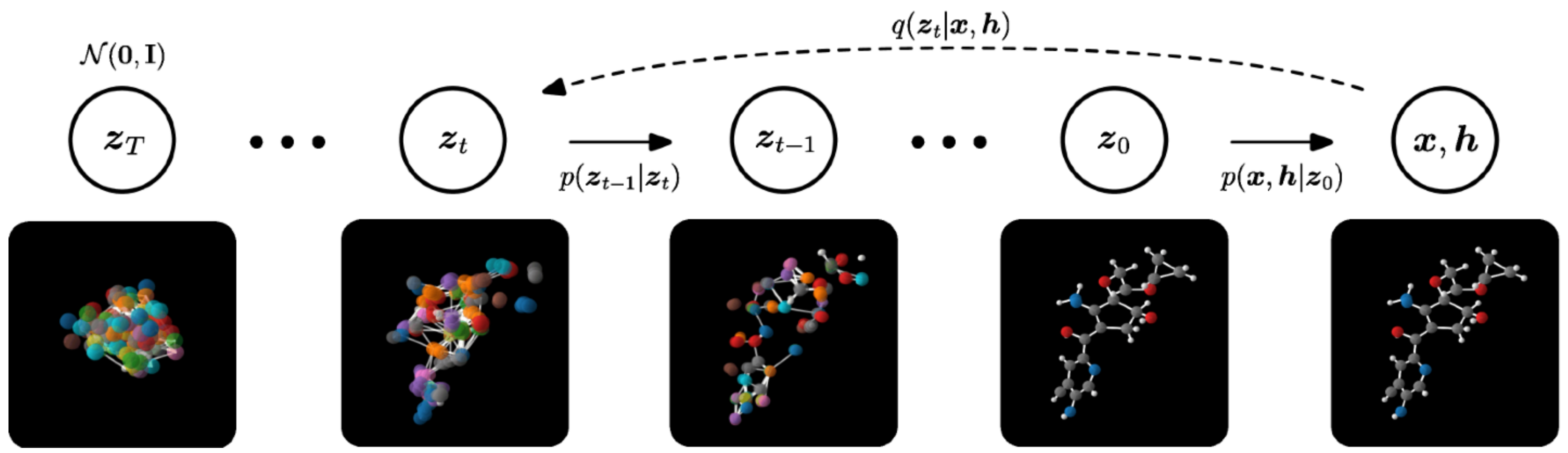
S. Osher

Zhang, Benjamin J., Siting Liu, Wuchen Li, Markos A. Katsoulakis, and Stanley J. Osher. "Wasserstein proximal operators describe score-based generative models and resolve memorization." *arXiv:2402.06162* (2024).

Generative models

Stable diffusion

Sora

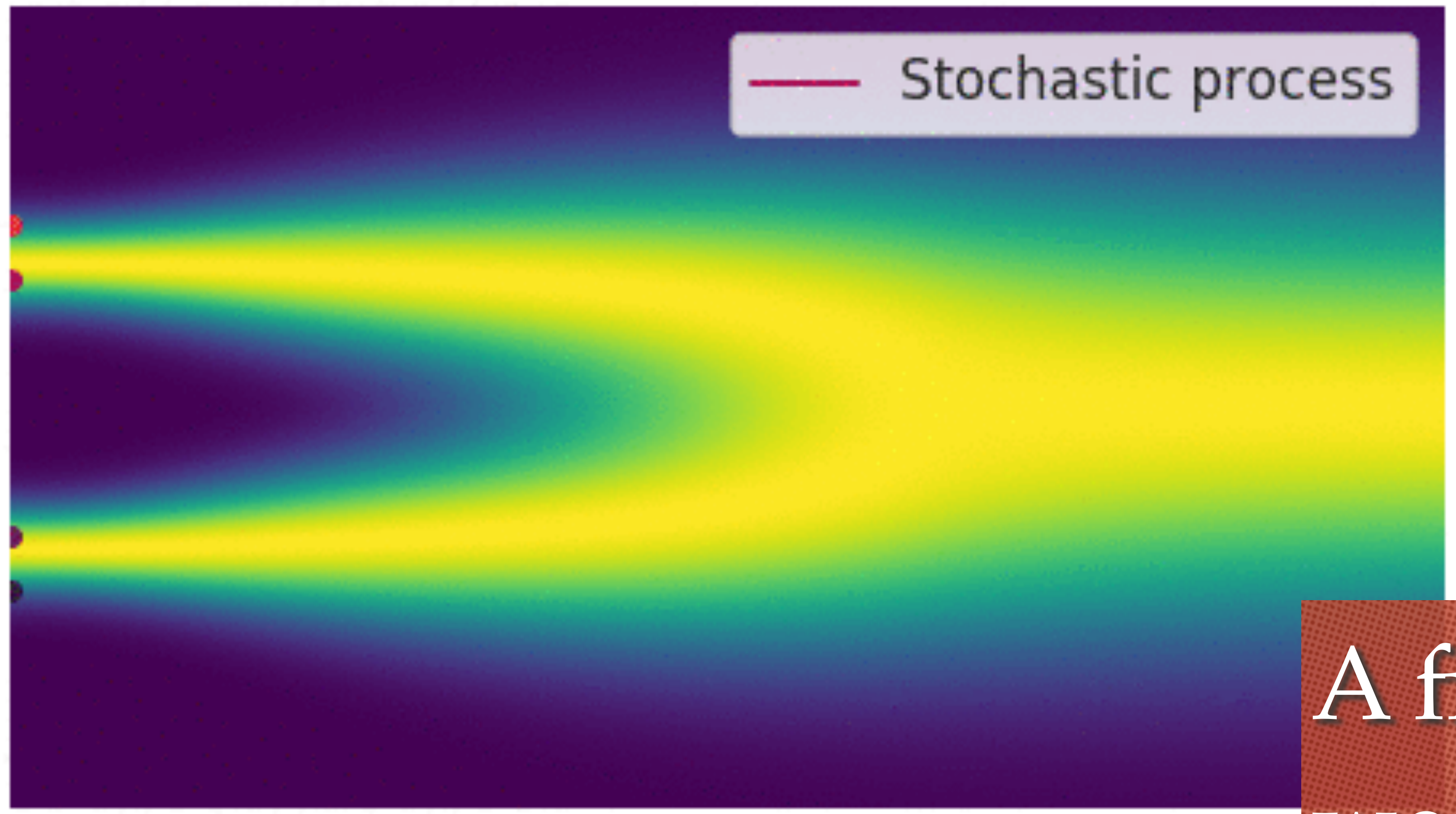


Molecular generation
Hoogeboom et al. 2022

Goal:

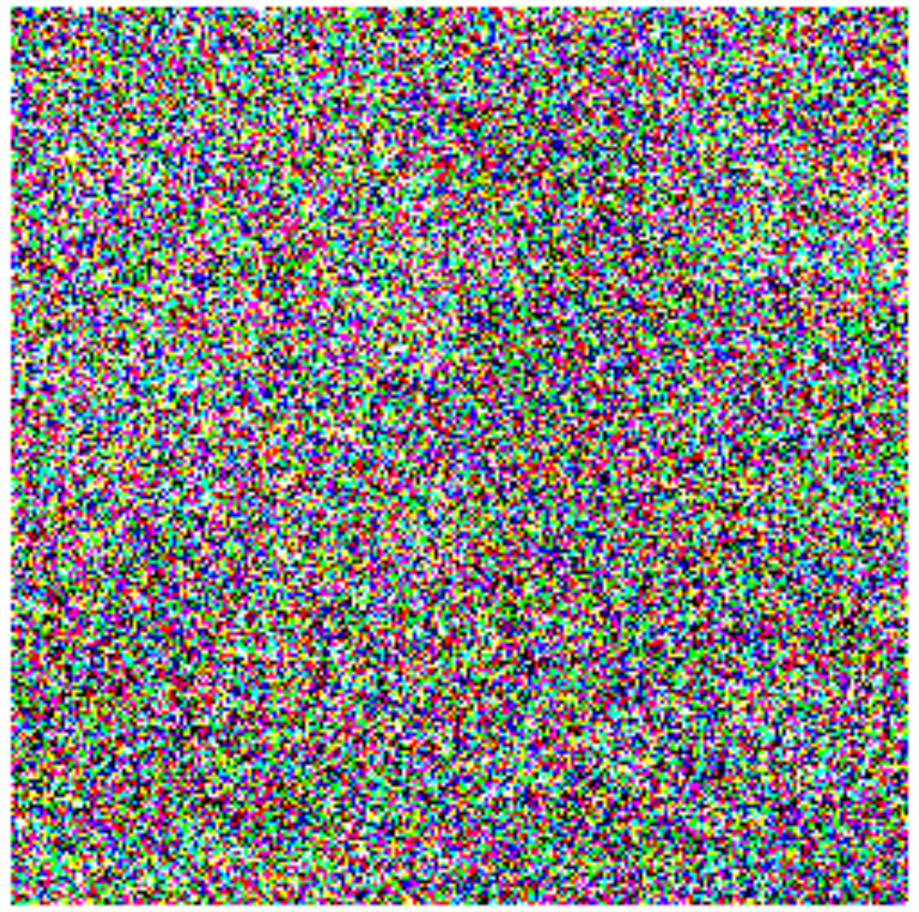
- given samples $\{x_i\}_{i=1}^N$ from some unknown distribution π
- generate more samples from the same measure

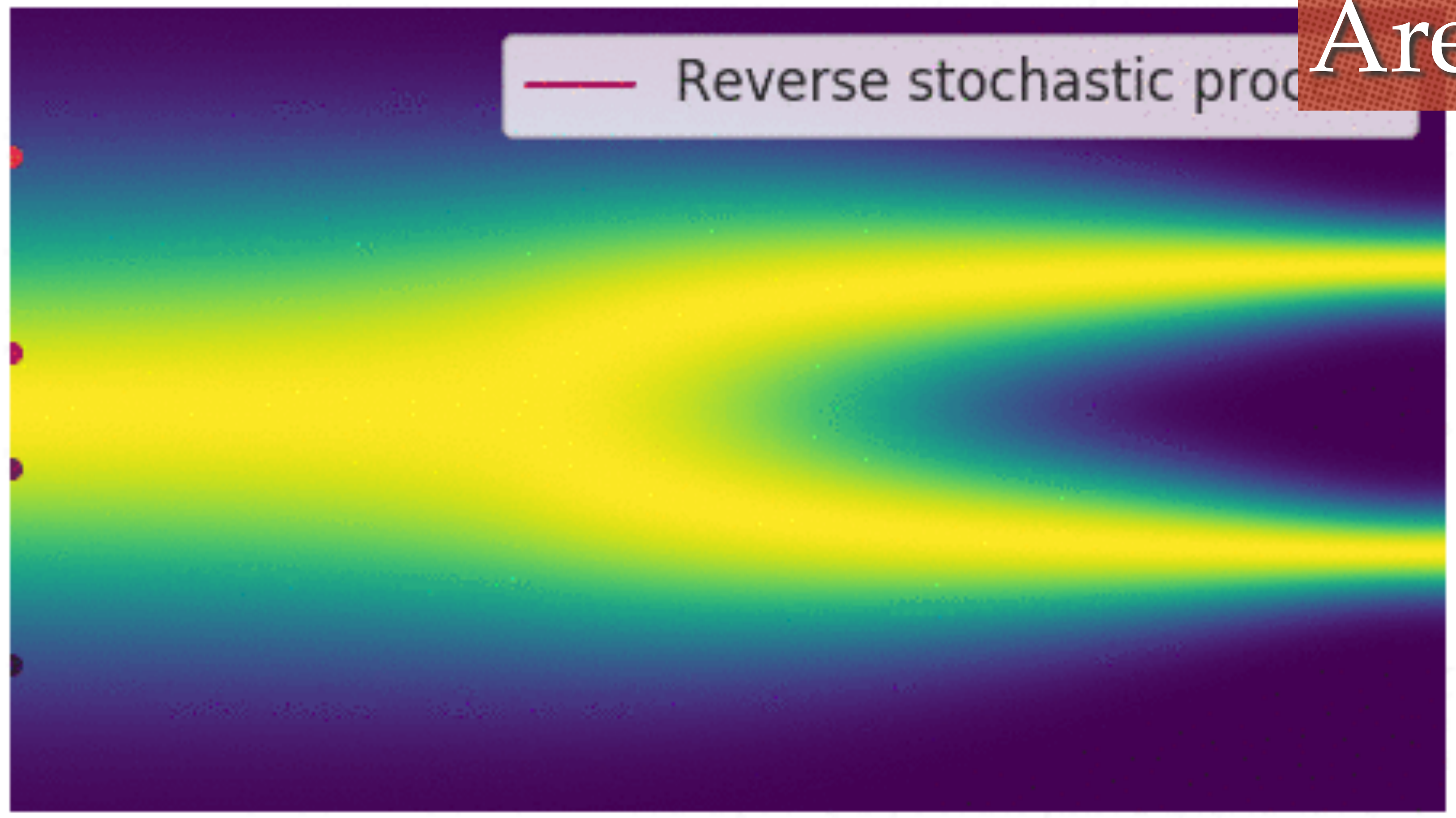
Diffusion-based generative models



Perturbing data to noise with a continuous-time stochastic process.

A figure is a data point $x \in R^d$,
we apply diffusion process by adding noise.
Are we reversing a heat equation?





Generate data from noise by **reversing the perturbation procedure.**

Videos from Song Yang

Score-based generative model (SGM)

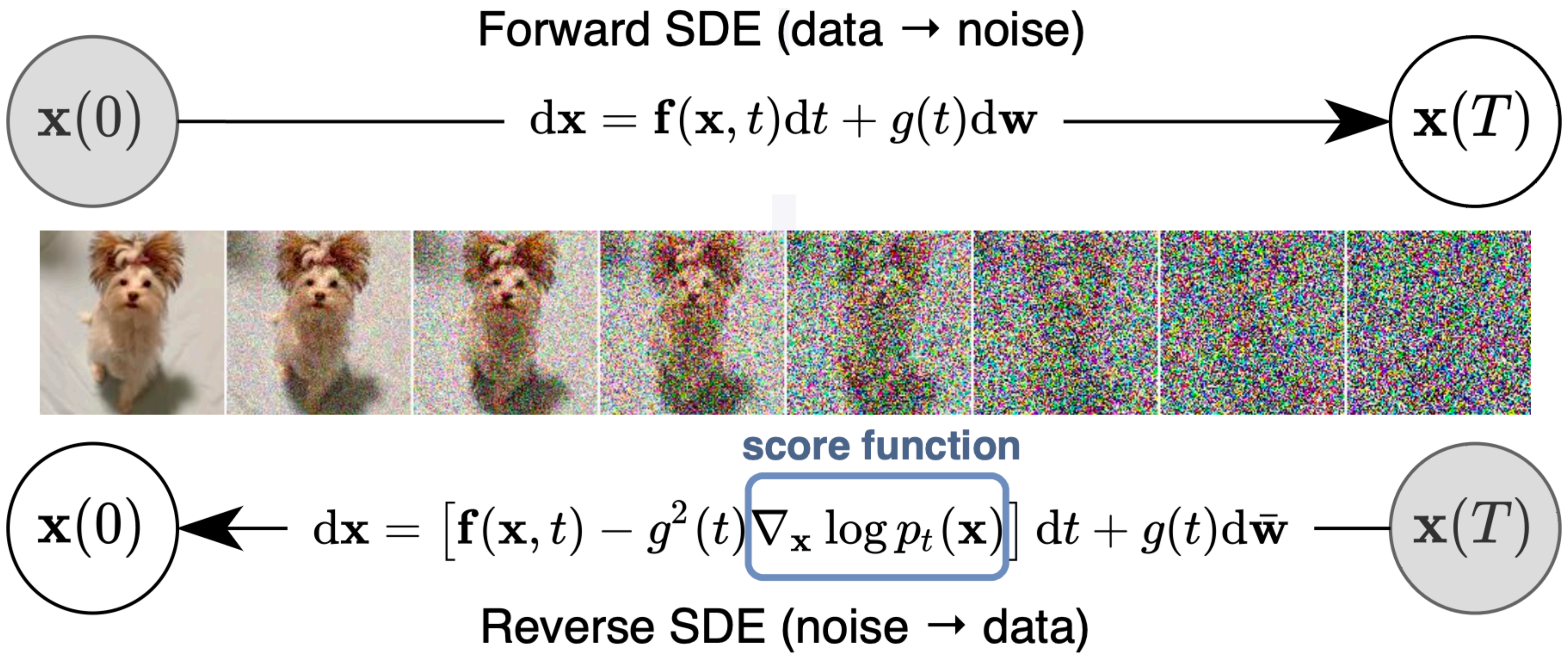


Image from Song Yang

Example 1:

Forward SDE (data → noise)
$dx = \sigma dW$

Reverse SDE (noise → data)
$dx = -\sigma^2 s dt + \sigma dW$

Example 2:

Forward SDE (data → noise)
Ornstein-Uhlenbeck Process

Reverse SDE (noise → data)
$dx = -\sigma^2 s dt + \sigma dW$

Song, Y., et al. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Score-based generative model (SGM)

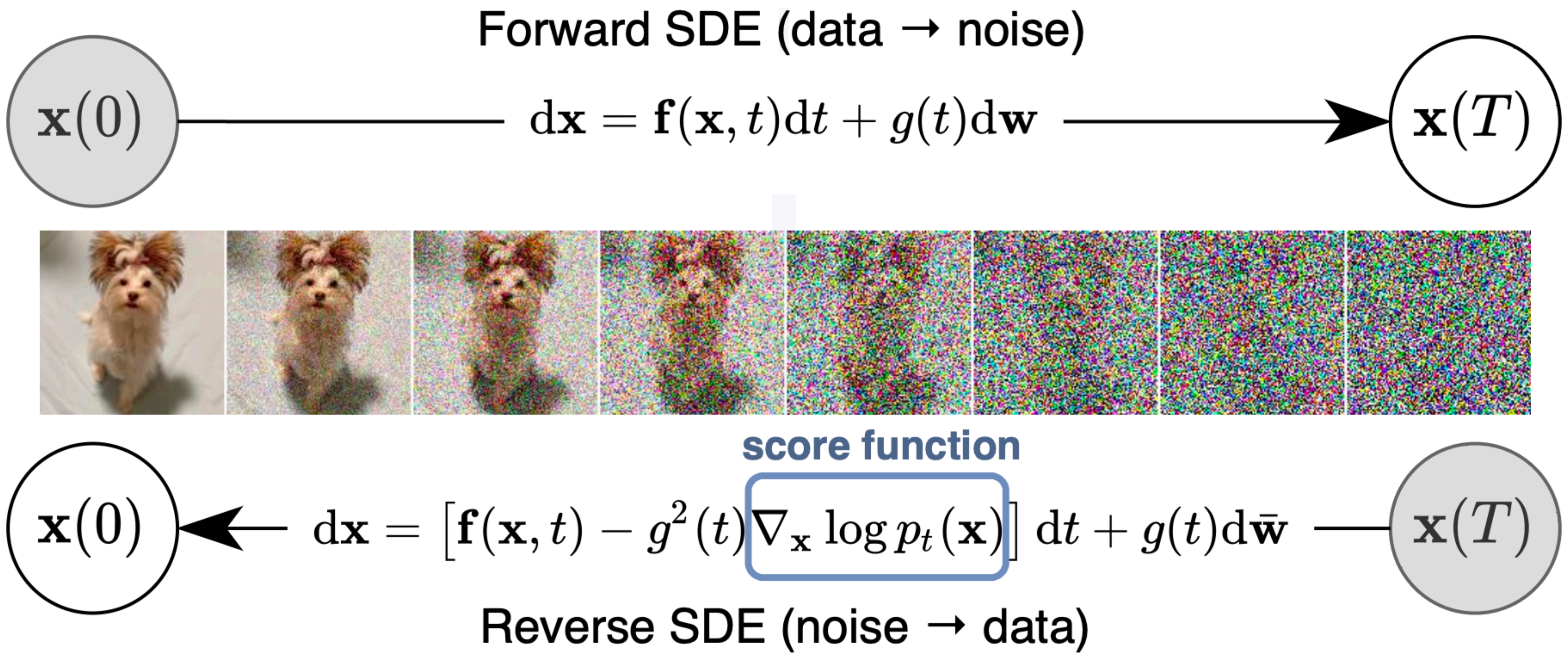


Image from Song Yang

- Reversing guided by score function $s(x,t) = \nabla_x \log p(x,t)$, p: probability density function.
- If we know the **score of the distribution at each intermediate time step**, we can generate samples from noise.
- Use neural net $s_\theta : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ is trained by minimizing a score-matching loss function.

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

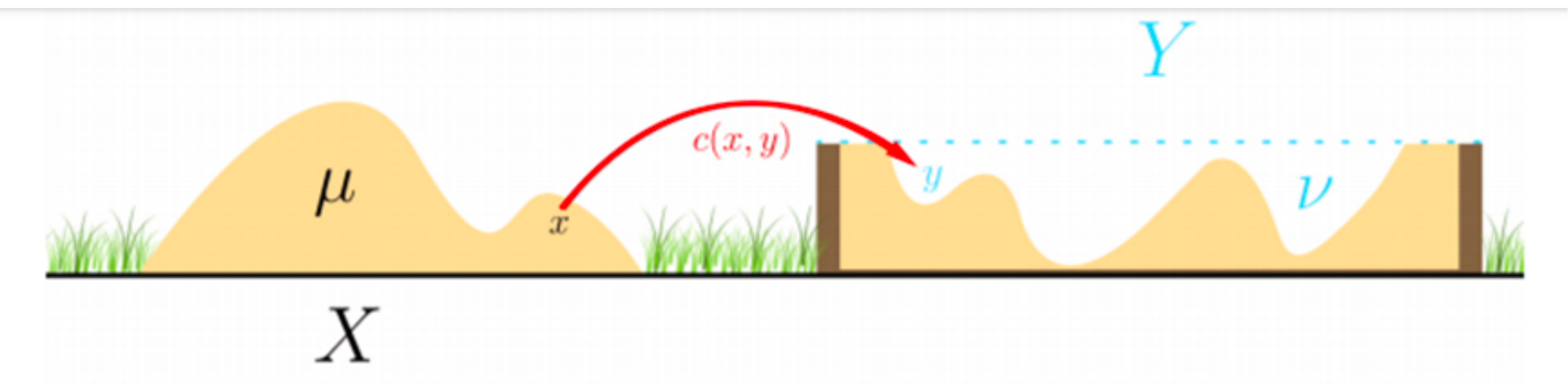
$$\min_\theta C_{ISM}(\theta) = \min_\theta \int_0^T \int_{\mathbb{R}^d} \sigma(T-s)^2 \left[\frac{1}{2}\|s_\theta(y,s)\|^2 + \nabla \cdot s_\theta(y,s)\right] \eta(y,s)\, dy\, ds$$

Fundamental mathematical nature of SGMs

- A fundamental characterization of score-based generative models as **Wasserstein proximal operators (WPO) of cross-entropy**
- Mean-field games build a bridge to mathematically equivalent alternative formulations of SGM
- Yield**s explainable** formulations of SGMs grounded in theories of information, optimal transport, manifold learning, and optimization
- Uncovering mathematical structure of SGMs explains **memorization**, and informs practical models to **generalize better**; suggests new practical models with interpretable mathematically-informed structure that **train faster** with **less data.**

Optimal Transport and Wasserstein metric

- Wasserstein metric is a distance function defined between **probability distributions**, also known as earth mover's distance $W(\mu, \nu)$



- **Monge**: soil-transportation problem; **Kantorovich**: applications in plywood industry
- Applications: Economics, Industrial Engineering, Data Sciences, etc.
- By **Benamou-Brenier**, *A computational fluid mechanics solution of the Monge-Katonrovich mass transfer problem*

$$\inf_{\rho, v} \left\{ \int_0^1 \int_\Omega \frac{1}{2} \rho(x,t) \|v(x,t)\|^2 dx dt \right\}$$

$$\text{s.t. } \rho_t + \nabla \cdot (\rho v(x,t)) = 0, \ \rho(x,0) = \mu(x), \ \rho(x,1) = \nu(x)$$

Wasserstein proximal operator

- Given a probability density ρ_0 , we consider the Wasserstein proximal operator (WPO) of the some function $V(x)$:

$$\rho := \mathrm{WProx}_{\tau V}(\rho_0) := \arg \min_{q \in \mathscr{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} V(x) q(x) dx + \frac{W(\rho_0, q)^2}{2\tau}$$

where $W(\rho_0, q)$ is the Wasserstein-2 distance.

- Set $V(x) = -\log \pi(x)$ of a distribution π, the first term is the cross-entropy of π with respect to ρ.

ρ_0 (source) $\mapsto$ π (target), redistribution + transport

Wasserstein proximal operator

- Given a probability density ρ_0 , we consider the Wasserstein proximal operator (WPO) of the some function $V(x)$:

$$\rho := \mathrm{WProx}_{\tau V}(\rho_0) := \arg \min_{q \in \mathscr{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} V(x)q(x)dx + \frac{W(\rho_0, q)^2}{2\tau}$$

where $W(\rho_0, q)$ is the Wasserstein-2 distance.

- Computing the WPO requires solving an optimization problem.

- Equivalent to solving the following variational problem

$$\inf_{\rho,v} \left\{ \int_0^1 \int_\Omega \frac{1}{2}\rho(x,t)\|v(x,t)\|^2 dxdt + \int_\Omega V(x)\rho(x,1)dx \right\}$$

$$\text{s.t. } \rho_t + \nabla \cdot (\rho v(x,t)) = 0, \ \rho(x,0) = \rho_0(x)$$

A potential mean-field game

Wasserstein proximal operator

- Given a probability density ρ_0 , we consider the Wasserstein proximal operator (WPO) of the some function $V(x)$:

$$\rho := \mathrm{WProx}_{\tau V}(\rho_0) := \arg \min_{q \in \mathscr{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} V(x)q(x)dx + \frac{W(\rho_0, q)^2}{2\tau}$$

where $W(\rho_0, q)$ is the Wasserstein-2 distance.

$$\inf_{\rho, v} \left\{ \int_0^1 \int_\Omega \frac{1}{2} \rho(x,t) \| v(x,t) \|^2 dxdt + \int_\Omega V(x)\rho(x,1)dx \right\}$$

$$\text{s.t. } \rho_t + \nabla \cdot (\rho v(x,t)) = 0, \ \rho(x,0) = \rho_0(x)$$

Optimality condition

$$\begin{cases} -\dfrac{\partial U}{\partial t} + \dfrac{1}{2} |\nabla U|^2 = 0, \ U(x,h) = V(x) \\ \dfrac{\partial \rho}{\partial t} - \nabla \cdot (\rho \nabla U) = 0, \ \rho(x,0) = \rho_0(x). \end{cases}$$

Regularized WPO

- The regularization via adding **viscosity $\beta\Delta\rho$** through the dynamic formulation of the Optimal Transport.

$$\inf_{\rho,v}\left\{\int_0^1\int_\Omega \frac{1}{2}\rho(x,t)\|v(x,t)\|^2 dxdt + \int_\Omega V(x)\rho(x,1)dx\right\}$$
$$\text{s.t. } \rho_t + \nabla\cdot(\rho v(x,t)) = \boxed{\beta\Delta\rho}, \ \rho(x,0) = \rho_0(x)$$

- Regularized WPO:

$$\rho := \mathrm{WProx}_{\tau V,\beta}(\rho_0) := \arg\min_{q\in\mathscr{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} V(x)q(x)dx + \frac{W_\beta(\rho_0,q)^2}{2\tau}.$$

- We obtain **a closed-form formulation**, which allows fast computation of the WPO.

Wuchen Li, Siting Liu and Stanley Osher, "A kernel formula for regularized Wasserstein proximal operators." *Research in the Mathematical Sciences*

Regularized WPO

- Regularized WPO:

$$\rho := \mathrm{WProx}_{\tau V,\beta}(\rho_0) := \arg \min_{q \in \mathscr{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} V(x)q(x)dx + \frac{W_\beta(\rho_0, q)^2}{2\tau}.$$

Optimality condition

$$\begin{cases} -\frac{\partial U}{\partial t} + \frac{1}{2}|\nabla U|^2 = \gamma \Delta U, \ U(x,T) = V(x) \\ \frac{\partial \rho}{\partial t} - \nabla \cdot (\rho \nabla U) = \gamma \Delta \rho, \ \rho(x,0) = \rho_0(x), \end{cases}$$

With **Cole-Hopf transform** (log-transform), G :heat kernel.

$$U(x,t) = -2\gamma \log \left(G_{\gamma,T-t} * e^{-\frac{V(x)}{2\gamma}} \right)$$

Deriving SGM from regularized WPO

- The **cross-entropy** of a distribution π with respect to μ is $H(\mu, \pi) := -\int_{R^d} \mu(x) \log \pi(x) dx$.

- Set $V(x) = -\log \pi(x)$ in WPO $\displaystyle \min_{q \in \mathscr{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} V(x) q(x) dx + \frac{W_\beta(\rho_0, q)^2}{2\tau}$.

- Via **Cole-Hopf transform** (log-transform) with a time reparametrization, we obtain the system:

Forward SDE (data → noise)
$dx = \sigma dW$

Reverse SDE (noise → data)
$dx = \sigma^2 s dt + \sigma dW$

Score!

$$\begin{cases} \frac{\partial \eta}{\partial s} = \frac{\sigma^2}{2} \Delta \eta \\ \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \sigma^2 \boxed{\nabla \log \eta}) = \frac{\sigma^2}{2} \Delta \rho \\ \eta(y, 0) = \pi(y),\ \rho(x, 0) = \rho_0(x). \end{cases}$$

SGMs are WPOs of cross-entropy

$$\pi = \mathrm{WProx}_{\sigma^2 T\mathcal{H},\sigma^2/2}(\mathrm{WProx}^{-1}_{\sigma^2 T\mathcal{H},\sigma^2/2}(\hat{\pi})),$$

backward **forward**

where samples $\{Z_i\}_{i=1}^N$ drawn from distribution π, the empirical distribution $\pi(\cdot) \approx \hat{\pi}(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i}(\cdot)$

- Reveals **forward-backward** / **noising-denoising** nature of SGMs.
- It gives the **exact score function**:

$$\hat{s}(y,s) = \frac{(\nabla_y G_s * \hat{\pi})(y)}{(G_s * \hat{\pi})(y)} = -\frac{\sum_{i=1}^N \frac{y-Z_i}{\sigma^2 s} G_s(y, Z_i)}{\sum_{i=1}^N G_s(y, Z_i)}.$$

$G_t(y, y')$ is the heat kernel.

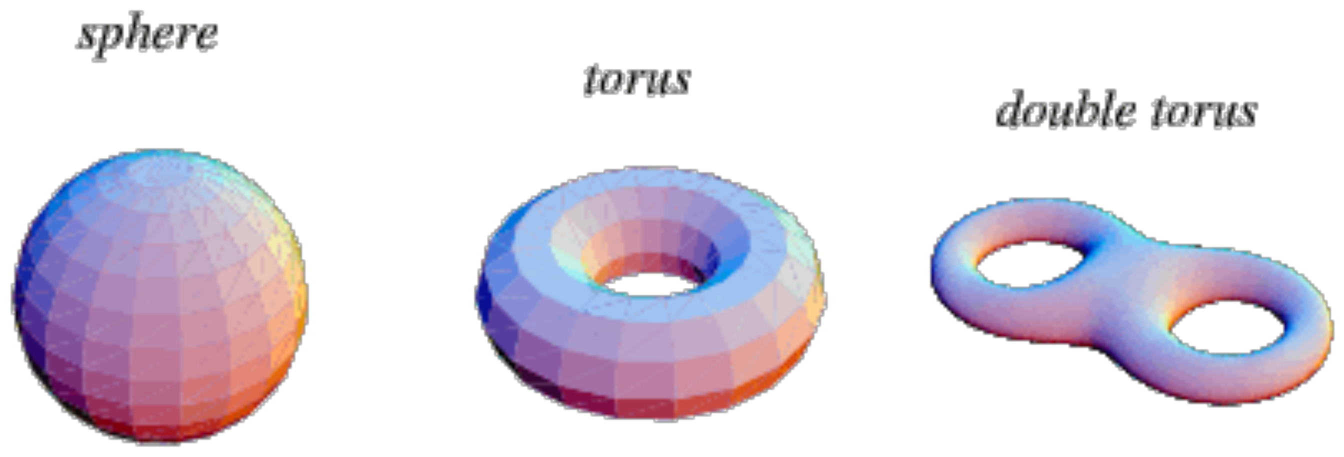
But overfit! We don't get new samples

A kernel model that generalizes

- Consider a generalization of the empirical distribution by Gaussian kernels.

$$\hat{\pi}_\theta(x; \{Z_i\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^{N} \frac{\det \mathbf{\Gamma}_\theta(Z_i)}{(2\pi)^{d/2}} \exp\left(-(x - Z_i)^\top \mathbf{\Gamma}_\theta(Z_i)(x - Z_i)\right)$$

- Learn **local covariance matrix Γ_θ** near each kernel center use neural networks.
- Enforce the terminal condition of HJ equation, which is equivalent to **implicit score-matching**.
- Learning local covariance matrix is akin to **manifold learning**, which is something SGM has been empirically observed to do.[J. Pidstrigach 2022]



Basic surfaces that are manifolds.
Figures from *Medium- Manifolds in Data Science*

Naïve kernel model

WPO-informed kernel model

Satisfies HJB alone

$$\pi(\,\cdot\,) \approx \hat{\pi}(\,\cdot\,) = \frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i}(\,\cdot\,)$$

$$\text{Reverse diffusive process with } \hat{\mathbf{s}}(\,\cdot\,,t) = \frac{(\nabla_y G_t * \hat{\pi})(\,\cdot\,)}{(G_t * \hat{\pi})(\,\cdot\,)}$$

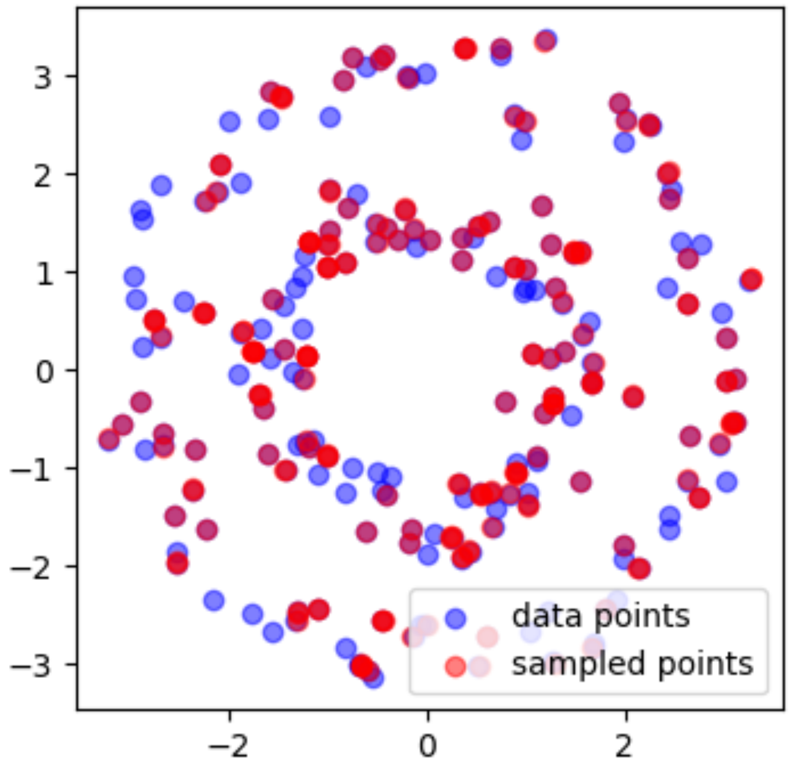
Also enforces terminal condition

$$\pi(\,\cdot\,) \approx \hat{\pi}_\theta(x;\theta,\{Z_i\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^{N} G_{t,\theta}(Z_i,\,\cdot\,)$$

$$G_{t,\theta}(Z,x) = \frac{\det \boldsymbol{\Gamma}_{T-t,\theta}(Z_i)}{(2\pi)^{d/2}} \exp\left(-(x-Z)^{\top}\boldsymbol{\Gamma}_{T-t,\theta}(Z)(x-Z)\right)$$

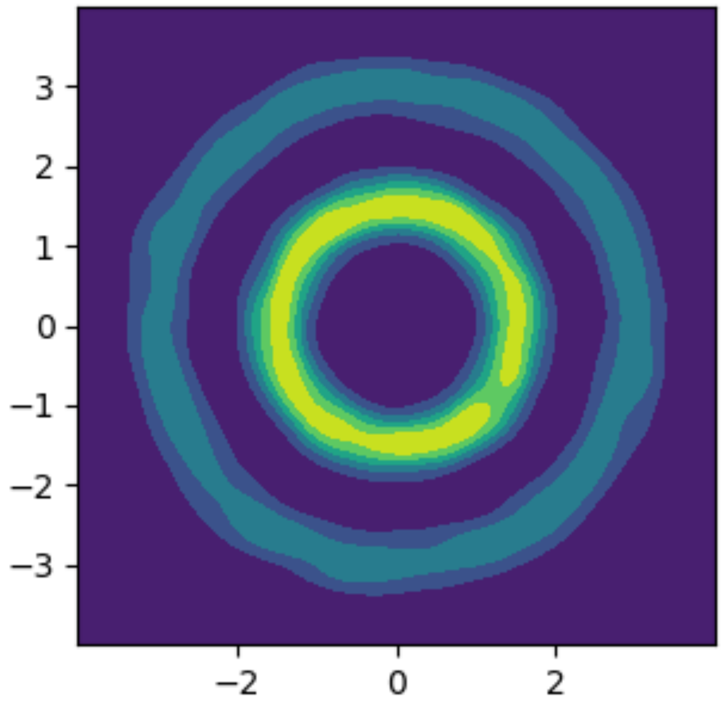
$\boldsymbol{\Gamma}_{t,\theta}(\,\cdot\,)$ is the learnt local covariance matrix informed by WPO.

$$\text{Reverse diffusive process with } \hat{\mathbf{s}}_\theta(\,\cdot\,,t) = \frac{(\nabla_y G_s * \hat{\pi}_\theta)(\,\cdot\,)}{(G_s * \hat{\pi}_\theta)(\,\cdot\,)}$$



Exact kernel formula **overfits**

memorize and resample!

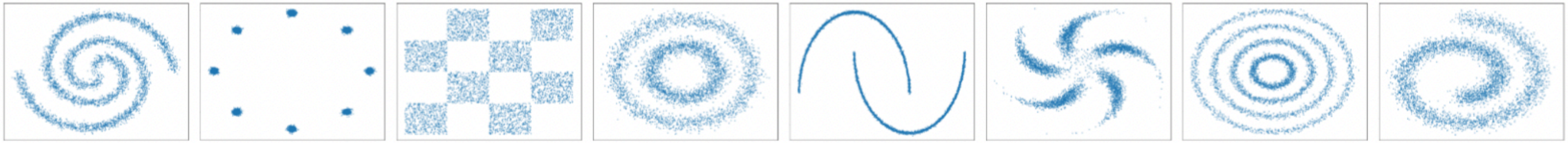


Learning local covariance matrices **generalizes**

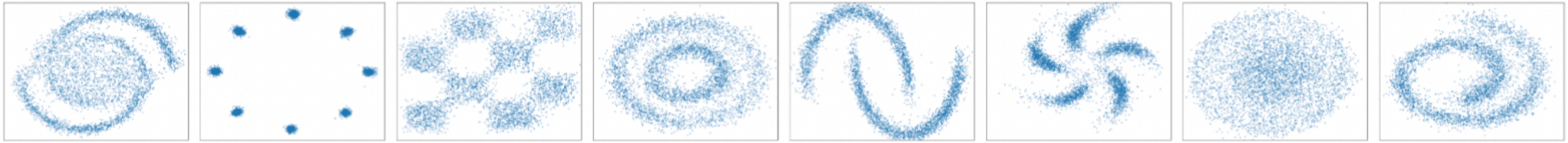
We directly learn a lower-dimensional representational space by enforcing the proper terminal condition of the HJ equation in one-step!

Illustrative examples: Deconstructing SGM

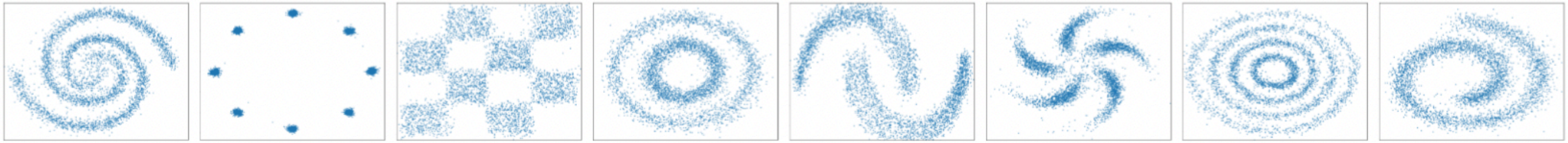
Truth



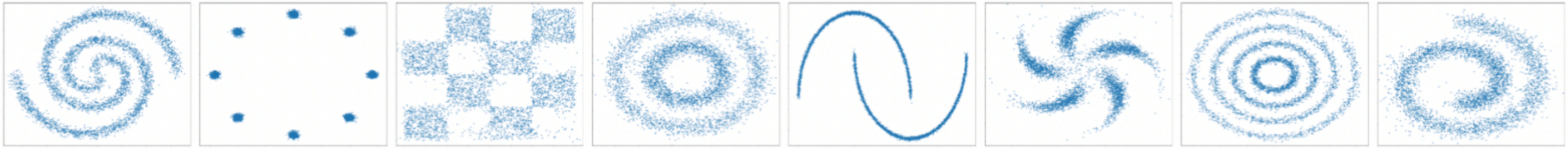
Denoising score matching with 50k epochs



Denoising score matching with 1000k epochs

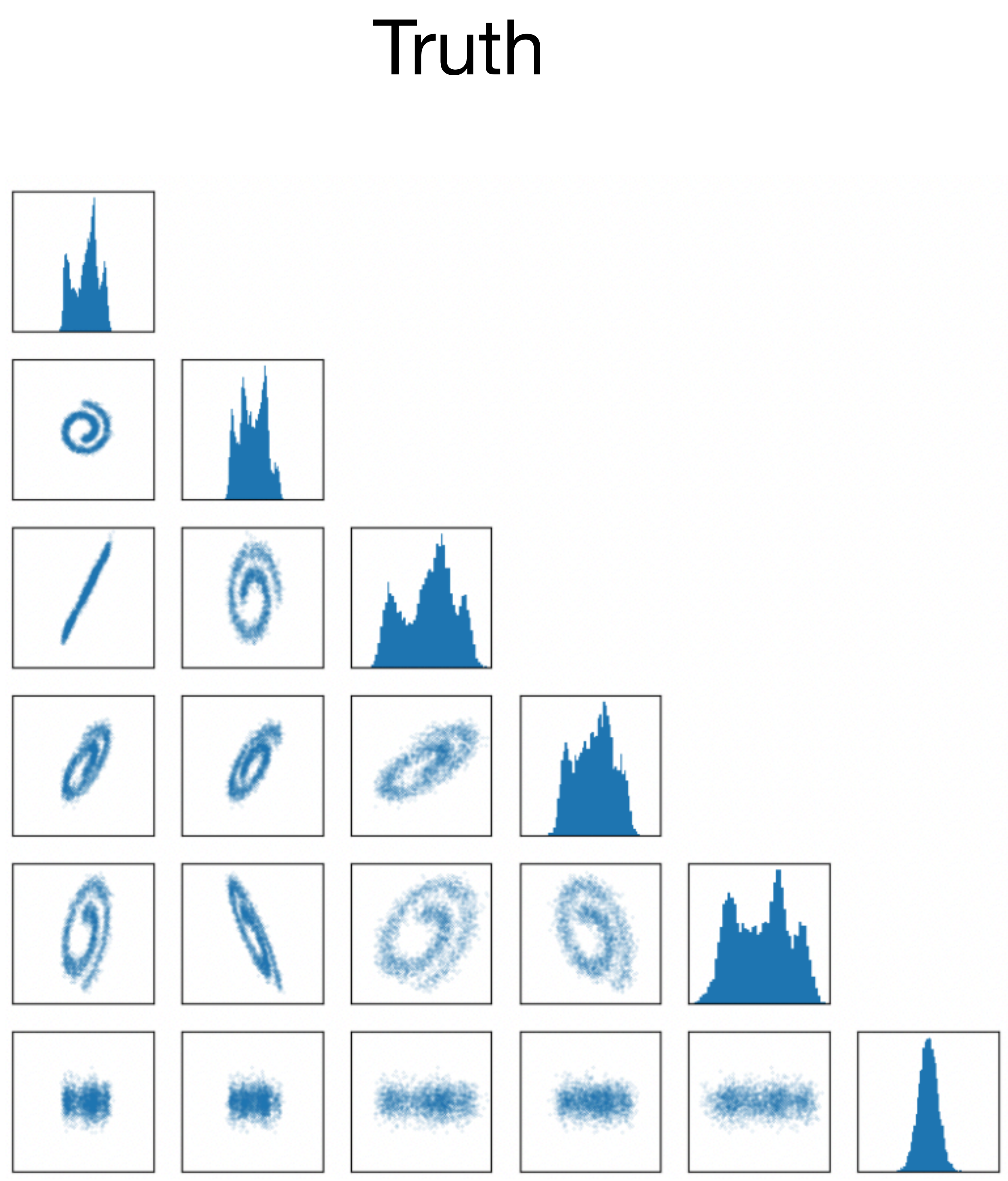


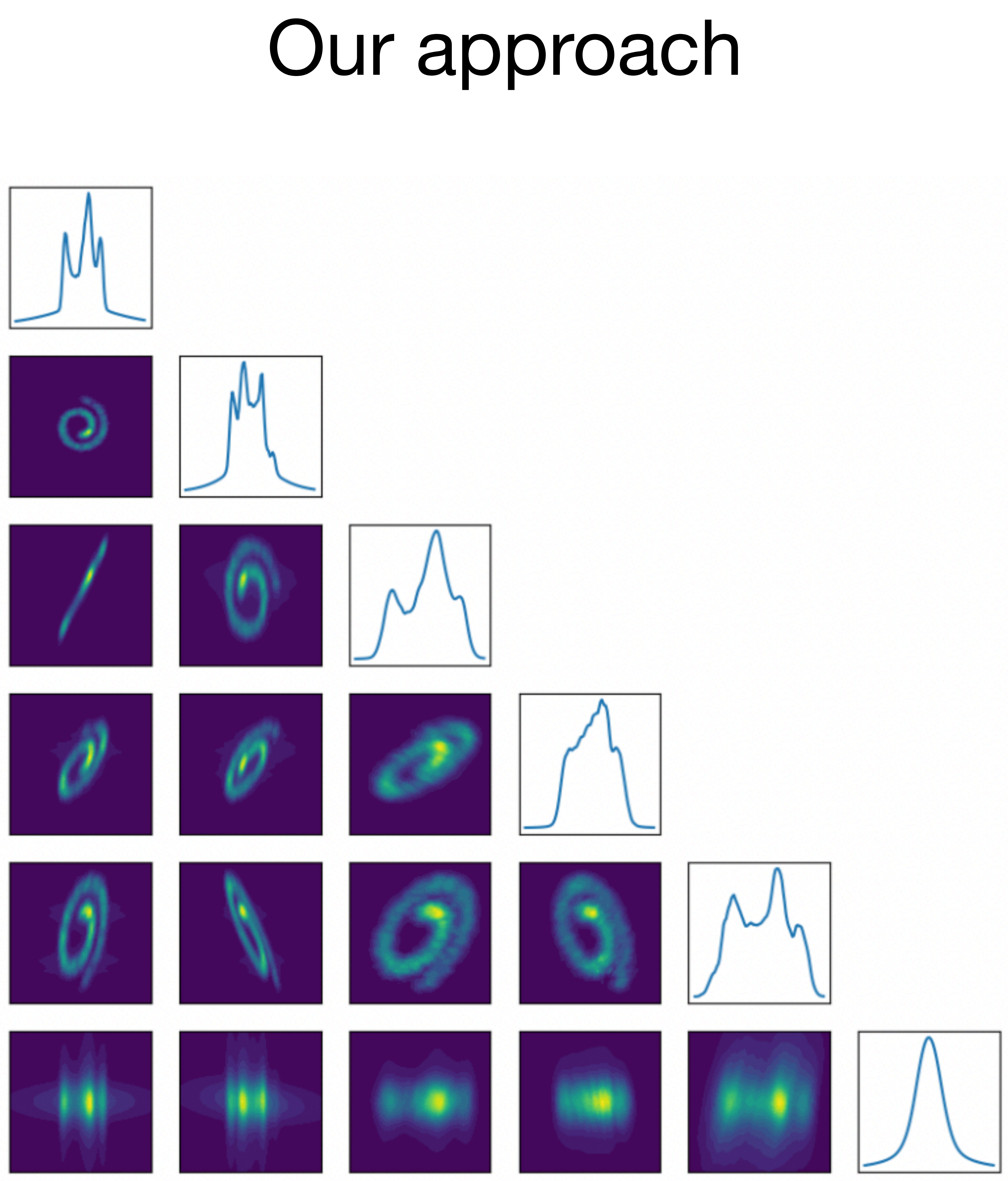
Our approach with 50k epochs



An informed mathematical structure learns score models faster

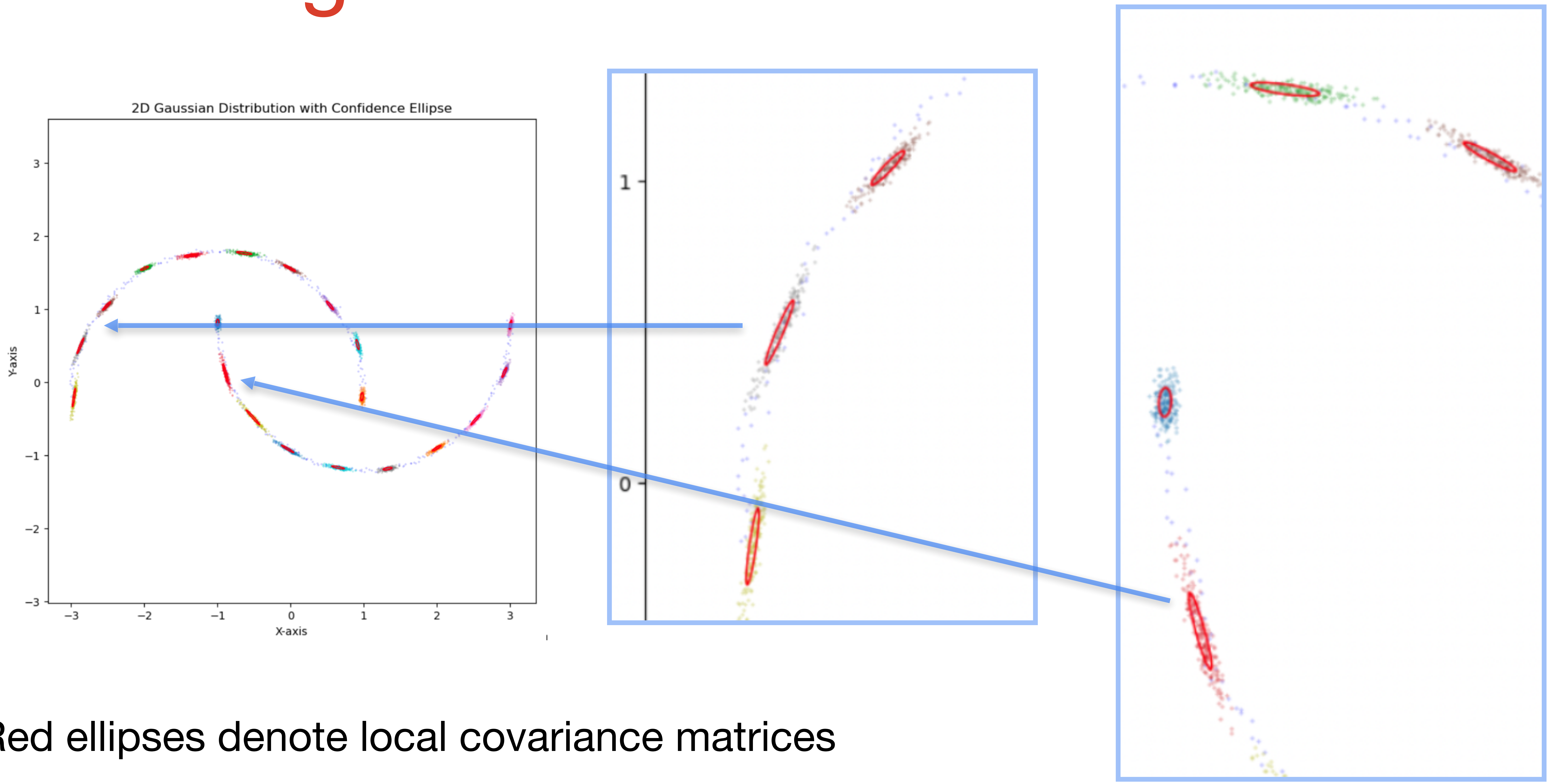
Illustrative examples: Deconstructing SGM





Six dimensional example: 3D swissroll noisily embedded in a 6D space.

Learning the data manifold



- Red ellipses denote local covariance matrices
- Set of local covariance matrices define Riemannian metric, and therefore a manifold

Takeaways

- **Faster training** with **less data** due to mathematically-informed structure of the kernel model, **resolving memorization**
    - Proper choice of kernel (solves HJB equation)
    - Manifold learning (terminal condition of HJB, proximal interpretation)

- **REQUIRES NO SIMULATION OF SDEs**
    - Kernel model can be sampled from directly

- **Formulation provides new ideas of implementations**
    - New **bespoke neural nets** for score-based models **for *scalable* implementations**
    - **Tensors** instead of neural networks in manifold learning

Thank you very much for the attention!