Are We Learning the Score Function?

Reinterpreting Diffusion Models Through Wasserstein Gradient Flow Matching

An Vuong[1], Michael T. McCann[2], Javier E. Santos[2], Yen Ting Lin[2,*]

[1] Oregon State University [2] Los Alamos National Laboratory

NEURAL INFORMATION PROCESSING SYSTEMS

Standard Interpretation of Diffusion Models

Current understanding of the mechanism of diffusion model relies on the notion of reverse-time process. Take Orstein-Uhlenbeck process (as in DDPM or as in Variance-Preserving SDE[1]) for example, the **forward process**

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}W_t \tag{1}$$

is used as the data-corruption process from $t : 0 \to \infty$ to transport $\mu_0 \to \mu_\infty$, and the **reverse-time** process from $\tau : -\infty \to 0$ to revert the process

$$\mathrm{d}X_\tau = \left[X_\tau + 2\nabla_x \log p\left(X_\tau, -\tau\right)\right]\mathrm{d}\tau + \sqrt{2}\,\mathrm{d}W_\tau, \quad X_{-\infty} \sim \mu_\infty. \tag{2}$$

is utilized for inference/sampling.

The key idea is to use the samples generated by the forward process to learn the **score-function** $S(x,t) := \nabla_x \log \rho(x,t)$, where $\rho(x,t) := (\mathrm{d}/\mathrm{d}x)\mathbb{P}\left\{X_t > x\right\}$ is the probability density of the forward process.

[1] Santos & Lin, *Using Ornstein-Uhlenbeck Procss to understand Denoising Diffusion Probabilistic Model and its Noise Schedules*, 2023

Otto Calculus and Wasserstein Gradient Flow

Jordan, Kinderlehrer, and Otto[2] and Otto[3] established the equivalence of the marginal distribution induced by the forward diffusion process (1) and that by a flow described by an ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = v_{\mathrm{WGF}}\left(x\left(t\right)\right) := -x(t) - \nabla_x \log \rho(x(t), t). \tag{WGF}$$

The equivalence can be esbalished by observing the same forward Chapman–Komogorov equations (CKE); for diffusion process, the CK equation is the Fokker–Planck Equation:

$$\partial_t \rho(\xi, t|\zeta, 0) = \nabla_\xi\left[\xi\,\rho(\xi, t|\zeta, 0)\right] + \nabla_\xi^2 \rho(\xi, t|\zeta, 0),$$

whereas for WGF, CKE is a Liouville Equation:

$$\partial_t \rho(\xi, t|\zeta, 0) = \nabla_\xi[(\xi + \nabla_\xi \log \rho\left(\xi, t|\zeta, 0\right))\,\rho(\xi, t|\zeta, 0)]\,.$$

[2] Jordan, Kinderlehrer, and Otto, *The Variational Formulation of the Fokker–Planck Equation*, 1998

[3] Otto, *The geometry of dissipative evolution equations: The porous medium equation*, 2001

Are We Really Learning the Score Function?

Because the score function $S(x,t)$ is a gradient of a scalar field $\log \rho(\cdot, t)$, it is a **conservative field**. Specifically, at any given time t, it must satisfy the **integral constraint**

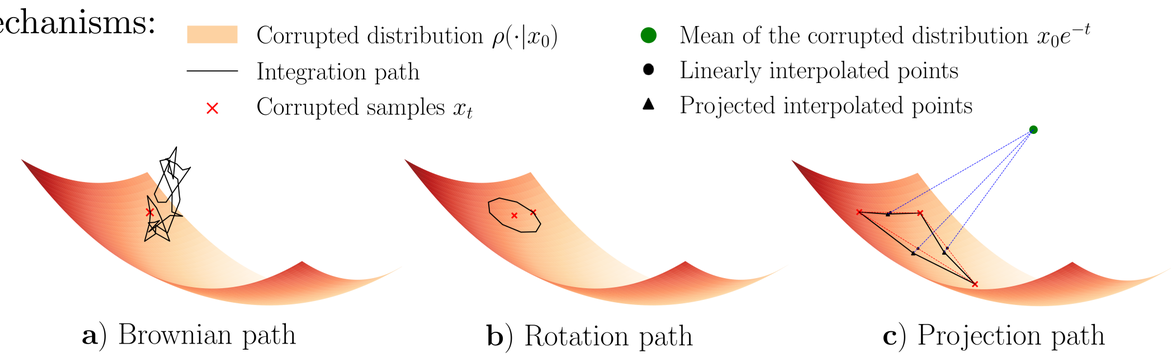
$$\oint_{\mathcal{C}} S(x,t)\,\mathrm{d}x = 0, \text{ for any closed path } \mathcal{C}, \tag{IC}$$

as well as the **differential constraint**

$$\partial_{x_j} S_i(x,t) = \partial_{x_i} S_j(x,t). \tag{DC}$$

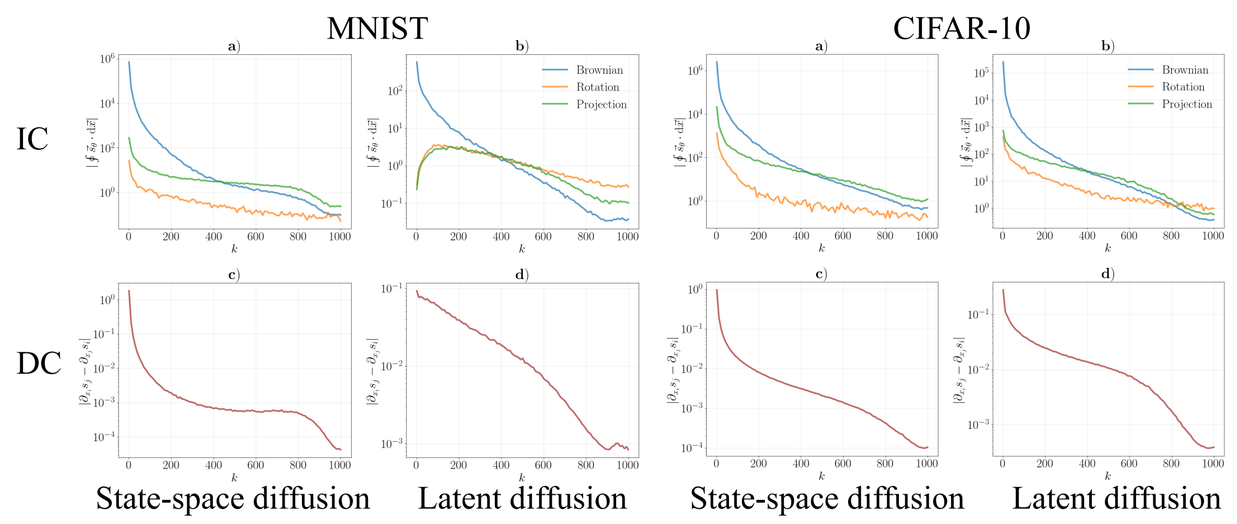
We are interested in checking if the trained diffusion models respect IC and DC.

For differential constraints, we rely on reverse-mode automatic differentiation. For integral constraints, we preform three different closed path-generation mechanisms:



a) Brownian path **b**) Rotation path **c**) Projection path

We examined the violation of integral and differential constraints of trained diffusion models, including plain and latent diffusion, on different datasets. The results indicated that both IC and DC are violated by a significant margin:



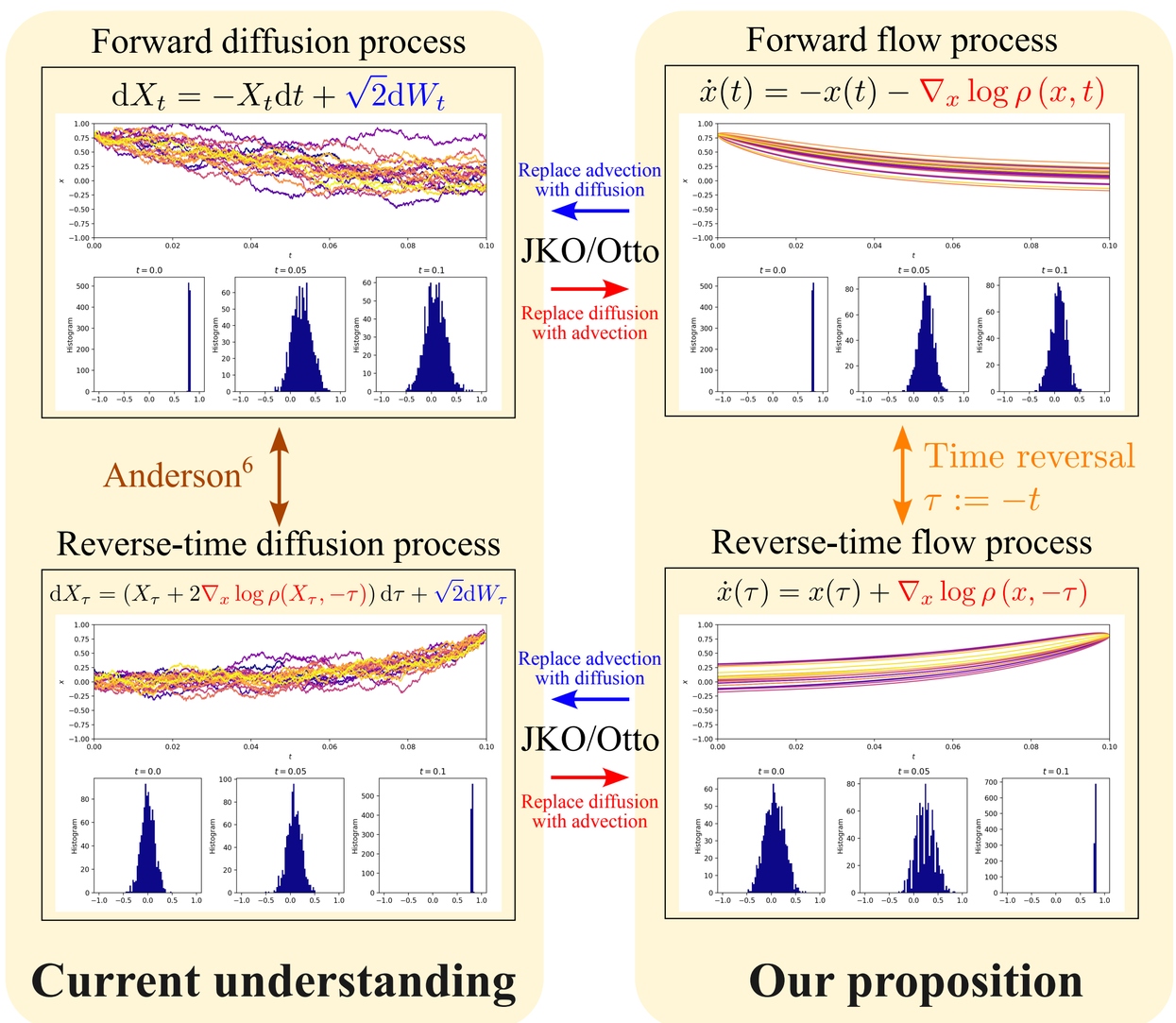
The violations raised an interesting question: Clearly, the neural network is not capable of learning the conservative score function needed for the actual reverse-time process. Why is existing computational procedure remains effective for generative modeling?

Understanding Diffusion Models as Wasserstein Gradient Flow Matching

We propose a hypothesis leveraging WGF theory to clarify what diffusion models actually learn:

> Existing diffusion modeling is better understood as modeling a normalizing flow[4], through performing flow matching[5] to the WGF velocity, rather than learning the reverse stochastic differential equation established by Anderson[6].

Our proposition can be tersely expressed as the following diagram:



Reinterpreting the existing algorithm

Training:

- Flow-matching: $\min_\theta \mathbb{E}_{t\sim\mathrm{Unif}(0,T)}\mathbb{E}_{x\sim\rho(\cdot,t)}\left\|v_\theta(\cdot,t) - S\left(\cdot,t\right)\right\|_2^2$

Sampling/Inference: $\tau := -t$, $v_{\mathrm{reverse}} := x - v_{\theta^*}$

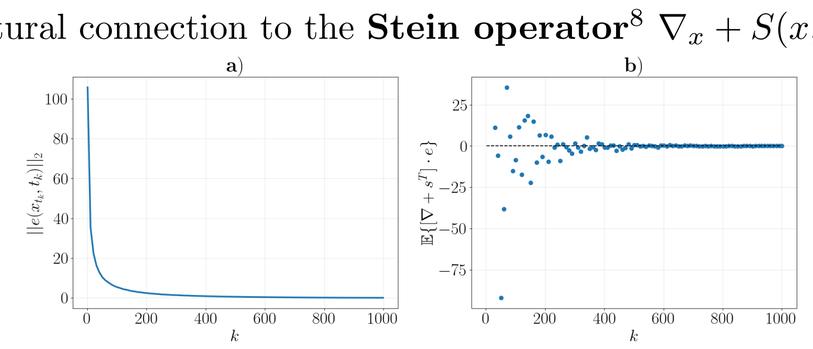
- Time-reversed Wasserstein Gradient Flow ≡ Probability Flow[7]
- WGF $\xrightarrow{\text{Reverse JKO/Otto}}$ Diffusion ≡ reversed-time SDE[6]

The advantage of the reinterpretation includes:

- Natural formulation and no need to explain the missing coefficients in DDPM (Cf. Sohl-Dickstein et al.[9]) and no need to reinvent the so-called "Probability Flow[7]";
- Involving the existence of reversed-time SDE[6] is not necessary: generative modeling only needs the **marginal distribution** (in contrast, reversed-time SDE[6] ensures the **joint distribution** and the **path measure**);
- An error/null-kernel analysis could elucidate why existing models are robust even failing to learn the conservative $S(x,t)$, showing that generative modeling can be effective if

$$0 = \left[\nabla_x + S\left(x,t\right)\right] \cdot \left[\mathrm{NN}_\theta\left(x,t\right) - S(x,t)\right].$$

A natural connection to the **Stein operator**[8] $\nabla_x + S(x,t)$ emerged.



[4] Chen et al. *Neural Ordinary Differential Equations*, 2019

[5] Lipman et al. *Flow Matching for Generative Modeling*, 2022

[6] Anderson, *Reverse-time diffusion equation models*, 1982

[7] Song et al. *Score-based generative modeling through stochastic differential equations*, 2021

[8] Liu & Wang, *Stein variational gradient descent*, 2016

[9] Sohl-Dickstein et al. *Deep Unsupervised Learning using Non-Eq. Thermodynamics*, 2015