Improving the Euclidean Diffusion Generation of Manifold Data by Mitigating Score Function Singularity

Zichen Liu[1], Wei Zhang[2], Tiejun Li[134]

[1]Center for Data Science, Peking University
[2]Zuse Institute Berlin
[3]School of Mathematical Sciences, Peking University
[4]Center for Machine Learning Research, Peking University

October 20, 2025

Motivation: Distributions on Riemannian Manifolds

- Diffusion models operate through a forward process, which gradually perturbs data into noise, and a reverse process, which reconstructs the data from noise.
- Many scientific fields involve data distributions constrained to Riemannian manifolds.
  - Sphere: Geographical Sciences (Karpatne et al., 2018; Mathieu & Nickel, 2020)
  - SO(3), SE(3): Protein Structures (Jumper et al., 2021; Watson et al., 2022); Robotic Movements (Simeonov et al., 2022)
  - SU(3): Lattice Quantum Chromodynamics (Lou et al., 2023)
  - Triangular Meshes: 3D Computer Graphics Shapes (Hoppe et al., 1992)
  - Poincaré Disk: Cell Development (Klimovskaia et al., 2020)
- **Problem Setting:** Given samples from an unknown data distribution $p_0(x)d\sigma_{\mathcal{M}}(x)$ defined on a known manifold $\mathcal{M}$, the goal is to learn a generator for this distribution. $\mathcal{M} = \{x \in \mathbb{R}^n \mid \xi(x) = 0_{n-d}\}$ is a d-dimensional submanifold of $\mathbb{R}^n$.
- Several prior studies have highlighted the divergence of score functions in diffusion models under the manifold hypothesis.

The Origin of the Scale Discrepancy

- We consider the Variance Exploding SDE (VESDE; Song et al., 2021):

$$\mathrm{d}X_t = \sqrt{\frac{\mathrm{d}\sigma_t^2}{\mathrm{d}t}}\mathrm{d}W_t,$$

  where σ_t is a predefined noise scale. $p_{\sigma_t}(x_t|x)$ follows a Gaussian distribution $\mathcal{N}(x_t|x, \sigma_t^2 I)$.

- The perturbed data deviate from their original confinement to the d-dimensional submanifold $\mathcal{M}$.

$$p_\sigma(\tilde{x}) = \int_{\mathcal{M}} p_0(x) p_\sigma(\tilde{x}|x) \mathrm{d}\sigma_{\mathcal{M}}(x) = (2\pi\sigma^2)^{-\frac{n}{2}} \int_{\mathcal{M}} p_0(x) e^{-\frac{|x-\tilde{x}|^2}{2\sigma^2}} \mathrm{d}\sigma_{\mathcal{M}}(x)$$

- $p_0(x)$ is defined only on $\mathcal{M}$, while the perturbed density $p_\sigma(\tilde{x})$ is defined on $\mathbb{R}^n$.
- As $\sigma \to 0$, the perturbed distribution becomes tightly concentrated around its mean, resulting in a steep gradient landscape for $-\log p_\sigma(\tilde{x})$.

Scale Discrepancy of the Score Function

Theorem (Scale discrepancy under the isotropic noise)

Under mild assumptions, the following two asymptotic expansions for $p_\sigma(\tilde{x})$ hold:

1. *For $\tilde{x} \notin \mathcal{M}$, assume that x^* is the unique minimizer of $\min_{x\in\mathcal{M}} \|x - \tilde{x}\|$. As $\sigma \to 0$, we have*

$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = \frac{x^* - \tilde{x}}{\sigma^2} + O(1)\,.$$

2. *For $\tilde{x} \in \mathcal{M}$, as $\sigma \to 0$, we have*

$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = \nabla_{\tilde{x}}^{\mathcal{M}} \log p_0(\tilde{x}) - \frac{1}{2}\sum_{j,j'=1}^{n} \frac{\partial P_{\cdot j}}{\partial x_{j'}}(\tilde{x}) P_{jj'}(\tilde{x}) + O(\sigma),$$
$$\nabla_{\tilde{x}}^{\mathcal{M}} \log p_\sigma(\tilde{x}) = \nabla_{\tilde{x}}^{\mathcal{M}} \log p_0(\tilde{x}) + O(\sigma),$$

where $P(x)$ denotes the projection matrix and $\frac{\partial P_{\cdot j}}{\partial x_{j'}}$ denotes the vector whose ith component is $\frac{\partial P_{ij}}{\partial x_{j'}}$.

- Part (1) shows that the score function explodes entirely due to its normal direction, represented by $x^* - \tilde{x}$. The remaining component is of order $O(1)$.
- Part (2) establishes a connection between the Riemannian score function and the perturbed score function in the ambient space, for points on the manifold.

Scale Discrepancy of the Loss Function

- We extend $P(x)$ to the entire space $\mathbb{R}^n$ and denote $P^\perp(\tilde{x}) = I - P(\tilde{x})$.
- For $\tilde{x} \in \mathbb{R}^n$, define the tangential and normal components of the score function as $P(\tilde{x})\nabla_{\tilde{x}} \log p_\sigma(\tilde{x})$ and $P^\perp(\tilde{x})\nabla_{\tilde{x}} \log p_\sigma(\tilde{x})$, respectively.
- The quadratic loss ℓ_{quad} can be decomposed into the tangential and normal parts:

$$
\begin{aligned}
\ell_{\text{quad}} =& \mathbb{E}_{\tilde{x}\sim p_\sigma(\tilde{x})} \left\| s_\theta(\tilde{x}, t) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) \right\|^2 \\
=& \mathbb{E}_{\tilde{x}\sim p_\sigma(\tilde{x})} \left\| P(\tilde{x}) s_\theta(\tilde{x}, t) - P(\tilde{x})\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) \right\|^2 \\
& + \mathbb{E}_{\tilde{x}\sim p_\sigma(\tilde{x})} \left\| P^\perp(\tilde{x}) s_\theta(\tilde{x}, t) - P^\perp(\tilde{x})\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) \right\|^2 \\
=: & \ \ell^{\|}_{\text{quad}} + \ell^{\perp}_{\text{quad}}.
\end{aligned}
$$

- $\ell^{\|}_{\text{quad}}$ and $\ell^{\perp}_{\text{quad}}$ have scales of $O(1)$ and $O(1/\sigma)$, respectively:

$$
\mathbb{E}_{\tilde{x}|x} P(\tilde{x})\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = \mathbb{E}_{\tilde{x}|x} \left(P(x^*) + O(\tilde{x} - x^*)\right) \left(\frac{x^* - \tilde{x}}{\sigma^2} + O(1)\right) = O(1),
$$

$$
\mathbb{E}_{\tilde{x}|x} P^\perp(\tilde{x})\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = \mathbb{E}_{\tilde{x}|x} \left(P^\perp(x^*) + O(\tilde{x} - x^*)\right) \left(\frac{x^* - \tilde{x}}{\sigma^2} + O(1)\right) = O\left(\frac{1}{\sigma}\right).
$$

Training under Scale Discrepancy

This multiscale singularity of the loss formulation poses challenges during training.

- Training tends first to fit larger-scale features aligned with the normal component, which mainly pulls samples back onto the manifold.
- The model underfits the tangential component, so it fails to capture finer, on-manifold details of the data distribution, reducing the accuracy of the generated distribution.

We propose the following two methods:

- **Niso-DM**: Perturb data with non-isotropic noise by introducing additional noise along the normal direction during the forward diffusion process.
- **Tango-DM**: Train only the tangential component of the score function using a tangential-only loss function.

Niso-DM: Perturb Data with Non-isotropic Noise

- To mitigate the scale discrepancy, we replace the isotropic noise in the forward process with non-isotropic noise.
- The perturbed data is generated as $\tilde{x}_t = x + \sigma_t \epsilon_1 + \sigma_t^{\alpha_t} N(x) \epsilon_2$, where $x \sim p_0(x)$, $\epsilon_1 \sim \mathcal{N}(0, I_n)$, $\epsilon_2 \sim \mathcal{N}(0, I_{n-d})$, $\alpha_t \in (0,1)$, and $N(x) \in \mathbb{R}^{n \times (n-d)}$ is given by $N(x) = \nabla \xi(x) \left(\nabla \xi(x)^T \nabla \xi(x) \right)^{-\frac{1}{2}}$.
- The conditional probability density is given by $p_{\sigma_t}(\tilde{x}|x) = \mathcal{N}(x, \Sigma_{\sigma_t}(x))$, where $\Sigma_{\sigma_t}(x) = \sigma_t^2 I + \sigma_t^{2\alpha_t} N(x) N(x)^T$.
- Noting that $\nabla_{x_t} \log p_{\sigma_t}(x_t|x) = -\Sigma_{\sigma_t}(x)^{-1}(x_t - x)$, the denoising score matching loss becomes:

$$\ell_{\text{Niso}}(t, \theta) = \mathbb{E}_{x, x_t} \left\| s_\theta(x_t, t) + \Sigma_{\sigma_t}(x)^{-1}(x_t - x) \right\|^2,$$

where $\Sigma_{\sigma_t}(x)^{-1}$ has a closed-form expression.

Niso-DM: Perturb Data with Non-isotropic Noise

By introducing additional noise along the normal direction, the scale of the normal component is reduced from $O(1/\sigma^2)$ to $O(1/\sigma^{2\alpha})$.

Theorem (Scale discrepancy under the non-isotropic noise)

Let $p_\sigma(\tilde{x})$ denote the distribution under non-isotropic perturbation, defined as:

$$p_\sigma(\tilde{x}) = (2\pi)^{-\frac{n}{2}} \int_{\mathcal{M}} p_0(x)(\det \Sigma_\sigma(x))^{-\frac{1}{2}} e^{-\frac{1}{2}(\tilde{x}-x)^T \Sigma_\sigma(x)^{-1}(\tilde{x}-x)} \mathrm{d}\sigma_{\mathcal{M}}(x),$$

where $\Sigma_\sigma(x) = \sigma^2 I + \sigma^{2\alpha} N(x)N(x)^T$ and $\alpha \in (0,1)$.

1. *For $\tilde{x} \notin \mathcal{M}$, assuming that $x^* \in \mathcal{M}$ is the unique minimizer of $\min_{x\in\mathcal{M}} \|x - \tilde{x}\|$, we have*
$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = \frac{x^* - \tilde{x}}{\sigma^{2\alpha}} \cdot \frac{1}{1+\sigma^{2-2\alpha}} + O(\sigma^{(1-2\alpha)\wedge 0}).$$
2. *For $\tilde{x} \in \mathcal{M}$, as $\sigma \to 0$, we have*
$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = \nabla_{\tilde{x}}^{\mathcal{M}} \log p_0(\tilde{x}) + O(\sigma^{(2-2\alpha)\wedge 1}).$$

Tango-DM: Learn Only the Tangential Component

- Recall that the loss function can be decomposed into two parts, $\ell^{\parallel}_{\text{quad}}$ and $\ell^{\perp}_{\text{quad}}$, and the singularity issue comes from $\ell^{\perp}_{\text{quad}}$.
- We propose training only the tangential component of the score function using the loss $\ell^{\parallel}_{\text{quad}}$ when the noise scale σ_t is sufficiently small, thereby avoiding the singularity associated with $\ell^{\perp}_{\text{quad}}$.
- We introduce the following Tango loss

$$\ell_{\text{tango}}(t, \theta) := \mathbb{E}_{x, x_t} \| s^{\parallel}_{\theta}(x_t, t) - P(x_t) \nabla_{x_t} \log p_{\sigma_t}(x_t | x) \|^2,$$

where we define $s^{\parallel}_{\theta}(x, t) := P(x) s_{\theta}(x, t)$ for $x \in \mathbb{R}^n$.

- The optimal score network $s^{\parallel}_{\theta^*}(x, t)$ satisfies $s^{\parallel}_{\theta^*}(x, t) = P(x) \nabla_x \log p_{\sigma_t}(x)$. This result ensures the validity of the Tango loss.
- When σ_t is not small (the singularity issue is less severe), we use the original denoising score matching loss to train the entire score function $s_{\theta}(x, t)$.

Experiments

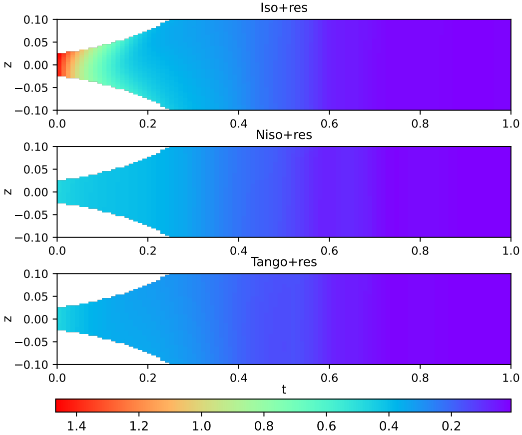
- **Rescaling technique** ($+res$): We consider using neural networks to approximate the normalized score functions, i.e. $s_\theta(x,t) = \hat{s}_\theta(x,t)/w_t$, where $\hat{s}_\theta(x,t)$ denotes a neural network and w_t is the scaling factor of the optimal score function.
    - It is designed to improve numerical stability, as neural networks can more effectively approximate an $O(1)$ term compared to one that grows explosively.
    - Notably, this technique is equivalent to the ϵ-parameterization introduced in prior studies.

- We perform experiments with the vanilla diffusion models, as well as our proposed Niso-DM and Tango-DM, denoted as *Iso*, *Niso*, and *Tango*, respectively.

- New samples are generated via two methods: Reverse SDE and Annealing SDE on manifolds.

Hyperplane in 3D Space

- The manifolds: $\mathcal{M} = \{(x, y, z) \in \mathbb{R}^3 | z = 0\}$.
- The target distribution is a mixture of Gaussian distributions with nine modes located on the plane.



- The average error of the tangential component of the learned score function in the $x - y$ plane, along the z-axis and t-axis. From top to bottom, the plots correspond to the vanilla algorithm (Iso-DM), our proposed Niso-DM and Tango-DM.

High-dimensional Special Orthogonal Group

- The manifold $\mathrm{SO}(10)$, a 45-dimensional submanifold embedded in $\mathbb{R}^{100}$.
- Target distribution: a multimodal distribution on $\mathrm{SO}(10)$, consisting of 5 modes.

Table: Results for $\mathrm{SO}(10)$: Sliced 1-Wasserstein distance under different training methods.

	Reversal	Annealing
Iso	1.76e-2 ±1.15e-2	1.86e-2 ±9.65e-3
Niso	5.58e-3 ±1.84e-3	1.16e-2 ±2.51e-3
Tango	-	1.89e-2 ±2.05e-3
Iso+res	9.49e-3 ±1.91e-3	1.17e-2 ±7.29e-4
Niso+res	**4.60e-3** ±8.24e-4	**6.00e-3** ±7.53e-4
Tango+res	-	6.42e-3 ±1.97e-3

Thank You!