A Geometric Perspective on Diffusion Models

Yuanzhi Zhu

18/Jun/2023

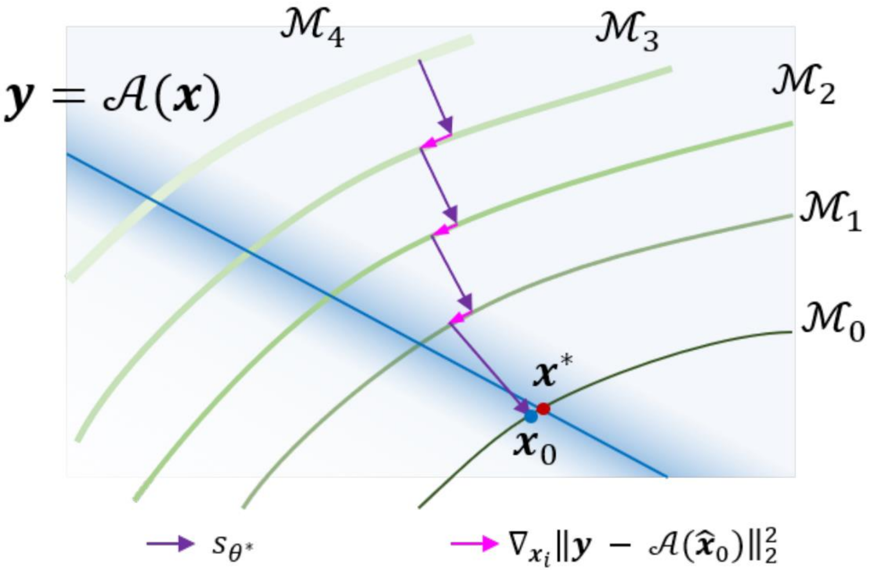
Content

- Motivation
- Geometric Perspective
- Conclusion

A Geometric Perspective on Diffusion Models*

(my) Motivation: Not related to non-Euclidean geometry

1. How to understand this graph?

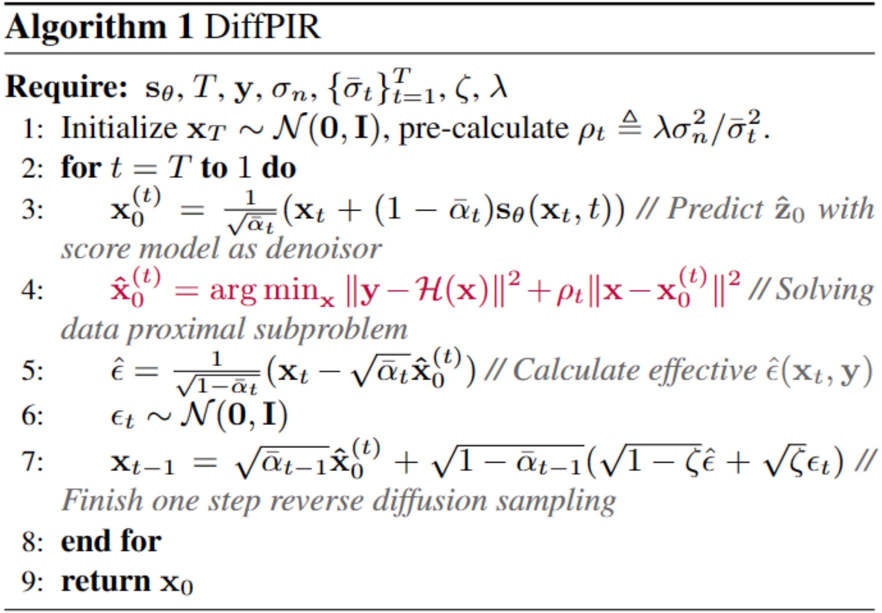


*[2305.19947] A Geometric Perspective on Diffusion Models (arxiv.org)

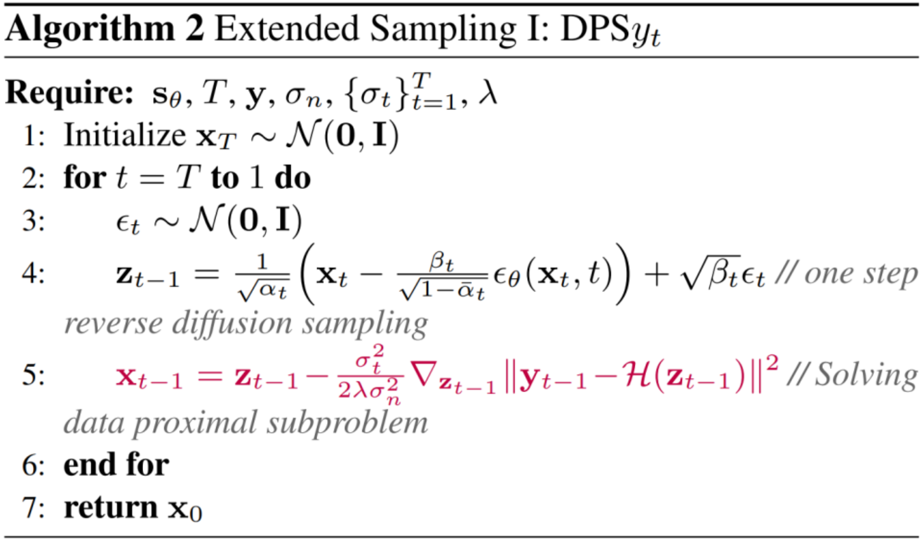
A Geometric Perspective on Diffusion Models

(my) Motivation: Not related to MOLECULAR generation

2. What's the difference between the following two algorithm?



```
Algorithm 1 DiffPIR
Require: $\mathbf{s}_\theta, T, \mathbf{y}, \sigma_n, \{\bar{\sigma}_t\}_{t=1}^{T}, \zeta, \lambda$
Initialize $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, pre-calculate $\rho_t \triangleq \lambda\sigma_n^2/\bar{\sigma}_t^2$.
for $t = T$ to $1$ do
    $\mathbf{x}_0^{(t)} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t + (1-\bar{\alpha}_t)\mathbf{s}_\theta(\mathbf{x}_t, t))$ // Predict $\hat{\mathbf{z}}_0$ with score model as denoisor
    $\hat{\mathbf{x}}_0^{(t)} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathcal{H}(\mathbf{x})\|^2 + \rho_t\|\mathbf{x} - \mathbf{x}_0^{(t)}\|^2$ // Solving data proximal subproblem
    $\hat{\epsilon} = \frac{1}{\sqrt{1-\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0^{(t)})$ // Calculate effective $\hat{\epsilon}(\mathbf{x}_t, \mathbf{y})$
    $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0^{(t)} + \sqrt{1-\bar{\alpha}_{t-1}}(\sqrt{1-\zeta}\hat{\epsilon} + \sqrt{\zeta}\epsilon_t)$ // Finish one step reverse diffusion sampling
end for
return $\mathbf{x}_0$
```

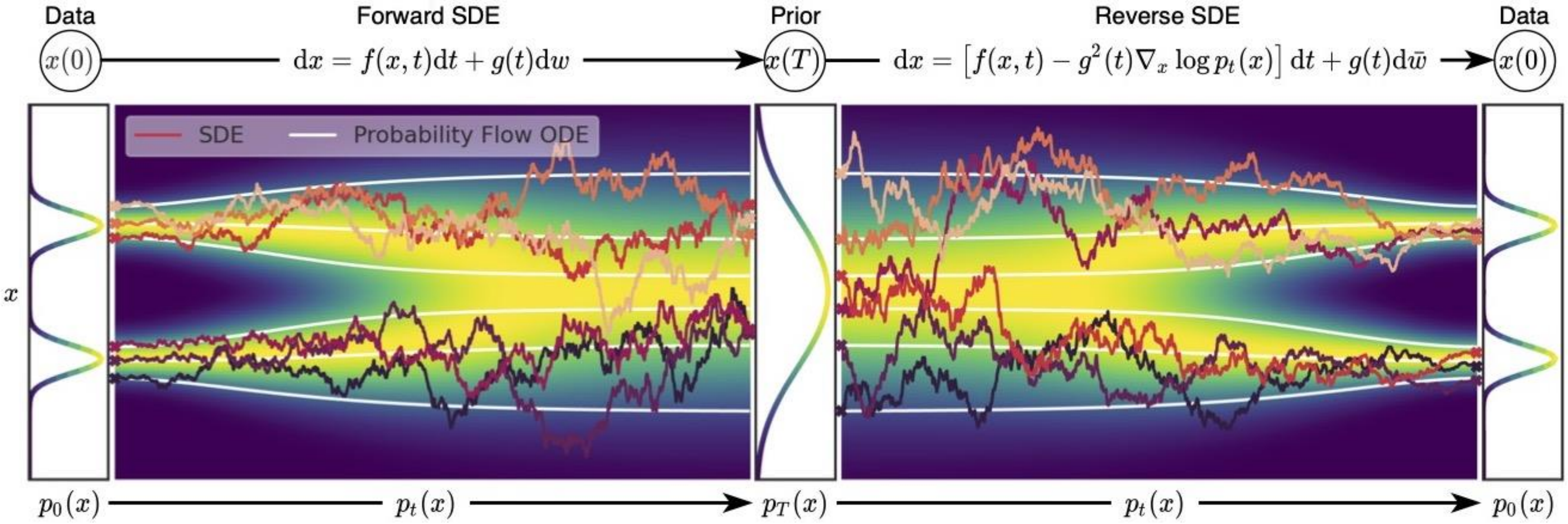


```
Algorithm 2 Extended Sampling I: DPS$y_t$
Require: $\mathbf{s}_\theta, T, \mathbf{y}, \sigma_n, \{\sigma_t\}_{t=1}^{T}, \lambda$
Initialize $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T$ to $1$ do
    $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sqrt{\beta_t}\epsilon_t$ // one step reverse diffusion sampling
    $\mathbf{x}_{t-1} = \mathbf{z}_{t-1} - \frac{\sigma_t^2}{2\lambda\sigma_n^2}\nabla_{\mathbf{z}_{t-1}}\|\mathbf{y}_{t-1} - \mathcal{H}(\mathbf{z}_{t-1})\|^2$ // Solving data proximal subproblem
end for
return $\mathbf{x}_0$
```

A Geometric Perspective on Diffusion Models

(my) Motivation: Use only VE-ODE for example

3. How to understand the diffusion trajectory better?

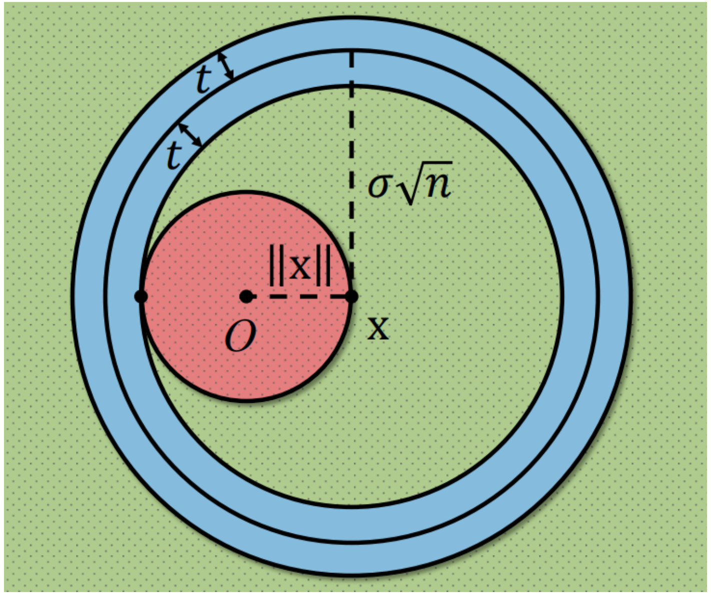

Data
Forward SDE
Prior
Reverse SDE
Data
$x(0)$
$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w$
$x(T)$
$\mathrm{d}x = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{w}$
$x(0)$
SDE
Probability Flow ODE
x
$p_0(x)$
$p_t(x)$
$p_T(x)$
$p_t(x)$
$p_0(x)$

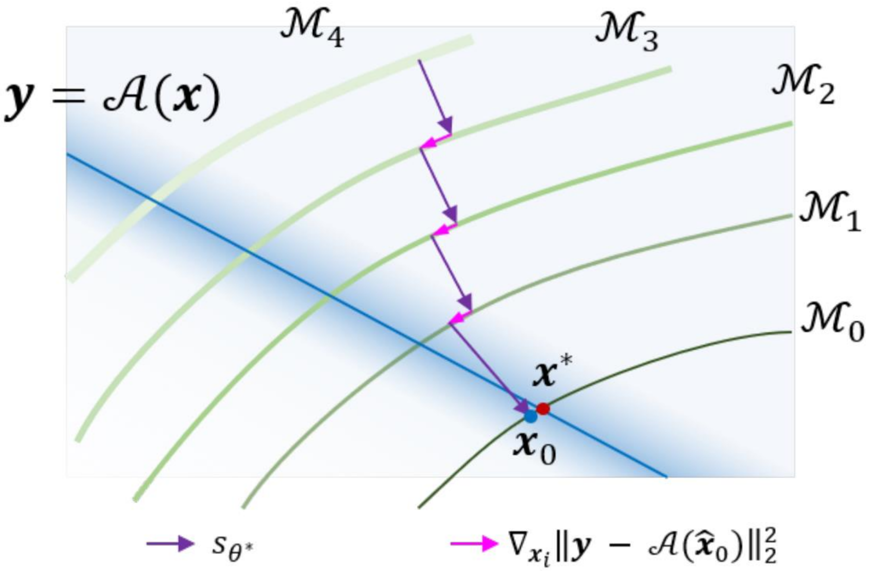
Content

- Motivation
- Geometric Perspective
- Conclusion

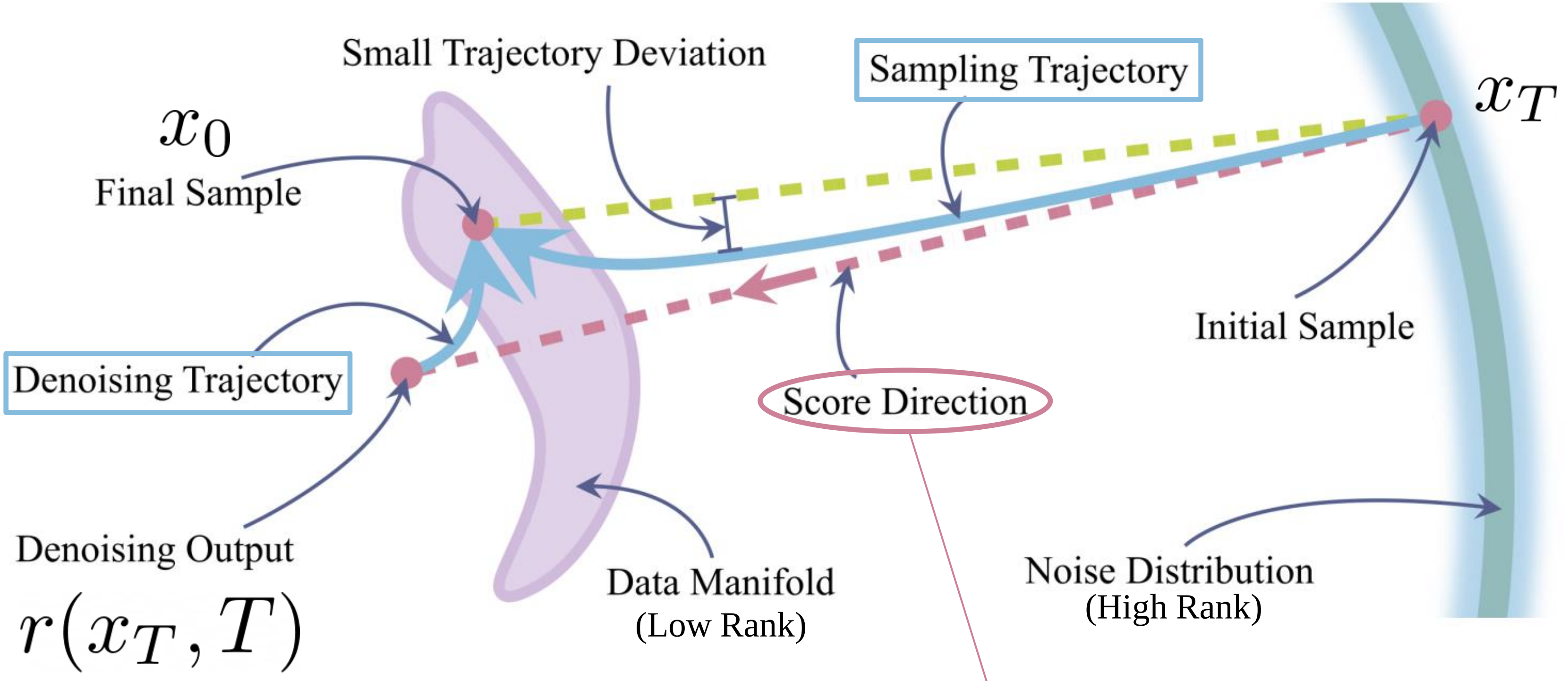
Visualization of High Dimensional Trajectory

Proposition 1. *Given a high-dimensional vector* $\mathbf{x} \in \mathbb{R}^d$ *and an isotropic Gaussian noise* $\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}; \sigma^2 \mathbf{I}_d\right)$, $\sigma > 0$, *we have* $\mathbb{E}\left\|\mathbf{z}\right\|^2 = \sigma^2 d$, *and with high probability,* $\mathbf{z}$ *stays within a "thin shell":* $\|\mathbf{z}\| = \sigma\sqrt{d} \pm O(1)$. *Additionally,* $\underbrace{\mathbb{E}\left[\|\mathbf{x}+\mathbf{z}\|^2 - \|\mathbf{x}\|^2\right] = \sigma^2 d}_{\text{Perpendicular}}$, $\lim_{d\to\infty} \mathbb{P}\left(\|\mathbf{x}+\mathbf{z}\| > \|\mathbf{x}\|\right) = 1$.





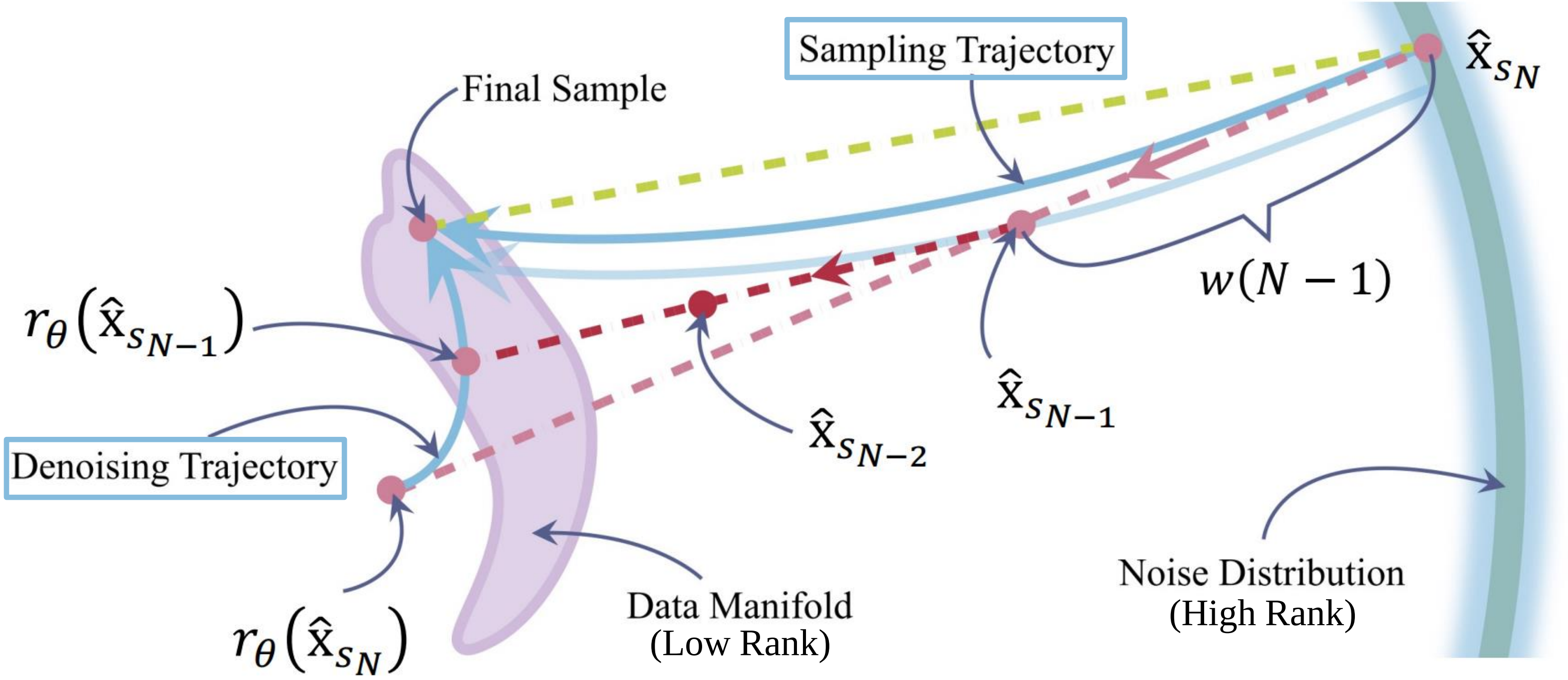
Visualization of High Dimensional Trajectory



$$\frac{dx}{dt} = -\dot{\sigma}(t)\sigma(t)\nabla \log p(x, \sigma(t)) = -\dot{\sigma}(t)\frac{r(x,\sigma(t)) - x}{\sigma(t)} \xrightarrow{\sigma(t)=t} -\frac{r(x,t) - x}{t}$$

Visualization of High Dimensional Trajectory

1. Straightness of the trajectories
2. Properties of denoising trajectory



Experiments on High Dimensional Trajectory

Notations:

sampling trajectory sequence	$\{\hat{\mathbf{x}}_s\}_{s_N}^{s_0}$	(reverse diffusion with trained model)
optimal sampling sequence	$\{\hat{\mathbf{x}}_s^\star\}_{s_N}^{s_0}$	(forward diffusion of image from dataset)
ℓ_2 distance	$d(\cdot,\cdot)$	
trajectory deviation	$d(\hat{\mathbf{x}}_s, [\hat{\mathbf{x}}_{s_0}, \hat{\mathbf{x}}_{s_N}])$	(straightness)
denoising trajectory sequence	$\{r_\theta(\hat{\mathbf{x}}_s, s)\}_{s_N}^{s_1}$	

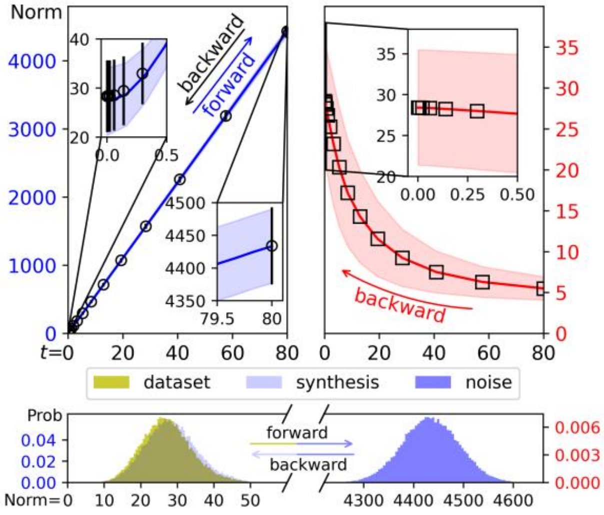
optimal denoiser

$$r_{\boldsymbol{\theta}}^\star(\hat{\mathbf{x}};\sigma_t) = \sum_i u_i \mathbf{x}_i = \sum_i \frac{\exp\left(-\|\hat{\mathbf{x}} - \mathbf{x}_i\|^2 / 2\sigma_t^2\right)}{\sum_j \exp\left(-\|\hat{\mathbf{x}} - \mathbf{x}_j\|^2 / 2\sigma_t^2\right)} \mathbf{x}_i, \quad \sum_i u_i = 1.$$

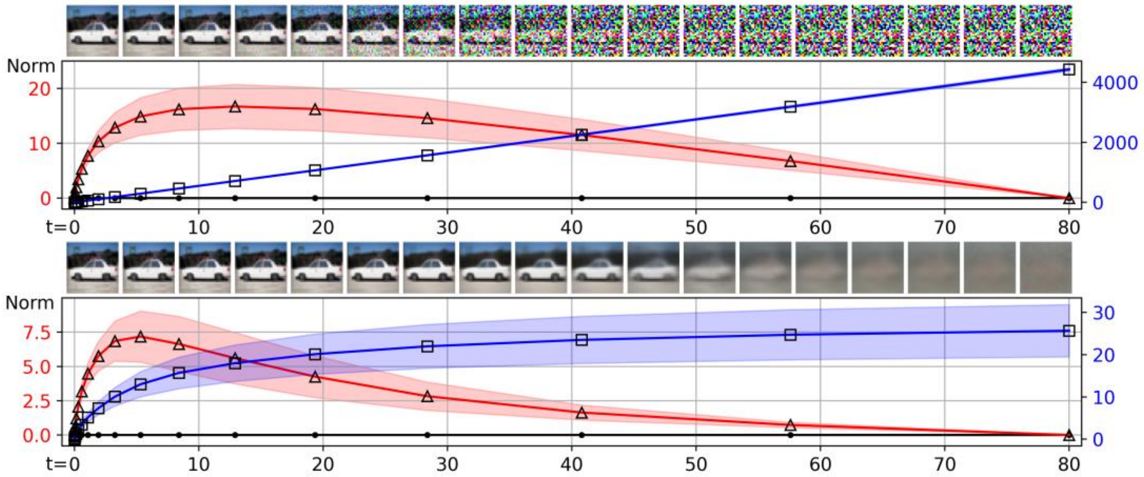
Experiments on High Dimensional Trajectory

Observation 1. *The sampling trajectory is almost straight while the denoising trajectory is bent.*

Observation 2. *The generated samples on the sampling trajectory and denoising trajectory both move monotonically from the initial points toward their converged points in expectation, i.e.,* $\{\mathbb{E}\left[d(\hat{\mathbf{x}}_s, \hat{\mathbf{x}}_{s_0})\right]\}_{s_N}^{s_0}$ *and* $\{\mathbb{E}\left[d\left(r_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_s), r_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_{s_1})\right)\right]\}_{s_N}^{s_1}$ *are monotone decreasing sequences.*



(a) The statistics of magnitude.



(b) Deviation in the sampling (top)/denoising (bottom) trajectories.

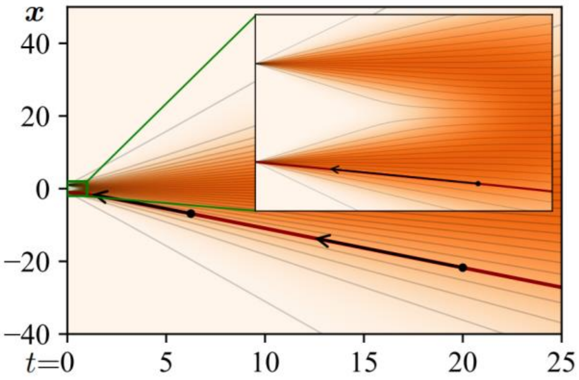
Experiments on High Dimensional Trajectory

Observation 3. *The sampling trajectory converges to the data distribution in a monotone magnitude shrinking way. Conversely, the denoising trajectory converges to the data distribution in a monotone magnitude expanding way. Formally, we have* $\{\mathbb{E}\|\hat{\mathbf{x}}_s\|\}_{s_N}^{s_0} \downarrow$ *and* $\boxed{\{\mathbb{E}\|r_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_s)\|\}_{s_N}^{s_1} \uparrow.}$

optimal denoiser: $r_{\boldsymbol{\theta}}^{\star}\left(\hat{\mathbf{x}};\sigma_t\right) = \sum_i u_i \mathbf{x}_i = \sum_i \frac{\exp\left(-\|\hat{\mathbf{x}}-\mathbf{x}_i\|^2/2\sigma_t^2\right)}{\sum_j \exp\left(-\|\hat{\mathbf{x}}-\mathbf{x}_j\|^2/2\sigma_t^2\right)}\mathbf{x}_i, \quad \sum_i u_i = 1.$

$$r_{\theta}^{\star}\left(\hat{\mathbf{x}};\sigma_t\right) \xrightarrow{\sigma_t \to \infty} \mathbb{E}[\mathbf{x}_i]$$

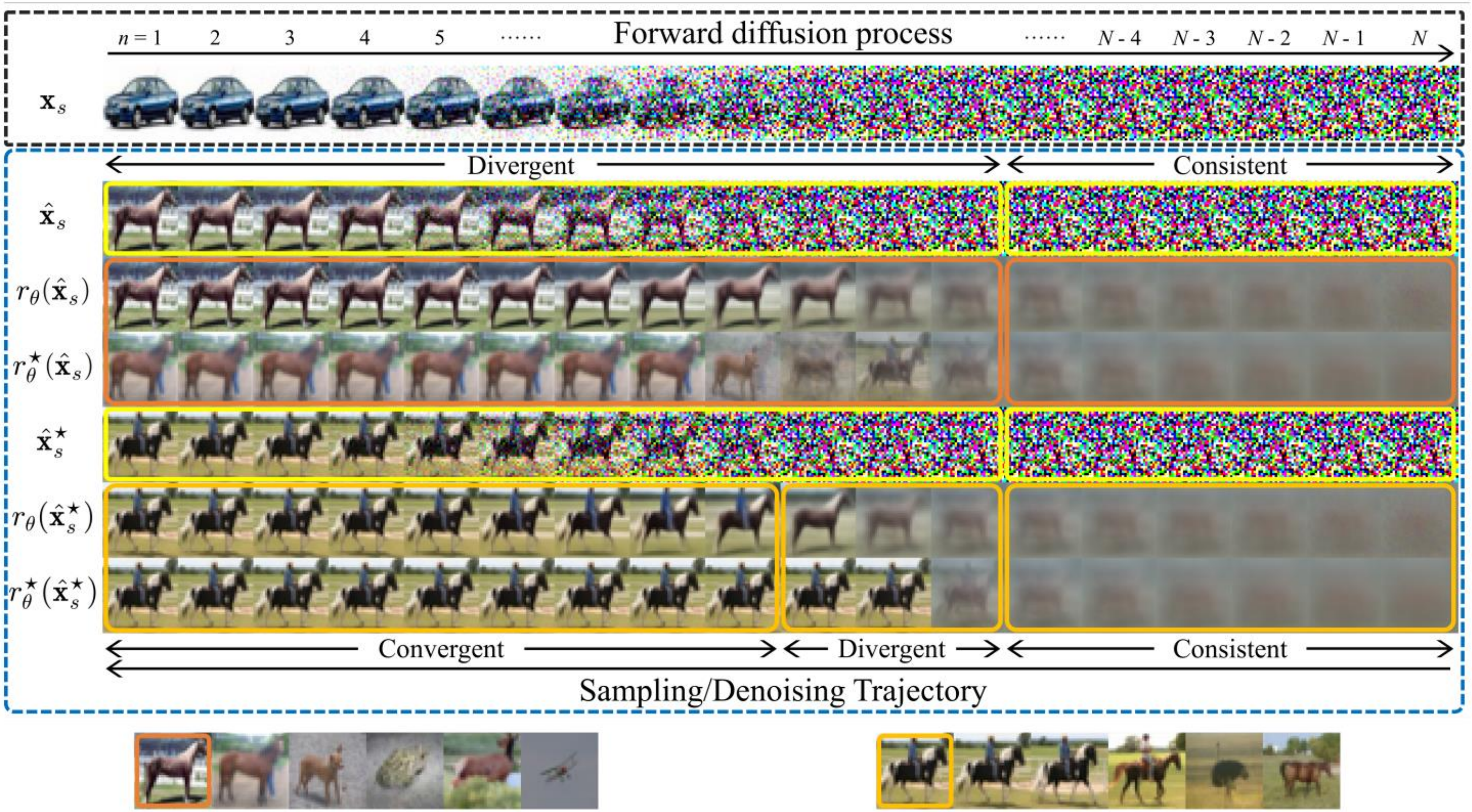
1D example with two source data point (-1) and (+1) (from EDM)



(c) DDIM [47] / Our ODE

$$\frac{dx}{dt} = -\frac{r(x,t)-x}{t}$$

Experiments on High Dimensional Trajectory



n = 1
2
3
4
5
......
Forward diffusion process
......
N - 4
N - 3
N - 2
N - 1
N
$\mathbf{x}_s$
Divergent
Consistent
$\hat{\mathbf{x}}_s$
$r_\theta(\hat{\mathbf{x}}_s)$
$r_\theta^\star(\hat{\mathbf{x}}_s)$
$\hat{\mathbf{x}}_s^\star$
$r_\theta(\hat{\mathbf{x}}_s^\star)$
$r_\theta^\star(\hat{\mathbf{x}}_s^\star)$
Convergent
Divergent
Consistent
Sampling/Denoising Trajectory

Experiments on High Dimensional Trajectory

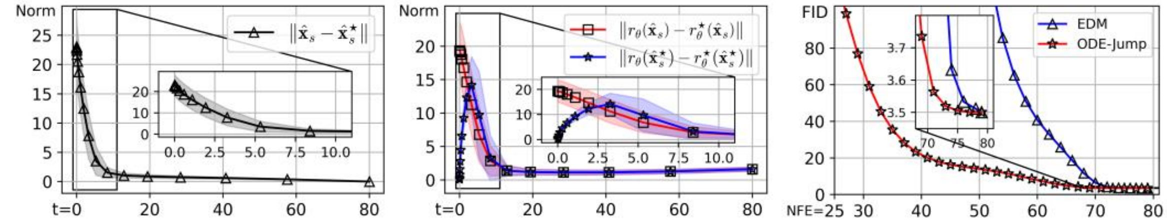


Figure 4: The score deviation in expectation (left and middle) and FID with different NFEs (right).

Observation 4. *The learned score is well-matched to the optimal score in the large-noise region (from* 80 *to around* 10*), otherwise they may diverge or almost coincide depending on different regions*

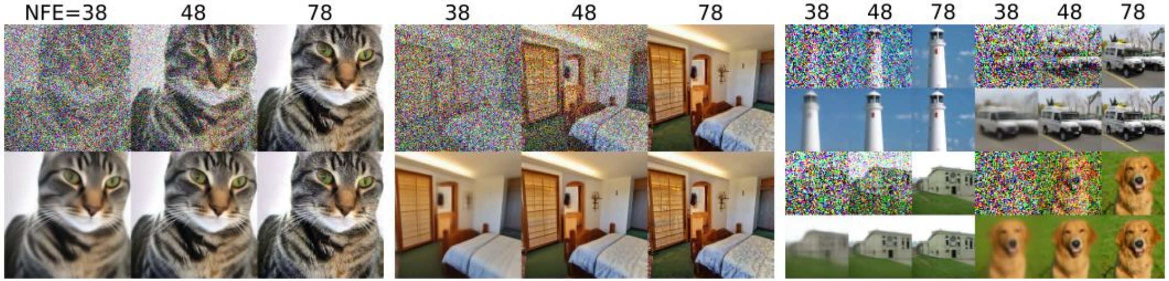


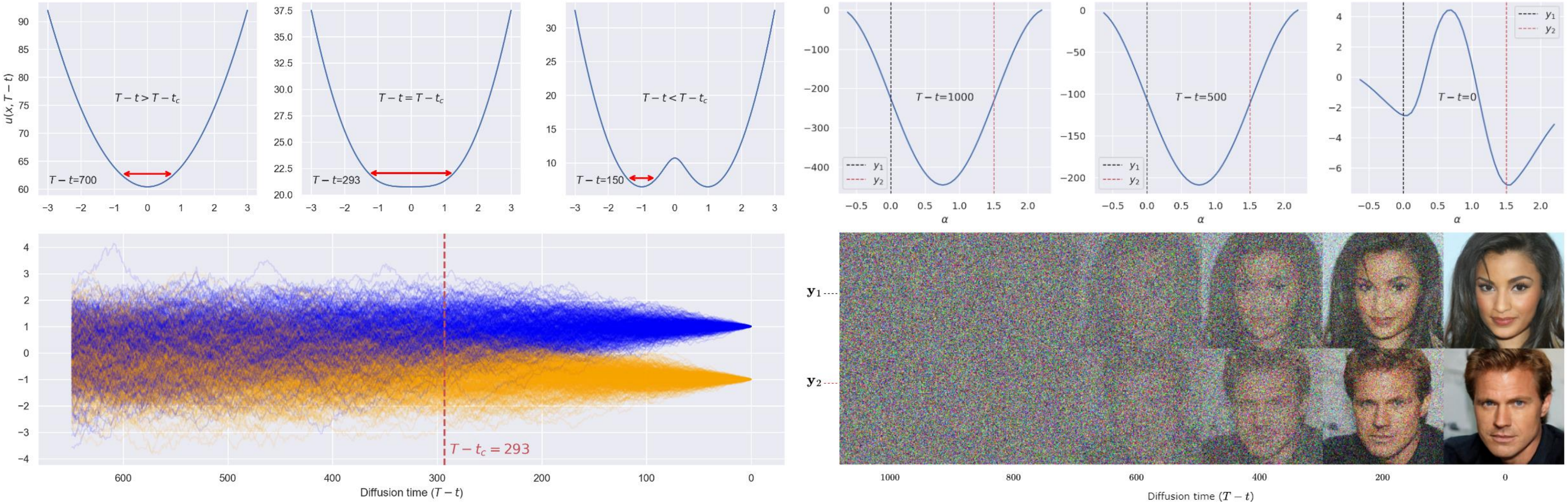
Figure 5: The synthesized images of our proposed ODE-Jump sampling (bottom) converge much faster than that of EDMs [KAAL22] (top) in terms of visual quality.

Observation 5. *The (optimal) denoising trajectory converges faster than the (optimal) sampling trajectory in terms of visual quality.*

"In fact, our learned score has to moderately diverge from the optimum to guarantee the generative ability."

Spontaneous symmetry breaking in generative diffusion models*

$$d\mathbf{X}_t = -\nabla_{\mathbf{x}} u(\mathbf{X}_t, T-t)\mathrm{d}t \; + g(T-t)\mathrm{d}\mathbf{W}_t$$



(a) Symmetry breaking in 1D diffusion model

(b) Symmetry breaking in CelebA HQ 256x256

$$u(\boldsymbol{x}, s) = -g^2(s)\log p(\boldsymbol{x}, s) + \int_{\mathbf{0}}^{\boldsymbol{x}} f(\boldsymbol{z}, s)\cdot \mathrm{d}\boldsymbol{z}$$

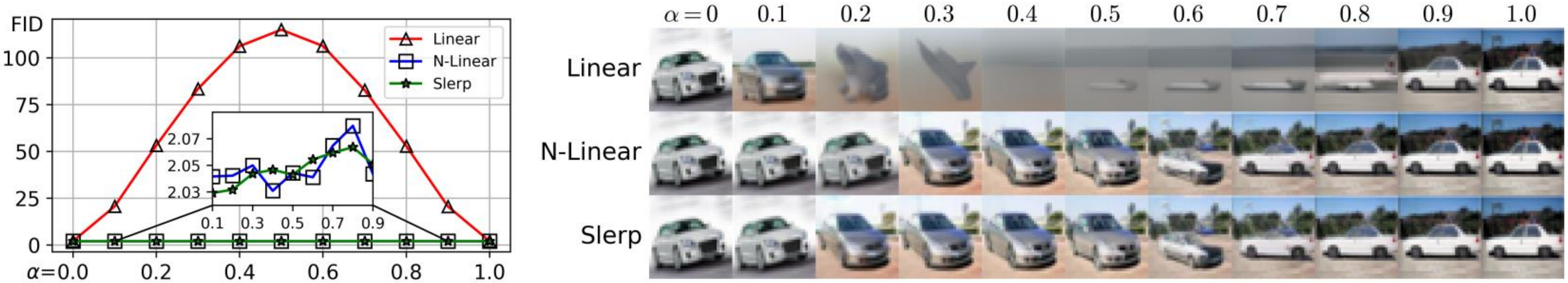
$$u(x,t) = \beta(T-t)\left(-\frac{1}{4}x^2 - \log\left(e^{-\frac{(x-\theta_{T-t})^2}{2(1-\theta_{T-t}^2)}} + e^{-\frac{(x+\theta_{T-t})^2}{2(1-\theta_{T-t}^2)}}\right)\right)$$

*[2305.19693] Spontaneous symmetry breaking in generative diffusion models (arxiv.org)

Gabriel Raya | Spontaneous symmetry breaking in generative diffusion models

In-Distribution Latent Interpolation

Proposition 5. *In high dimensions, linear interpolation [HJA20] shifts the latent distribution while spherical linear interpolation [SME21] asymptotically* ($d \to \infty$) *maintains the latent distribution.*

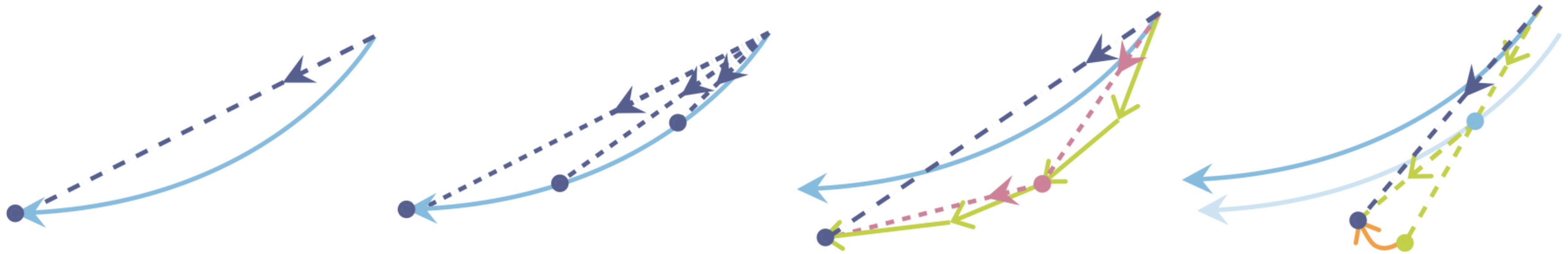


(a) The comparison of FID.

(b) Visualization of latent interpolation with different strategies.

Figure 6: Linear latent interpolation results in blurry images, while a simple re-scaling trick greatly preserves the fine-grained image details and enables a smooth traversal among different modes.

Rethinking Distillation-Based Fast Sampling Techniques



(a) KD [LL21]. (b) DFNO [ZNV+22]. (c) PD [SH22]. (d) CD [SDCS23].

Figure 7: The comparison of distillation-based techniques. The *offline* techniques first simulate a long ODE trajectory with the teacher score and then make the student score points to the final point (KD [LL21]) or also include intermediate points on the trajectory (DFNO [ZNV+22]). The *online* techniques iteratively fine-tune the student prediction to align with the target simulated by a few-step teacher model along the sampling trajectory (PD [SH22]) or the denoising trajectory (CD [SDCS23]).

Content

- Motivation
- Geometric Perspective
- Conclusion

Conclusion

- √ Geometric perspective on (VE) diffusion models
- √ Origin of generative ability

- × Theoretical results do not entirely substantiate the observations
- × Limited to VE-ODE