上海交通大学
约翰·霍普克罗夫特
计算机科学中心
John Hopcroft Center for Computer Science

Why Rectified Flow is Better? Elucidating VP, VE and RF-based diffusion models

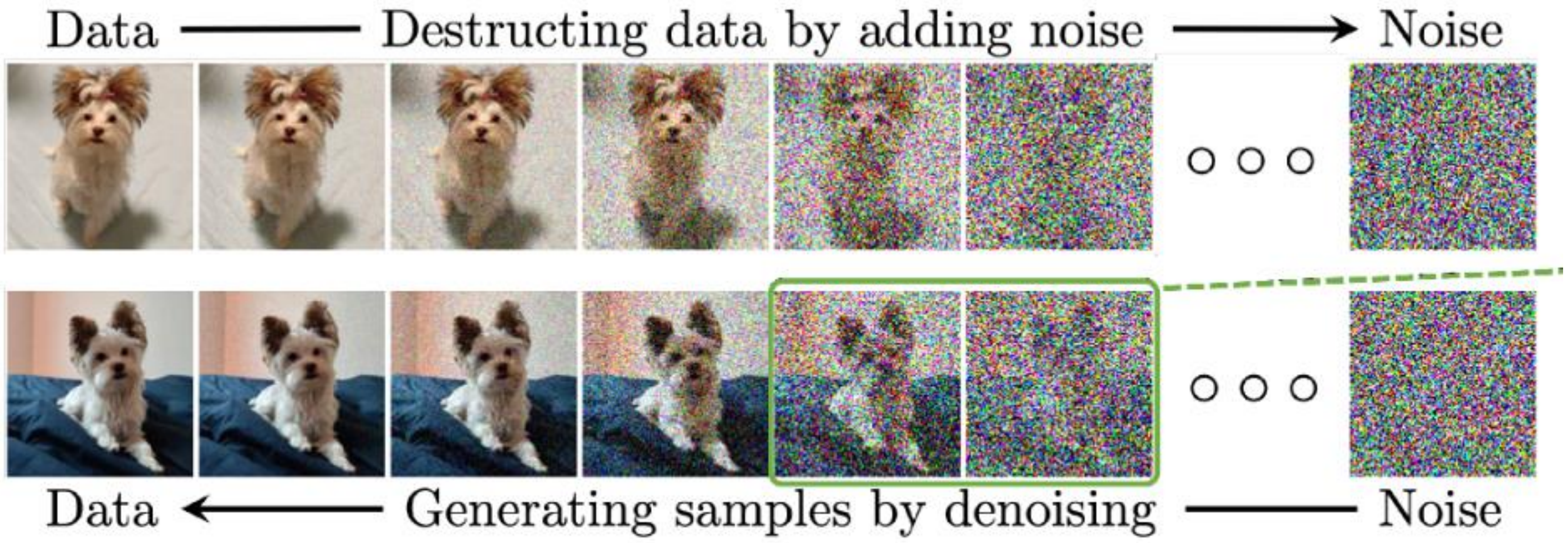
Ruofeng Yang

Shanghai Jiao Tong University

Supervisor: Shuai Li

The Paradigm of Diffusion Models

- A forward process and a Reverse Process

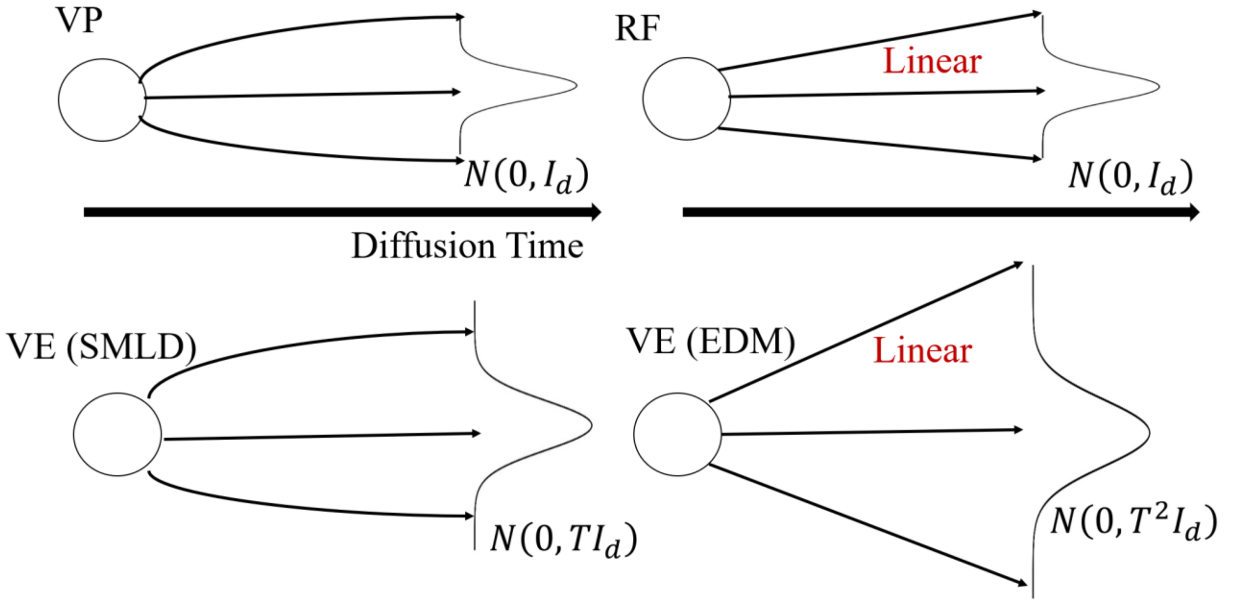


- The general forward process:

$$dX_t = f(X_t, t)dt + g(t)dB_t, X_0 \sim q_0 \in \mathbb{R}^d$$

Common forward processes:

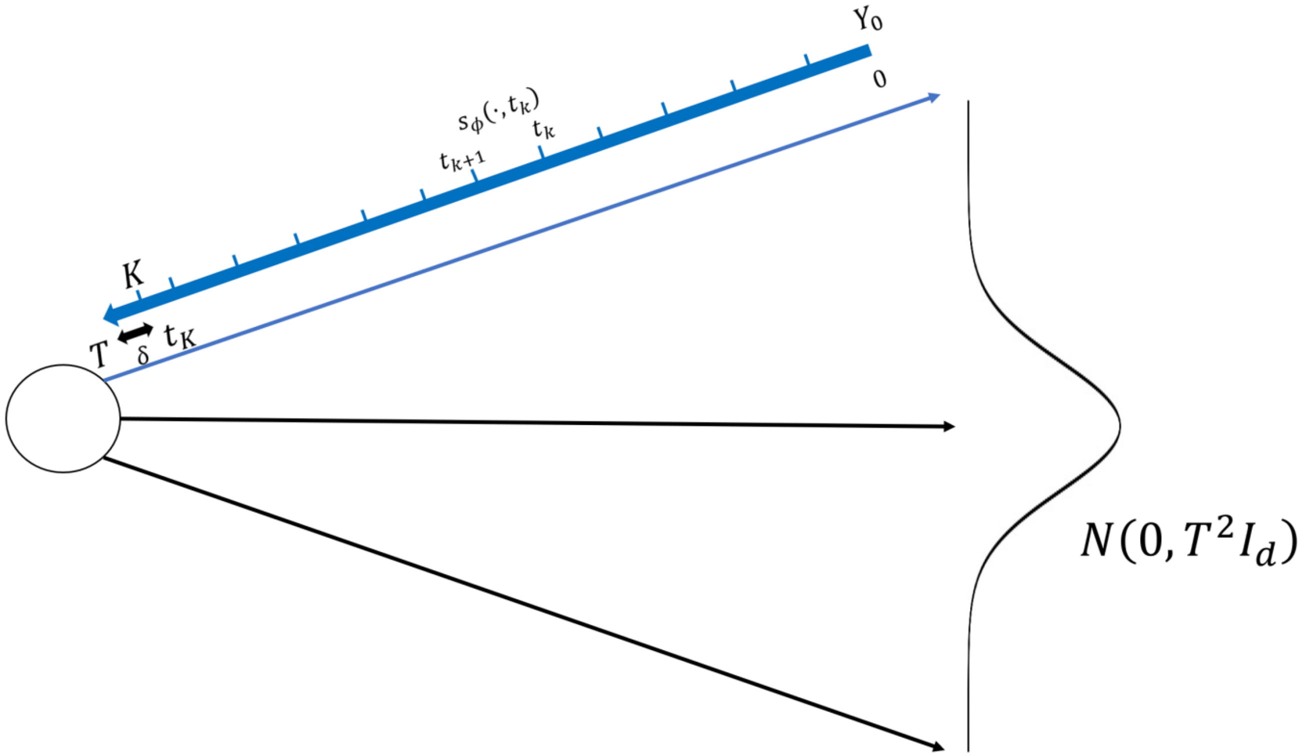
- Variance Preserving (VP): $f(X_t, t) = -\frac{1}{2}X_t, g(t) = 1$
- Variance Exploding (VE (SMLD)): $f(X_t, t) = 0, g(t) = \sqrt{2}$
- Variance Exploding (VE (EDM)): $f(X_t, t) = 0, g(t) = \sqrt{2t}$
- Rectified Flow: $X_t = (1 - t)X_0 + tZ, t \in [0,1]$



The Reverse Process

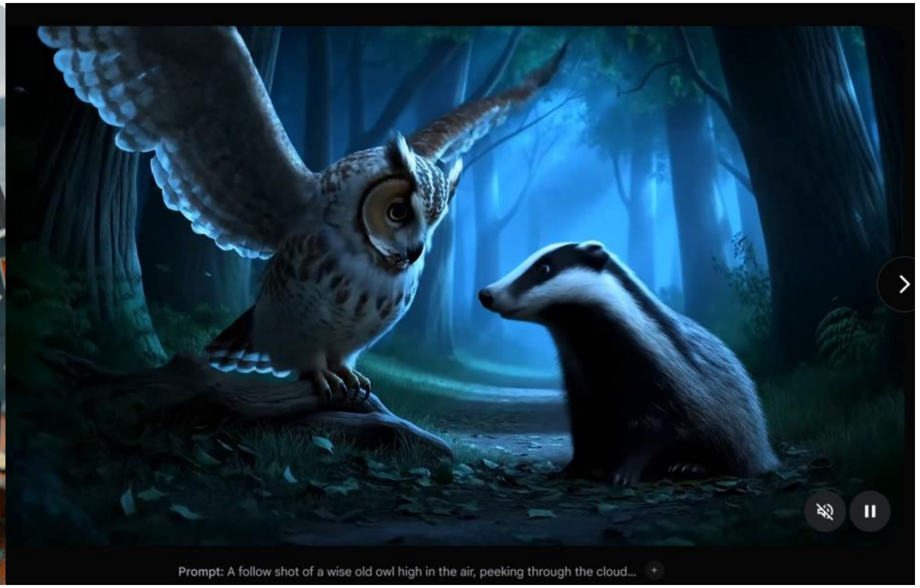
$$Y_{t'} = \left[f(Y_{t'}, T - t') - \frac{1+\eta^2}{2} g^2(T - t') \nabla \log q_{T-t'}(Y_{t'}) \right] dt' + \eta g(T - t') dB_{t'}, \eta \in [0,1]$$

- $\eta = 1 \rightarrow$ Reverse SDE; $\eta = 0 \rightarrow$ Reverse probability flow ODE (PFODE)

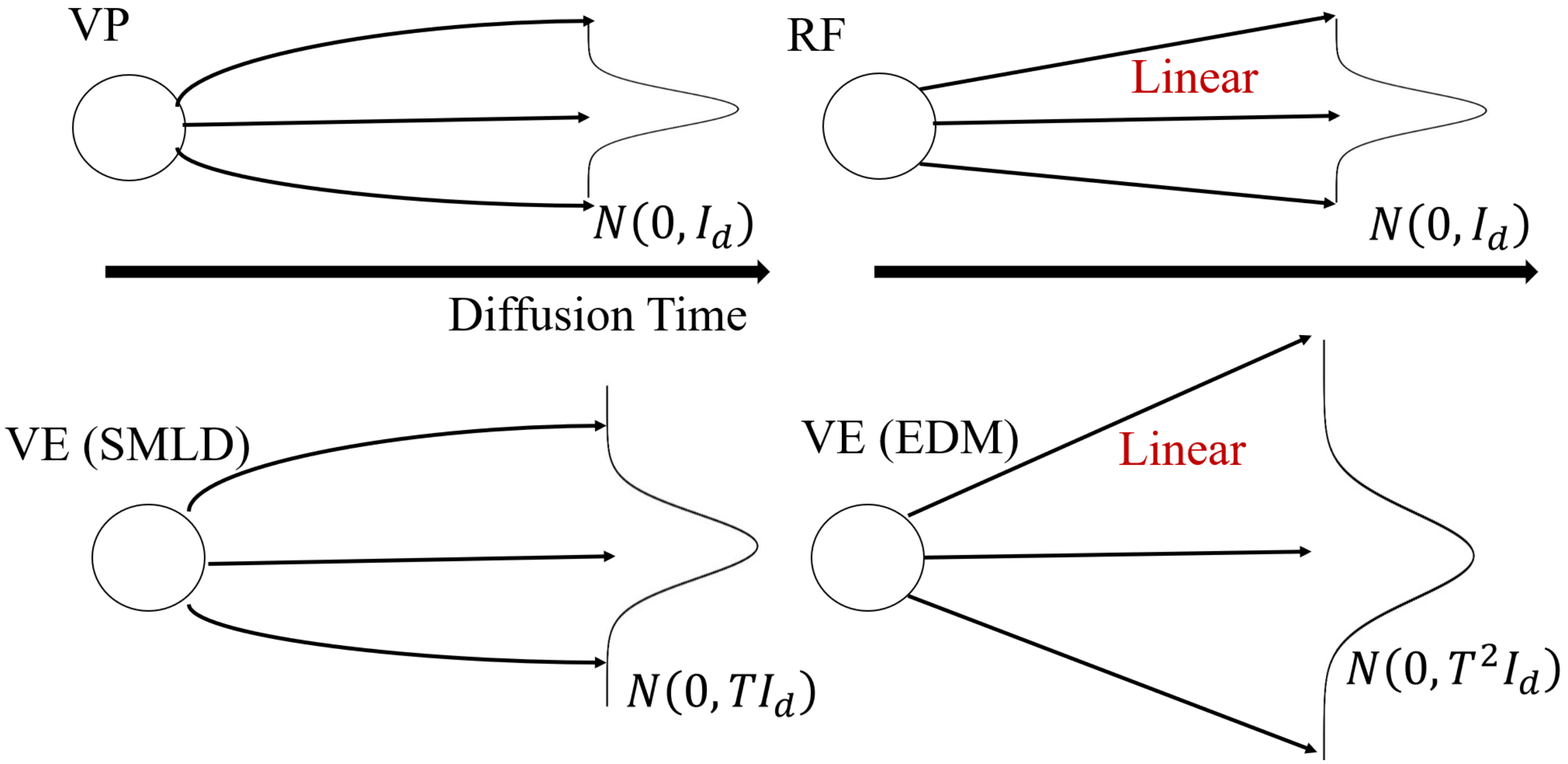


Motivations

- In the early year, many works adopt VP (SDXL) and VE (EDM).
- Since last year, RF becomes the main choice in computer vision and audio.
  - Image: SD 3, FLUX, Qwen-Image
  - Video: Seeddance, Wan 2.2
  - Video-audio joint generation: Veo3



Motivations



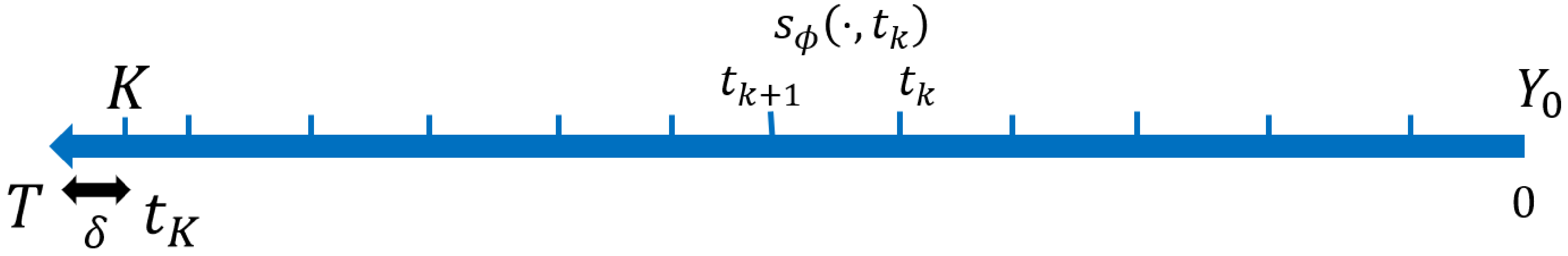
Why VE (EDM) is comparable with VP and RF is better?

Sample Complexity for Diffusion Models

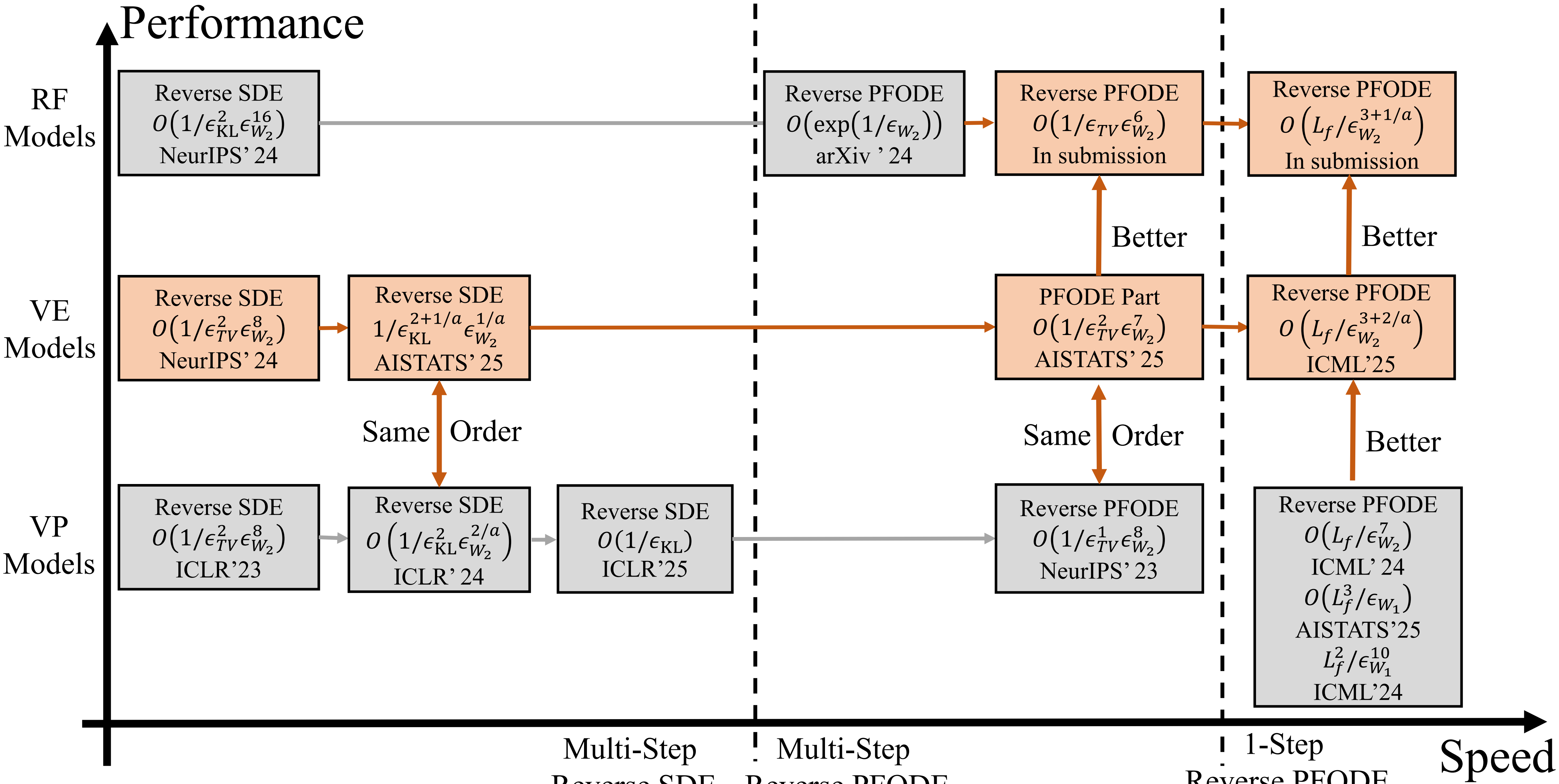
Assume an accurate enough score function

$$\left\|\log q_t(X,t) - s_\phi(X,t)\right\|_2^2 \leq \epsilon_{score}^2$$

The sample complexity K to guarantee $Dis\left(p_{t_K}, q_0\right) \leq \epsilon$.



Overview



Performance
RF Models
VE Models
VP Models
Reverse SDE
$O(1/\epsilon_{\mathrm{KL}}^2\epsilon_{W_2}^{16})$
NeurIPS' 24
Reverse PFODE
$O(\exp(1/\epsilon_{W_2}))$
arXiv ' 24
Reverse PFODE
$O(1/\epsilon_{TV}\epsilon_{W_2}^{6})$
In submission
Reverse PFODE
$O\left(L_f/\epsilon_{W_2}^{3+1/a}\right)$
In submission
Better
Better
Reverse SDE
$O(1/\epsilon_{TV}^2\epsilon_{W_2}^{8})$
NeurIPS' 24
Reverse SDE
$1/\epsilon_{\mathrm{KL}}^{2+1/a}\epsilon_{W_2}^{1/a}$
AISTATS' 25
PFODE Part
$O(1/\epsilon_{TV}^2\epsilon_{W_2}^{7})$
AISTATS' 25
Reverse PFODE
$O\left(L_f/\epsilon_{W_2}^{3+2/a}\right)$
ICML'25
Same Order
Same Order
Better
Reverse SDE
$O(1/\epsilon_{TV}^2\epsilon_{W_2}^{8})$
ICLR'23
Reverse SDE
$O\left(1/\epsilon_{\mathrm{KL}}^2\epsilon_{W_2}^{2/a}\right)$
ICLR' 24
Reverse SDE
$O(1/\epsilon_{\mathrm{KL}})$
ICLR'25
Reverse PFODE
$O(1/\epsilon_{TV}^1\epsilon_{W_2}^{8})$
NeurIPS' 23
Reverse PFODE
$O(L_f/\epsilon_{W_2}^7)$
ICML' 24
$O(L_f^3/\epsilon_{W_1})$
AISTATS'25
$L_f^2/\epsilon_{W_1}^{10}$
ICML'24
Multi-Step Reverse SDE
Multi-Step Reverse PFODE
1-Step Reverse PFODE
Speed

General Guarantee (Reverse SDE)

Theorem. Under the bounded support assumption, for diffusion models

$\mathrm{KL}(N(0,\sigma_T^2), q_T)$ Discretization

$$\mathrm{KL}(p_{T-\delta}, q_\delta) \leq \bar{D}^2 m_T/\sigma_T^2 + d^2 (T/\delta)^{\frac{1}{a}}/K \leq \tilde{O}\left(\epsilon_{\mathrm{KL}}^2\right)$$

?

$$\mathrm{W}_2^2(q_0, q_\delta) \leq \sigma_\delta^2 \leq \epsilon_{W_2}^2$$

?

- Balance: (a) T determined by the first term (b) discretization depends on T
- Influence by early stopping parameter δ

Discussion on Diffusion Time T

$$\mathrm{KL}(p_{T-\delta}, q_\delta) \le \bar{D}^2 m_T/\sigma_T^2 + d^2 (T/\delta)^{\frac{1}{a}}/K \overset{?}{\le} \tilde{O}(\epsilon_{\mathrm{KL}}^2)$$

$$\mathrm{W}_2^2(q_0, q_\delta) \le \sigma_\delta^2 \overset{?}{\le} \epsilon_{W_2}^2$$

$$\mathrm{KL}(p_{T-\delta}, q_\delta) \le \bar{D}^2 m_T/\sigma_T^2 + d^2 (T/\delta)^{\frac{1}{a}}/K$$

- VP enjoy an exponential-decay first term $m_T = e^{-T}$ and $\sigma_T = 1 \rightarrow$

  A logarithmic $T = \log(1/\epsilon_{TV})$

- VE has a polynomial-decay one $m_T = 1$ and $\sigma_T^2 = poly(T) \rightarrow$

  Large sample complexity

Discussion on Early Stopping δ

$$\mathrm{KL}(p_{T-\delta}, q_\delta) \le \bar{D}^2 m_T/\sigma_T^2 + d^2 (T/\delta)^{\frac{1}{a}}/K \le \tilde{O}(\epsilon_{\mathrm{KL}}^2)$$

?

$$\mathrm{W}_2^2(q_0, q_\delta) \le \sigma_\delta^2 \le \epsilon_{W_2}^2$$

?

$$\mathrm{W}_2^2(q_0, q_\delta) \le \sigma_\delta^2 \le \epsilon_{W_2}^2$$

- For VP, $\sigma_\delta^2 = \delta \to \delta = \epsilon_{W_2}^2$
- For VE (EDM), $\sigma_\delta^2 = \delta^2 \to \delta = \epsilon_{W_2}$
- VP better in T and VE (EDM) better in $\delta \to$ The same order results

VP: $K = O\left(1/\epsilon_{\mathrm{KL}}^2 \epsilon_{W_2}^{2/a}\right)$, VE (EDM) $K = O\left(1/\epsilon_{\mathrm{KL}}^{2+1/a} \epsilon_{W_2}^{1/a}\right)$

Worst of Both World: VE (SMLD)

$$\mathrm{KL}(p_{T-\delta}, q_\delta) \le \bar{D}^2 m_T/\sigma_T^2 + d^2 (T/\delta)^{\frac{1}{a}}/K \le \tilde{O}(\epsilon_{\mathrm{KL}}^2)$$

?

$$\mathrm{W}_2^2(q_0, q_\delta) \le \sigma_\delta^2 \le \epsilon_{W_2}^2$$

?

- For VE (SMLD), $\sigma_\delta^2 = \delta \to \delta = \epsilon_{W_2}^2$ and $m_T = 1, \sigma_T^2 = T$
- Bad in T and δ at the same time $\to O\left(1/\epsilon_{\mathrm{KL}}^{2+2/a} \epsilon_{W_2}^{2/a}\right)$

VP: $K = O\left(1/\epsilon_{KL}^2 \epsilon_{W_2}^{2/a}\right)$, VE (EDM) $K = O\left(1/\epsilon_{KL}^{2+1/a} \epsilon_{W_2}^{1/a}\right)$

Best of Both World: Rectified Flow

$$\mathrm{KL}(p_{T-\delta}, q_\delta) \leq \bar{D}^2 m_T/\sigma_T^2 + d^2 (T/\delta)^{\frac{1}{a}}/K \leq \tilde{O}\left(\epsilon_{\mathrm{KL}}^2\right)$$

?

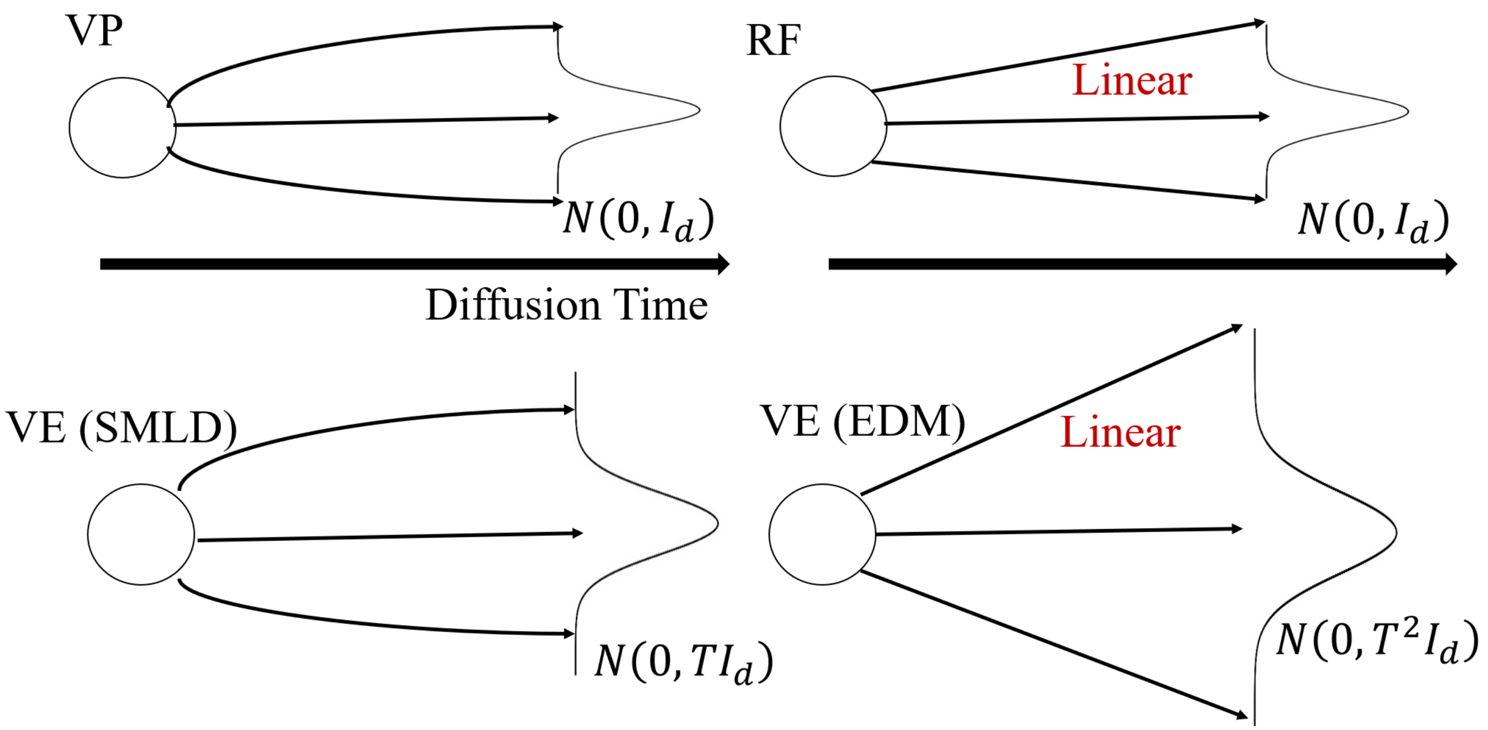
$$\mathrm{W}_2^2(q_0, q_\delta) \leq \sigma_\delta^2 \leq \epsilon_{W_2}^2$$

?

- $X_t = (1-t)X_0 + tZ, t \in [0,1] \rightarrow T = 1$
- Linear Interpolation: $\sigma_\delta^2 = \delta^2 \rightarrow \delta = \epsilon_{W_2}$
- Good in T and δ at the same time $\rightarrow O\left(1/\epsilon_{\mathrm{KL}}^2 \epsilon_{W_2}^{1/a}\right)$

VP: $K = O\left(1/\epsilon_{KL}^2 \epsilon_{W_2}^{2/a}\right)$, VE (EDM) $K = O\left(1/\epsilon_{KL}^{2+1/a} \epsilon_{W_2}^{1/a}\right)$

1-Step Consistency Models &InstaFlow



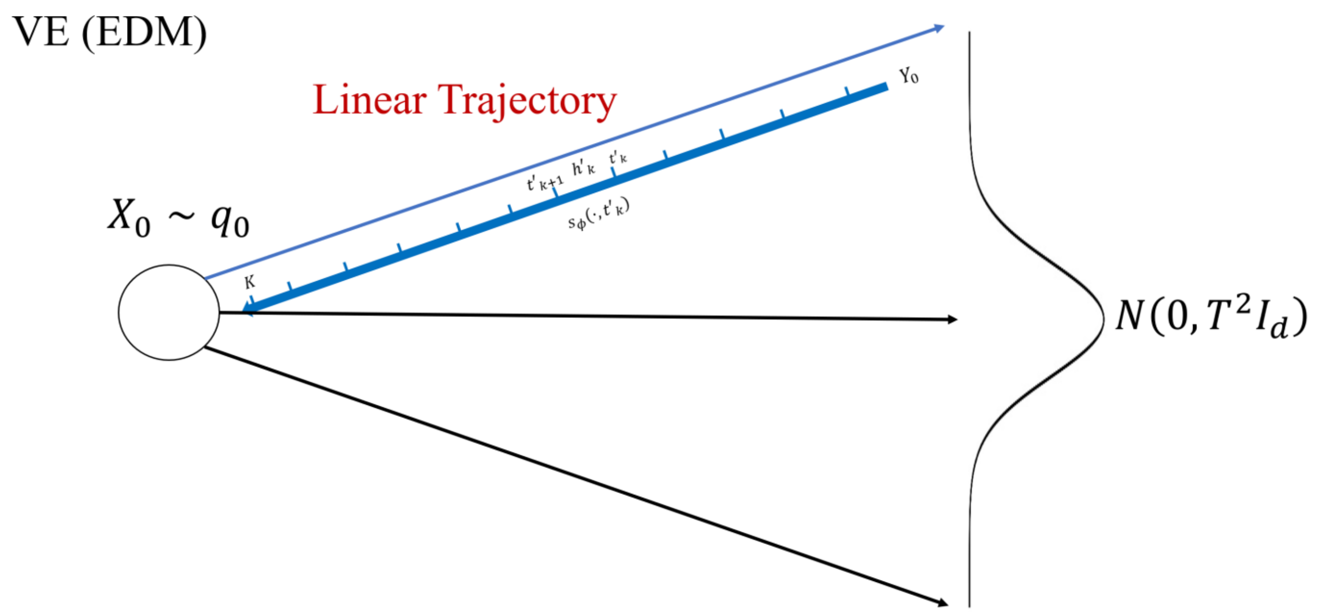
Due to the linear property, RF and VE (EDM) are used as the basic of One-step generation.

Recall: Reverse Process

- Reverse forward process → Reverse process ($t' = T - t$ and $Y_{t'} = X_{T-t'}$)

$$Y_{t'} = \left[f(Y_{t'}, T - t') - \frac{1+\eta^2}{2} g^2(T - t')\nabla \log q_{T-t'}(Y_{t'})\right] dt' + \eta g(T - t')dB_{t'}, \eta \in [0,1]$$

- $\eta = 0$ → Reverse probability flow ODE (PFODE, deterministic sampler)



The Paradigm of Consistency Models

- Based on diffusion models, to fast generate:

Consistency models, an one-step generation models

- For PFODE

$$\mathrm{d}Y_{t'} = v(Y_{t'}, t')\mathrm{d}t', Y_0 \sim q_T$$

the corresponding mapping function is

$$f^v(Y_{t'}, t') = Y_{T-\delta} = X_\delta, \forall t' \in [0, T-\delta]$$

- The property of mapping function:

$$f^v(Y_{t'}, t') = f^v(Y_{t''}, t''), \forall 0 \leq t'', t' \leq T-\delta$$
$$f^v(Y, T-\delta) = Y, \forall Y \in \mathbb{R}^d$$

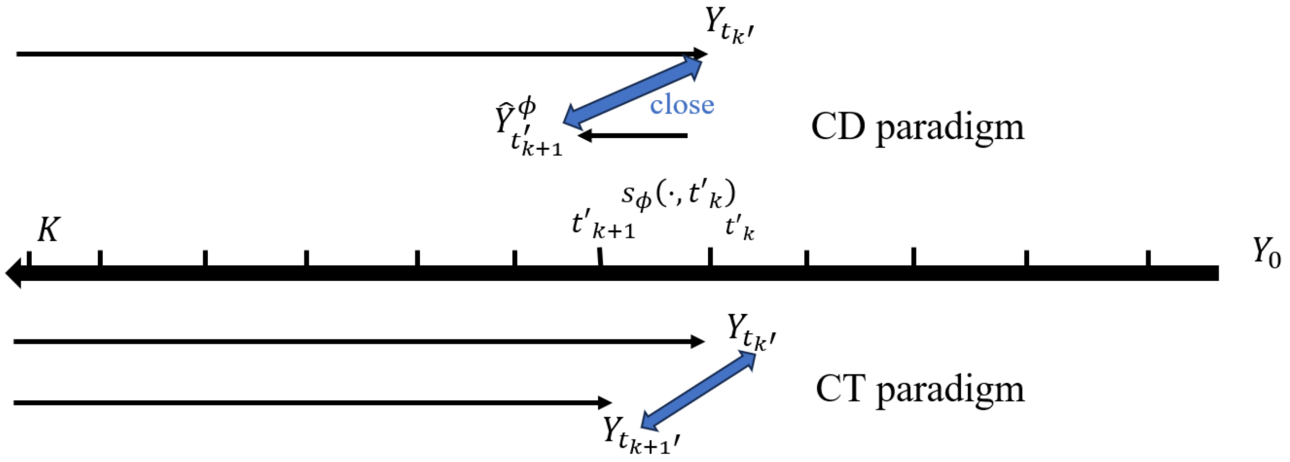
Goal: Consistency function $f_\theta(Y_t, t)$

- Consistency Distillation (CD) Paradigm:

Let $\hat{Y}^{\phi}_{t'_{k+1}}$ be the output running one step PFODE from$Y_{t'_k}$ with s_ϕ.

$$\mathcal{L}^K_{\text{CD}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \boldsymbol{\phi}) := \mathbb{E}_{X_0}\left[\mathbb{E}_{Y_{t'_k}|X_0}\left\|\boldsymbol{f_\theta}\left(Y_{t'_k}, t'_k\right) - \boldsymbol{f_{\theta^-}}\left(\hat{Y}^{\phi}_{t'_{k+1}}, t'_{k+1}\right)\right\|_2^2\right]$$

- Consistency Training: $\mathcal{L}^K_{\text{CT}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-) := \mathbb{E}_{X_0}\left[\mathbb{E}_{Y_{t'_k}|X_0}\left\|\boldsymbol{f_\theta}\left(Y_{t'_k}, t'_k\right) - \boldsymbol{f_{\theta^-}}\left(Y_{t'_{k+1}}, t'_{k+1}\right)\right\|_2^2\right]$



Discretization Complexity of Consistency Models

- Objective Function

$$\mathcal{L}_{\mathrm{CD}}^{K}(\boldsymbol{\theta}, \boldsymbol{\theta}^{-}; \boldsymbol{\phi}) := \mathbb{E}_{X_0}\left[\mathbb{E}_{Y_{t_k'}|X_0}\left\|\boldsymbol{f}_{\boldsymbol{\theta}}\left(Y_{t_k'}, t_k'\right) - \boldsymbol{f}_{\boldsymbol{\theta}^{-}}\left(\hat{Y}_{t_{k+1}'}^{\phi}, t_{k+1}'\right)\right\|_2^2\right]$$

- Large K: Training is time-consuming.

- Small K: Training is hard since $Y_{t_k'}$ and $\hat{Y}_{t_{k+1}'}^{\phi}$ is too far away.

- Choosing a suitable K in the training phase to guarantee

$$W_2\left(f_\theta\left(N(0, \sigma_T^2 I_d)\right), q_0\right) \le \epsilon_{W_2}$$

Current Discretization Complexity Results

- Many works focus on VP models instead of VE (EDM) and RF

  Assuming the consistency function f_θ (or f_v) is L_f Lipschitz

  They achieve discretization complexity with

  - (1) Bad dependence on ϵ: $L_f/\epsilon_{W_2}^{7}$ [1] and $L_f/\epsilon_{W_1}^{10}$ [2] or
  - (2) Large L_f dependence: L_f^3/ϵ_{W_1} [3]

- Far away from the SOTA sample complexity of $1/\epsilon_{W_2}^{4}$ of diffusion models.

Similar Balance Between δ and T

Theorem. For one-step generation models, using VE(EDM) as a example

$$W_2\left(f_\theta\left(N(0,\sigma_T^2 I_d)\right), q_0\right) \le \frac{R^2}{T} + \frac{L_f R^2 (R+\sqrt{d})(T/\delta)^{\frac{1}{a}}}{K\delta^2} + \sqrt{d}\delta$$

0 for RF

- Heavily influenced by δ → VE(EDM) and RF is great
- For VE (EDM) $O\left(L_f/\epsilon_{W_2}^{3+2/a}\right)$ → Better than previous $L_f/\epsilon_{W_2}^{7}$ and L_f^3/ϵ_{W_1}
- RF is free from T → Better $K = O\left(L_f/\epsilon_{W_2}^{3+1/a}\right)$

Reverse PFODE
$O(L_f/\epsilon_{W_2}^7)$
ICML' 24
$O(L_f^3/\epsilon_{W_1})$
AISTATS'25
$L_f^2/\epsilon_{W_1}^{10}$
ICML'24

→ Better →

Reverse PFODE
$O\left(L_f/\epsilon_{W_2}^{3+2/a}\right)$
ICML'25

→ Better →

Reverse PFODE
$O\left(L_f/\epsilon_{W_2}^{3+1/a}\right)$
In submission

Conclusion-Theory

- From the complexity perspective, RF is great in diffusion time T and early stopping δ.
- Future work (Theory):

Many people say flow-matching and score matching is totally equal:

Then, why FM training paradigm is better?

Conclusion-Application

- Future work (Application):

To design a better and efficient noising and denoising process with better theoretical guarantee and great performance.

Thanks!

Q&A