Characteristic Learning for Provable One Step Generation

Chenguang Duan
Wuhan University

Joint work with **Zhao Ding** (WHU), **Yuling Jiao** (WHU), **Ruoxuan Li** (WHU), **Jerry Zhijian Yang** (WHU), and **Pingwen Zhang** (WHU, PKU)

Outline

Backgrounds of the Generative Learning

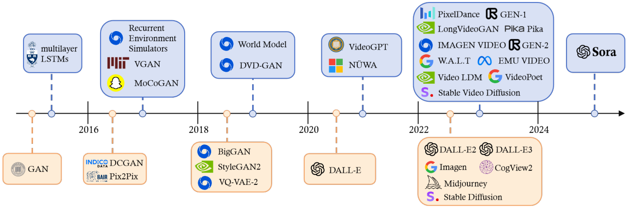


Figure: History of Vision Generative AI

Prompt:
This ink sketch-style illustration depicts a small hedgehog holding a piece of watermelon with its tiny paws, happily closing its eyes and taking small bites.

→ DALL-E 3 →

Generative Learning

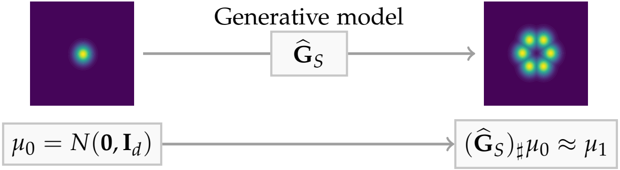
Problem Setup

- Known initial distribution μ_0 on $\mathbb{R}^m$.
- Samples drawn from an unknown target dsitribution on $\mathbb{R}^d$:

$$S = \{\mathbf{X}_i\}_{i=1}^n \sim^{\text{i.i.d.}} \mu_1.$$

- Learning a push-forward map $\widehat{\mathbf{G}}_S : \mathbb{R}^m \to \mathbb{R}^d$ using data set S such that:

$$(\widehat{\mathbf{G}}_S)_\sharp \mu_0 \approx \mu_1.$$



Stochastic Interpolations

$\mathbf{X}_0 \sim \mu_0$ → $\mathbf{X}_t := \alpha_t \mathbf{X}_0 + \beta_t \mathbf{X}_1$ → $\mathbf{X}_1 \sim \mu_1$

Interpolation between μ_0 and μ_1:

$$p_t := \mathrm{Law}(\mathbf{X}_t) = \frac{1}{\alpha_t} \int \mu_0\Big(\frac{\mathbf{x} - \beta_t \mathbf{x}_1}{\alpha_t}\Big)\mu_1(\mathbf{x}_1)\mathrm{d}\mathbf{x}_1, \quad t \in (0,1).$$

- Interpolation coefficients: $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = 0$.

	α_t	β_t
Linear	$1 - t$	t
Föllmer flow	$\sqrt{1 - t^2}$	t

Flow-based Generative Models

The interpolation density p_t satisfies the **transport equation**:

$$\begin{cases} \partial_t p_t(\mathbf{x}) + \nabla \cdot (\mathbf{b}^*(t,\mathbf{x}) p_t(\mathbf{x})) = 0, \quad \mathbf{x} \in \mathbb{R}^d, \ t \in (0,1), \\ p_0(\mathbf{x}) = \mu_0(\mathbf{x}), \end{cases}$$

where the **velocity field** $\mathbf{b}^* : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is defined as $\mathbf{b}^*(t,\mathbf{x}) = \mathbb{E}\left[\dot{\alpha}_t \mathbf{X}_0 + \dot{\beta}_t \mathbf{X}_1 \middle| \mathbf{X}_t = \mathbf{x}\right]$.

Flow-based Generative Models

The interpolation density p_t satisfies the **transport equation**:

$$\begin{cases} \partial_t p_t(\mathbf{x}) + \nabla \cdot (\mathbf{b}^*(t,\mathbf{x}) p_t(\mathbf{x})) = 0, \quad \mathbf{x} \in \mathbb{R}^d,\ t \in (0,1), \\ p_0(\mathbf{x}) = \mu_0(\mathbf{x}), \end{cases}$$

where the **velocity field** $\mathbf{b}^* : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is defined as $\mathbf{b}^*(t,\mathbf{x}) = \mathbb{E}\left[\dot{\alpha}_t \mathbf{X}_0 + \dot{\beta}_t \mathbf{X}_1 \middle| \mathbf{X}_t = \mathbf{x}\right]$.

The **characteristic** of the transport equation: **Probability Flow ODE**

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{b}^*(t,\mathbf{x}(t)), \quad t \in (0,1), \quad \mathbf{x}(0) = \mathbf{x}_0 \sim \mu_0.$$

The marginal density of $\mathbf{x}(t)$ is p_t

- *How can we estimate the velocity fields $\mathbf{b}^* : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$?*
- *How can we sample from the target distribution given an estimator of the velocity field?*

Training Phase

Learn the velocity field $\mathbf{b}^* : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$.

$$\mathbf{b}^*(t, \mathbf{x}) = \mathbb{E}\left[\dot{\alpha}_t \mathbf{X}_0 + \dot{\beta}_t \mathbf{X}_1 \middle| \mathbf{X}_t = \mathbf{x}\right]$$

Training Phase

Learn the velocity field $\mathbf{b}^* : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$.

$$\mathbf{b}^*(t, \mathbf{x}) = \mathbb{E}\left[\dot{\alpha}_t \mathbf{X}_0 + \dot{\beta}_t \mathbf{X}_1 \middle| \mathbf{X}_t = \mathbf{x}\right]$$

- Population risk:

$$\mathbf{b}^* = \operatorname*{arg\,min}_{\mathbf{b}:\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}^d} L(\mathbf{b}) := \int_0^T \mathbb{E}_{(\mathbf{X}_0,\mathbf{X}_1)\sim\mu_0\times\mu_1}\left[\|\dot{\alpha}_t \mathbf{X}_0 + \dot{\beta}_t \mathbf{X}_1 - \mathbf{b}(t, \alpha_t \mathbf{X}_0 + \beta_t \mathbf{X}_1)\|_2^2\right] \mathrm{d}t.$$

Training Phase

Learn the velocity field $\mathbf{b}^* : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$.

$$\mathbf{b}^*(t, \mathbf{x}) = \mathbb{E}\left[\dot{\alpha}_t \mathbf{X}_0 + \dot{\beta}_t \mathbf{X}_1 \middle| \mathbf{X}_t = \mathbf{x}\right]$$

- Population risk:

$$\mathbf{b}^* = \underset{\mathbf{b}:\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}^d}{\arg\min} L(\mathbf{b}) := \int_0^T \mathbb{E}_{(\mathbf{X}_0,\mathbf{X}_1)\sim\mu_0\times\mu_1}\left[\|\dot{\alpha}_t \mathbf{X}_0 + \dot{\beta}_t \mathbf{X}_1 - \mathbf{b}(t, \alpha_t \mathbf{X}_0 + \beta_t \mathbf{X}_1)\|_2^2\right] \mathrm{d}t.$$

- Empirical risk minimization:

$$\widehat{\mathbf{b}} \in \underset{\mathbf{b}\in\mathscr{B}}{\arg\min}\, \widehat{L}(\mathbf{b}) := \frac{1}{n}\sum_{i=1}^n \|\dot{\alpha}_{t^{(i)}} \mathbf{X}_0^{(i)} + \dot{\beta}_{t^{(i)}} \mathbf{X}_1^{(i)} - \mathbf{b}(t, \alpha_{t^{(i)}} \mathbf{X}_0^{(i)} + \beta_{t^{(i)}} \mathbf{X}_1^{(i)})\|_2^2.$$

  - $\mathscr{B}$ is a neural network class.
  - $(\mathbf{X}_0^{(1)}, \mathbf{X}_1^{(1)}), \ldots, (\mathbf{X}_0^{(n)}, \mathbf{X}_1^{(n)}) \sim^{\text{i.i.d.}} \mu_0 \times \mu_1$, and $t^{(1)}, \ldots, t^{(n)} \sim^{\text{i.i.d.}} \mathrm{unif}(0, T)$

Inference Phase

Euler discretization of the probability flow ODE:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \widehat{\mathbf{b}}(t, \mathbf{x}(t)), \quad t \in (0,1), \quad \mathbf{x}(0) = \mathbf{x}_0 \sim \mu_0.$$

- Time points: $0 = t_0 < t_1 < \ldots < t_K = T$ with $t_k = hk$.
- Euler approximation:

$$\frac{\mathrm{d}\widehat{\mathbf{x}}(t)}{\mathrm{d}t} = \widehat{\mathbf{b}}(t, \widehat{\mathbf{x}}(t_{k-1})), \quad t \in (t_{k-1}, t_k), \quad 1 \leq k \leq K, \quad \widehat{\mathbf{x}}(0) = \widehat{\mathbf{x}}_0 \sim \mu_0.$$

  or, equivalently,

$$\widehat{\mathbf{x}}(t_k) = \widehat{\mathbf{x}}(t_{k-1}) + h\widehat{\mathbf{b}}(t_{k-1}, \widehat{\mathbf{x}}(t_{k-1})), \quad 1 \leq k \leq K, \quad \widehat{\mathbf{x}}_0 \sim \mu_0. \qquad \mathrm{Law}(\widehat{\mathbf{x}}_K) \approx \mu_1$$

Inference Phase

Euler discretization of the probability flow ODE:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \widehat{\mathbf{b}}(t, \mathbf{x}(t)), \quad t \in (0,1), \quad \mathbf{x}(0) = \mathbf{x}_0 \sim \mu_0.$$

- Time points: $0 = t_0 < t_1 < \ldots < t_K = T$ with $t_k = hk$.
- Euler approximation:

$$\frac{\mathrm{d}\widehat{\mathbf{x}}(t)}{\mathrm{d}t} = \widehat{\mathbf{b}}(t, \widehat{\mathbf{x}}(t_{k-1})), \quad t \in (t_{k-1}, t_k), \quad 1 \leq k \leq K, \quad \widehat{\mathbf{x}}(0) = \widehat{\mathbf{x}}_0 \sim \mu_0.$$

  or, equivalently,

$$\widehat{\mathbf{x}}(t_k) = \widehat{\mathbf{x}}(t_{k-1}) + h\widehat{\mathbf{b}}(t_{k-1}, \widehat{\mathbf{x}}(t_{k-1})), \quad 1 \leq k \leq K, \quad \widehat{\mathbf{x}}_0 \sim \mu_0. \qquad \mathrm{Law}(\widehat{\mathbf{x}}_K) \approx \mu_1$$

Sampling from flow-based generative models is computationally expensive because it requires numerous evaluations of the large-scale velocity neural network $\widehat{\mathbf{b}}$.

Outline

Characteristic Learning: Distillation for One-Step Generation

The probability flow ODE specify a **deterministic** flow map:

$$\mathbf{g}^*_{t,s} : \mathbb{R}^d \to \mathbb{R}^d, \quad \mathbf{x}_t \mapsto \mathbf{x}_s, \quad 0 \leq t \leq s < 1.$$

$$(\mathbf{g}^*_{t,s})_\sharp p_t \approx p_s$$

Goal: Estimate the two-parameter flow map $\mathbf{g}^*_{t,s}$ by a deep neural network.

Characteristic Learning: Distillation for One-Step Generation

The probability flow ODE specify a **deterministic** flow map:

$$\mathbf{g}^*_{t,s} : \mathbb{R}^d \to \mathbb{R}^d, \quad \mathbf{x}_t \mapsto \mathbf{x}_s, \quad 0 \le t \le s < 1.$$

$(\mathbf{g}^*_{t,s})_\sharp p_t \approx p_s$

Goal: Estimate the two-parameter flow map $\mathbf{g}^*_{t,s}$ by a deep neural network.

Characteristic Matching:

Approximate characteristics generated by Euler discretization: $\{\widehat{\mathbf{x}}_k^{(i)} := \widehat{\mathbf{x}}^{(i)}(t_k)\}_{i=1}^n$.

$$\widehat{\mathbf{g}} \in \operatorname*{arg\,min}_{\mathbf{g}\in\mathscr{G}} \widehat{R}(\mathbf{g}) := \frac{2}{mK^2} \sum_{i=1}^m \sum_{k=0}^{K-1} \Big\{ \frac{1}{2} \|\widehat{\mathbf{x}}_k^{(i)} - \mathbf{g}(t_k, t_k, \widehat{\mathbf{x}}_k^{(i)})\|_2^2 + \sum_{\ell=k+1}^{K-1} \|\widehat{\mathbf{x}}_\ell^{(i)} - \mathbf{g}(t_k, t_\ell, \widehat{\mathbf{x}}_k^{(i)})\|_2^2 \Big\}.$$

Characteristic Learning: Distillation for One-Step Generation

The probability flow ODE specify a **deterministic** flow map:

$$\mathbf{g}^*_{t,s} : \mathbb{R}^d \to \mathbb{R}^d, \quad \mathbf{x}_t \mapsto \mathbf{x}_s, \quad 0 \leq t \leq s < 1.$$

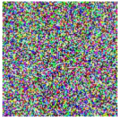
$(\mathbf{g}^*_{t,s})_\sharp p_t \approx p_s$

Goal: Estimate the two-parameter flow map $\mathbf{g}^*_{t,s}$ by a deep neural network.

Characteristic Matching:

Approximate characteristics generated by Euler discretization: $\{\widehat{\mathbf{x}}^{(i)}_k := \widehat{\mathbf{x}}^{(i)}(t_k)\}_{i=1}^n$.

$$\widehat{\mathbf{g}} \in \operatorname*{arg\,min}_{\mathbf{g}\in\mathscr{G}} \widehat{R}(\mathbf{g}) := \frac{2}{mK^2} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \left\{ \frac{1}{2} \|\widehat{\mathbf{x}}^{(i)}_k - \mathbf{g}(t_k, t_k, \widehat{\mathbf{x}}^{(i)}_k)\|_2^2 + \sum_{\ell=k+1}^{K-1} \|\widehat{\mathbf{x}}^{(i)}_\ell - \mathbf{g}(t_k, t_\ell, \widehat{\mathbf{x}}^{(i)}_k)\|_2^2 \right\}.$$



One-Step Generation

$\widehat{\mathbf{g}}_{0,T}$

$\mathbf{X}_0 \sim \mu_0 = N(\mathbf{0}, \mathbf{I}_d)$ → $\widehat{\mathbf{g}}_{0,T}(\mathbf{X}_0) \sim \widehat{\mu}_T \approx \mu_1$

Outline

Assumptions

- **A1.** (Prior distribution) The prior distribution $\mu_0 = N(\mathbf{0}, \mathbf{I}_d)$.
- **A2.** (Target distribution) There exists an unknown constant $\sigma > 0$, such that

$$\mu_1(\mathbf{x}) = N(\mathbf{0}, \sigma^2 \mathbf{I}_d) * \nu := \int \varphi_d\left(\frac{\mathbf{x} - \mathbf{x}'}{\sigma}\right) \mathrm{d}\nu(\mathbf{x}'),$$

  where φ_d represents the d-dimensinoal density of the standard Gaussian distribution, and $\mathrm{d}\nu(\mathbf{x}) = p(\mathbf{x})\mathrm{d}\mathbf{x}$ with $\mathrm{supp}(\nu) \subseteq [0, 1]^d$.

Regularity of Probability ODE

Under Assumptions **A1.** and **A2**.

- Local bounded velocity:

$$\|\mathbf{b}^*(t,\mathbf{x})\|_\infty \le B_{\mathrm{vel}} R, \quad 0 \le t \le 1,\ \mathbf{x} \in \mathbb{B}_R^\infty.$$

- Bounded spatial gradient of velocity:

$$\|\nabla \mathbf{b}^*(t,\mathbf{x})\|_{\mathrm{op}} \le G, \quad 0 \le t \le 1,\ \mathbf{x} \in \mathbb{R}^d.$$

- Bounded time derivative of velocity:

$$\|\partial_t \mathbf{b}^*(t,\mathbf{x})\|_2 \le D \sup_{t\in[0,T]} \left(\frac{\dot{\alpha}_t^2}{\alpha_t^2} + \frac{|\ddot{\alpha}_t|}{\alpha_t} \right) R, \quad 0 \le t \le T < 1,\ \mathbf{x} \in \mathbb{B}_R^\infty.$$

Regularity of Probability ODE

Under Assumptions **A1.** and **A2**.

- Local bounded flow map:

$$\|\mathbf{g}^*(t,s,\mathbf{x})\|_\infty \leq B_{\mathrm{flow}}R, \quad 0 \leq t \leq s \leq 1,\ \mathbf{x} \in \mathbb{B}_R^\infty.$$

- Bounded spatial gradient of flow map:

$$\|\nabla\mathbf{g}^*(t,s,\mathbf{x})\|_{\mathrm{op}} \leq \exp(G(s-t)), \quad 0 \leq t \leq 1,\ \mathbf{x} \in \mathbb{B}_R^\infty.$$

- Local bounded time derivative of flow map:

$$\max\left\{\|\partial_t\mathbf{g}^*(t,s,\mathbf{x})\|_2, \|\partial_s\mathbf{g}^*(t,s,\mathbf{x})\|_2\right\} \leq B'_{\mathrm{flow}}R, \quad 0 \leq t \leq s \leq 1,\ \mathbf{x} \in \mathbb{B}_R^\infty.$$

Error Analysis for Velocity Matching

Time-average L^2-risk:

$$\mathcal{E}_T(\mathbf{b}) = \frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{X}_t}\left[\|\mathbf{b}(t,\mathbf{X}_t) - \mathbf{b}^*(t,\mathbf{X}_t)\|_2^2\right]\mathrm{d}t$$

Time-average truncated L^2-risk:

$$\mathcal{E}_{T,R}(\mathbf{b}) = \frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{X}_t}\left[\|\mathbf{b}(t,\mathbf{X}_t) - \mathbf{b}^*(t,\mathbf{X}_t)\|_2^2 \mathbb{1}\{\|\mathbf{X}_t\|_\infty \leq R\}\right]\mathrm{d}t$$

Oracle inequality of velocity matching (Ding-Duan-Jiao-Li-Yang-Zhang)
Under Assumptions **A1** and **A2**. Consider the linear interpolant $\alpha_t = 1 - t$ and $\beta_t = t$.

$$\mathbb{E}[\mathcal{E}_T(\mathbf{b})] \lesssim \underbrace{\inf_{b \in \mathscr{B}} \mathcal{E}_{T,R}(\mathbf{b})}_{\text{approximation}} + \underbrace{R^2 \max_{1 \leq k \leq d} \frac{\mathrm{VCdim}(\Pi_k \mathscr{B})}{n \log^{-1}(n)}}_{\text{generallization}} + \underbrace{R^2 \exp(-\theta R^2)}_{\text{truncation}}$$

Error Analysis for Velocity Matching

Let $T \in (1/2, 1)$ and $R \in (1, +\infty)$. Set the hypothesis class $\mathscr{B}$ as

$$\mathscr{B} = \left\{ \mathbf{b} \in N(L,S) : \begin{array}{l} \|\mathbf{b}(t,\mathbf{x})\|_\infty \leq B_{\mathrm{vel}} R, \ \|\partial_t \mathbf{b}(t,\mathbf{x})\|_2 \leq 3D \sup_{t\in[0,T]} \left(\frac{\dot{\alpha}_t^2}{\alpha_t^2} + \frac{|\ddot{\alpha}_t|}{\alpha_t} \right) R, \\ \|\nabla \mathbf{b}(t,\mathbf{x})\|_{\mathrm{op}} \leq 3G, \ (t,\mathbf{x}) \in [0,T] \times \mathbb{R}^d \end{array} \right\},$$

where the depth and the width of the neural network are given, respectively, as $L = C$ and $S = CN^{d+1}$. Then the following inequalities holds for each $N \in \mathbb{N}_+$,

- approximation error

$$\inf_{b \in \mathscr{B}} \mathcal{E}_{T,R}(\mathbf{b}) \lesssim \sup_{t\in[0,T]} \left(\frac{\dot{\alpha}_t^2}{\alpha_t^2} + \frac{|\ddot{\alpha}_t|}{\alpha_t} \right)^2 R^2 N^{-2}.$$

- generalization error

$$\mathrm{VCdim}(\Pi_k \mathscr{B}) \lesssim N^{d+1} \log N.$$

Error Analysis for Velocity Matching

Convergence rate of velocity matching (Ding-Duan-Jiao-Li-Yang-Zhang)
Under Assumptions **A1** and **A2**. Let $T \in (1/2, 1)$. Set the hypothesis class $\mathscr{B}$ as

$$\mathscr{B} = \left\{ \mathbf{b} \in N(L,S) : \begin{array}{c} \|\mathbf{b}(t,\mathbf{x})\|_\infty \le B_{\mathrm{vel}} \log^{1/2} n, \\ \|\partial_t \mathbf{b}(t,\mathbf{x})\|_2 \le 3D \sup_{t\in[0,T]} \left(\frac{\dot{\alpha}_t^2}{\alpha_t^2} + \frac{|\ddot{\alpha}_t|}{\alpha_t} \right) \log^{1/2} n, \\ \|\nabla \mathbf{b}(t,\mathbf{x})\|_{\mathrm{op}} \le 3G, \ (t,\mathbf{x}) \in [0,T]\times\mathbb{R}^d \end{array} \right\},$$

where the depth and the width of the neural network are given, respectively, as $L = C$ and $S = Cn^{\frac{d+1}{d+3}}$. Then

$$\mathbb{E}[\mathcal{E}_T(\mathbf{b})] \lesssim \sup_{t\in[0,T]} \left(\frac{\dot{\alpha}_t^2}{\alpha_t^2} + \frac{|\ddot{\alpha}_t|}{\alpha_t} \right)^2 n^{-\frac{2}{d+3}} \log^2 n.$$

Error Analysis for Characteristic Learning

Convergence rate of characteristic learning (Ding-Duan-Jiao-Li-Yang-Zhang)
Under Assumptions **A1** and **A2**. Consider the linear interpolant $\alpha_t = 1 - t$ and $\beta_t = t$.

- Set the stopping time T as $T = 1 - Cn^{-\frac{1}{3(d+3)}} \log^{\frac{1}{2}} n$.
- The number of iterations of Euler sampling satisfies $K \geq Cn^{\frac{1}{d+3}}$.
- The depth $L_{\mathscr{B}}$ and the number of non-zero parameters $S_{\mathscr{B}}$ of the velocity network class $\mathscr{B}$ are set, respectively, as $L_{\mathscr{B}} = C$ and $S_{\mathscr{B}} = Cn^{\frac{(d+1)(3d+7)}{3(d+3)^2}}$.
- The depth $L_{\mathscr{G}}$ and the number of non-zero parameters $S_{\mathscr{G}}$ of the characteristic network class $\mathscr{G}$ are set, respectively, as $L_{\mathscr{G}} = C$ and $S_{\mathscr{G}} = Cm^{\frac{d+2}{d+4}}$.

Then the time-average squared 2-Wasserstein error can be bounded as:

$$\mathbb{E}\left[\frac{2}{T^2}\int_0^T\int_t^T W_2^2((\widehat{\mathbf{g}}_{t,s})_\sharp\mu_t,\mu_s)\mathrm{d}s\mathrm{d}t\right]$$
$$\lesssim n^{-\frac{2}{3(d+3)}}\log^2 n + m^{-\frac{2}{d+4}}\log^2 m + \frac{\max\{\log n, \log m\}}{K}.$$

Mitigate the Curse of Dimensionality

- **A3.** There exists an unknown constant $\sigma > 0$ and $d^* \ll d$, such that

$$\mu_1 = N(\mathbf{0}, \sigma^2 \mathbf{I}_d) * (\mathbf{P}_\sharp \tilde{\nu}),$$

where $\mathbf{P} \in \mathbb{R}^{d \times d^*}$ is a matrix whose column vectors are orthonormal in $\mathbb{R}^d$, and $\tilde{\nu}$ is a distribution with $\mathrm{supp}(\tilde{\nu}) \subseteq [0,1]^{d^*}$.

Mitigate the Curse of Dimensionality

Low-dimensional decomposition

For each $(t, s, \mathbf{x})$ with $0 \leq t \leq s \leq 1$ and $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbf{b}^*(t, \mathbf{x}) \equiv \mathbf{P}\,\widetilde{\mathbf{b}}^*(t, \mathbf{P}^\top \mathbf{x}) + \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2} (\mathbf{I}_d - \mathbf{P}\mathbf{P}^\top)\mathbf{x},$$

$$\mathbf{g}^*(t, s, \mathbf{x}) \equiv \mathbf{P}\,\widetilde{\mathbf{g}}^*(t, s, \mathbf{P}^\top \mathbf{x}) + \sqrt{\frac{\alpha_s^2 + \sigma^2 \beta_s^2}{\alpha_t^2 + \sigma^2 \beta_t^2}} (\mathbf{I}_d - \mathbf{P}\mathbf{P}^\top)\mathbf{x}.$$

The vector field $\widetilde{\mathbf{b}}^* : \mathbb{R} \times \mathbb{R}^{d^*} \to \mathbb{R}^{d^*}$ is defined as:

$$\widetilde{\mathbf{b}}^*(t, \widetilde{\mathbf{x}}) := \mathbb{E}[\dot{\alpha}_t \widetilde{\mathbf{X}}_0 + \dot{\beta}_t \widetilde{\mathbf{X}}_1 | \widetilde{\mathbf{X}}_t = \widetilde{\mathbf{x}}], \quad \widetilde{\mathbf{x}} \in \mathbb{R}^{d^*},$$

where $\widetilde{\mathbf{X}}_0 \sim N(\mathbf{0}, \mathbf{I}_{d^*})$ and $\widetilde{X}_1 \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{d^*}) * \tilde{\nu}$ are independent, and $\widetilde{\mathbf{X}}_t := \alpha_t \widetilde{\mathbf{X}}_0 + \beta_t \widetilde{\mathbf{X}}_1$. Further, the vector field $\widetilde{\mathbf{g}}^* : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d^*} \to \mathbb{R}^{d^*}$ is defined as the flow map of $\mathrm{d}\widetilde{\mathbf{x}}(t) = \widetilde{\mathbf{b}}^*(t, \widetilde{\mathbf{x}}(t))\mathrm{d}t$.

Mitigate the Curse of Dimensionality

Convergence rate under manifold assumption (Ding-Duan-Jiao-Li-Yang-Zhang)
Under Assumptions **A1** and **A3**. Consider the linear interpolant $\alpha_t = 1 - t$ and $\beta_t = t$.

- Set the stopping time T as $T = 1 - Cn^{-\frac{1}{3(d^*+3)}} \log^{\frac{1}{2}} n$.
- The number of iterations of Euler sampling satisfies $K \geq Cn^{\frac{1}{d^*+3}}$.
- The depth $L_{\mathscr{B}}$ and the number of non-zero parameters $S_{\mathscr{B}}$ of the velocity network class $\mathscr{B}$ are set, respectively, as $L_{\mathscr{B}} = C$ and $S_{\mathscr{B}} = Cn^{\frac{(d^*+1)(3d^*+7)}{3(d^*+3)^2}}$.
- The depth $L_{\mathscr{G}}$ and the number of non-zero parameters $S_{\mathscr{G}}$ of the characteristic network class $\mathscr{G}$ are set, respectively, as $L_{\mathscr{G}} = C$ and $S_{\mathscr{G}} = Cm^{\frac{d^*+2}{d^*+4}}$.

Then the time-average squared 2-Wasserstein error can be bounded as:

$$\mathbb{E}\Big[\frac{2}{T^2}\int_0^T\int_t^T W_2^2((\widehat{\mathbf{g}}_{t,s})_\sharp\mu_t,\mu_s)\mathrm{d}s\mathrm{d}t\Big] \lesssim n^{-\frac{2}{3(d^*+3)}}\log^2 n + m^{-\frac{2}{d^*+4}}\log^2 m + \frac{\max\{\log n, \log m\}}{K}.$$

Outline

Performance Comparisons on CIFAR-10

Model	NFE ↓	FID ↓
Diffusion + Sampler		
DDPM	1000	3.17
DDIM	100	4.16
Score SDE	2000	2.20
EDM	35	2.01
Diffusion + Distillation		
KD	1	9.36
DFNO	1	5.92
Rectified Flow	1	4.85
PD	1	9.12
CD	1	10.53
CTM (without GAN)	1	5.19
CG (ours)	1	**4.59**
PD	2	4.51
CTM (without GAN)	18	3.00
CG (ours)	2	**3.50**
CG (ours)	4	**2.83**

High-Resolution Images Generation

Conclusions

- End-to-end error analysis for flow-based generative models and distilled model.

[1] Zhao Ding, Chenguang Duan, Yuling Jiao, Ruoxuan Li, Jerry Zhijian Yang, and Pingwen Zhang. Characteristic Learning for Provable One Step Generation. arXiv:2405.05512