Школа анализа данных

Diffusion Distillation

Nikita Starodubcev

Yandex Research

Lecture 4 | ODE-Based Diffusion Distillation

01 Motivation

02 Knowledge Distillation

03 Consistency Distillation

04 Multi-Boundary Consistency Distillation

1. Motivation

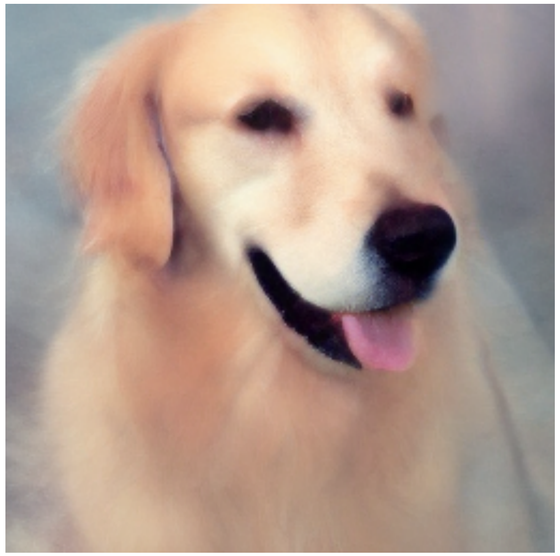
Diffusion Models are slow. Why?

30 steps

...

7 steps

5 steps



3 steps

1. PF-ODE, VE case

Remind the reverse PF-ODE equation

$$\mathrm{d}\boldsymbol{x} = \left(\boldsymbol{f}(\boldsymbol{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \right) \mathrm{d}t, t \in [0,1]$$

Variance Exploding case

$$\boldsymbol{f}(\boldsymbol{x}, t) = 0,\ g(t) = \sqrt{\frac{\mathrm{d}\sigma_t^2}{\mathrm{d}t}}$$

$$\mathrm{d}\boldsymbol{x} = -\sigma_t \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \mathrm{d}\sigma_t$$

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) = -\frac{1}{\sigma_t^2} \left(\boldsymbol{x} - \mathbb{E}\ \boldsymbol{x}_0 | \boldsymbol{x} \right)$$

1. Diffusion loss

Remind the diffusion loss function

$$\mathrm{L}(\boldsymbol{f}_\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_t(\boldsymbol{x}|\boldsymbol{x}_0)}\|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 \to \min_\theta$$

$\mathcal{U}(0,1)$ – Uniform distribution of time

$p_{data}(\boldsymbol{x}_0)$ – Data distribution (e.g., images)

$q_t(\boldsymbol{x}\,|\,\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{x}_0, \sigma_t^2)$ – Transition kernels (how to noise the data)

1. Why Diffusion Models are slow?

Potential sources of low speed?

1. Low capacity of a neural network that approximates $\mathbb{E}\left[\boldsymbol{x}_0 \,|\, \boldsymbol{x}\right]$?

2. Maybe we need a solver with lower discretization error?

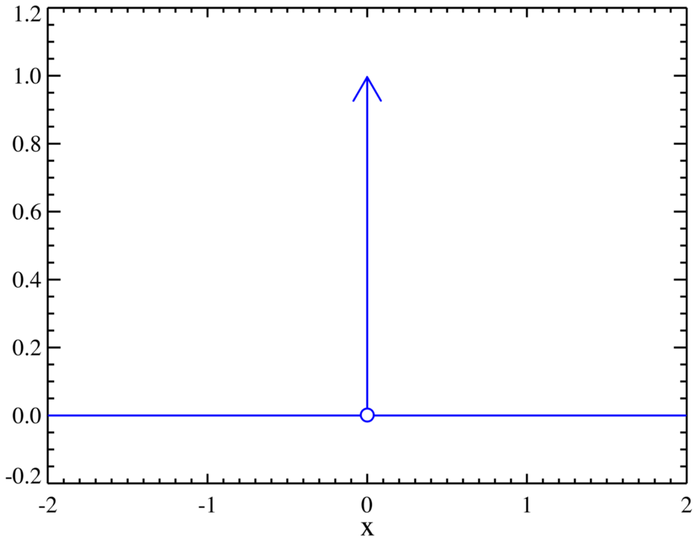
1. Towards ideal approximation

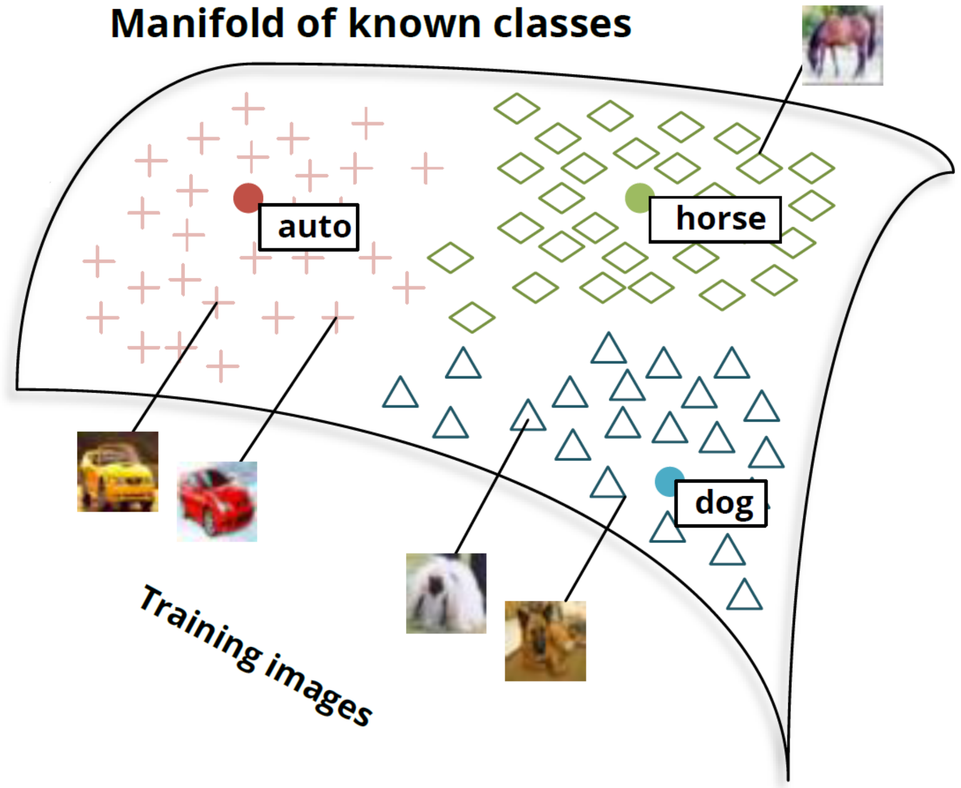
What if our neural network perfectly approximates $\mathbb{E}\left[\boldsymbol{x}_0 \,|\, \boldsymbol{x}\right]$?

$$\mathrm{L}(\boldsymbol{f}_\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_t(\boldsymbol{x}|\boldsymbol{x}_0)}\|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 - \text{optimum?}$$

Assume that

$$p_{data}(\boldsymbol{x}_0) = \frac{1}{N}\sum_{j=1}^{N}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j);\ \{x_0^j\}_{j=1}^N - \text{our dataset}$$





Expand the expectations:

$$\mathrm{L}(\boldsymbol{f}_\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_t(\boldsymbol{x}|\boldsymbol{x}_0)}\|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

$$= \int_0^1 \mathrm{d}t \; \mathcal{U}(t\,|\,0,1)$$

Expand the expectations:

$$\mathrm{L}(\boldsymbol{f}_\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_t(\boldsymbol{x}|\boldsymbol{x}_0)}\|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

$$= \int_0^1 \mathrm{d}t\ \mathcal{U}(t\,|\,0,1) \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x}_0\ \frac{1}{N}\sum_{j=1}^{N} \delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j)$$

Expand the expectations:

$$\mathrm{L}(\boldsymbol{f}_\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_t(\boldsymbol{x}|\boldsymbol{x}_0)}\|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

$$= \int_0^1 \mathrm{d}t\ \mathcal{U}(t\,|\,0,1) \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x}_0\ \frac{1}{N}\sum_{j=1}^{N} \delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j) \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x}\ \mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0, \sigma_t^2\right) \|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

Expand the expectations:

$$\mathrm{L}(\boldsymbol{f}_\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_t(\boldsymbol{x}|\boldsymbol{x}_0)}\|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

$$= \int_0^1 \mathrm{d}t\ \mathcal{U}(t\,|\,0,1)\frac{1}{N}\sum_{j=1}^{N}\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}_0\ \delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j)\left[\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}\ \mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0, \sigma_t^2\right)\|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2\right] =$$

We can use this property $\int_{\mathbb{R}^n}\mathrm{d}x\ \delta(x-y)g(x) = g(y)$

Expand the expectations:

$$\mathrm{L}(\boldsymbol{f}_\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_t(\boldsymbol{x}|\boldsymbol{x}_0)}\|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

$$= \int_0^1 \mathrm{d}t\ \mathcal{U}(t\,|\,0,1)\frac{1}{N}\sum_{j=1}^{N}\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}_0\ \delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j)\left[\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}\ \mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0, \sigma_t^2\right)\|\boldsymbol{x}_0 - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2\right] =$$

We can use this property $\int_{\mathbb{R}^n}\mathrm{d}x\ \delta(x-y)g(x) = g(y)$

$$= \int_0^1 \mathrm{d}t\ \mathcal{U}(t\,|\,0,1)\frac{1}{N}\sum_{j=1}^{N}\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}\ \mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)\|\boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

$$= \int_0^1 \mathrm{d}t \; \mathcal{U}(t \mid 0,1) \frac{1}{N} \sum_{j=1}^{N} \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x} \; \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{x}_0^j, \sigma_t^2\right) \left\| \boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t) \right\|_2^2 =$$

$$= \int_0^1 \mathrm{d}t \int_{\mathbb{R}^n} \frac{1}{N} \sum_{j=1}^{N} \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{x}_0^j, \sigma_t^2\right) \left\| \boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t) \right\|_2^2 \mathrm{d}\boldsymbol{x} =$$

$$= \int_0^1 \mathrm{d}t \int_{\mathbb{R}^n} \mathrm{J}(\boldsymbol{f}_\theta, \boldsymbol{x}, t) \mathrm{d}\boldsymbol{x}$$

We can minimize $\mathrm{L}(\boldsymbol{f}_\theta)$ by minimizing $\mathrm{J}(\boldsymbol{f}_\theta, \boldsymbol{x}, t)$ independently for each $\boldsymbol{x}, t$

$$\nabla_{f_\theta} \mathrm{J}(f_\theta) = 0$$

$$\nabla_{f_\theta} \mathrm{J}(\boldsymbol{f}_\theta) = 0$$

$$\nabla_{\boldsymbol{f}_\theta} \left(\frac{1}{N} \sum_{j=1}^{N} \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{x}_0^j, \sigma_t^2 \right) \| \boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t) \|_2^2 \right) =$$

$$\nabla_{f_\theta} \mathrm{J}(\boldsymbol{f}_\theta) = 0$$

$$\nabla_{\boldsymbol{f}_\theta}\left(\frac{1}{N}\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)\|\boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2\right) = \frac{1}{N}\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)\nabla_{\boldsymbol{f}_\theta}\|\boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

$$\nabla_{f_\theta} \mathrm{J}(\boldsymbol{f}_\theta) = 0$$

$$\nabla_{\boldsymbol{f}_\theta}\left(\frac{1}{N}\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)\|\boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2\right) = \frac{1}{N}\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)\nabla_{\boldsymbol{f}_\theta}\|\boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

$$= -\frac{2}{N}\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)(\boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t)) = 0$$

$$\nabla_{f_\theta} \mathrm{J}(\boldsymbol{f}_\theta) = 0$$

$$\nabla_{\boldsymbol{f}_\theta}\left(\frac{1}{N}\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)\|\boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2\right) = \frac{1}{N}\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)\nabla_{\boldsymbol{f}_\theta}\|\boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t)\|_2^2 =$$

$$= -\frac{2}{N}\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)(\boldsymbol{x}_0^j - \boldsymbol{f}_\theta(\boldsymbol{x}, t)) = 0$$

$$\boldsymbol{f}_\theta(\boldsymbol{x}, t) = \frac{\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)\boldsymbol{x}_0^j}{\sum_{j=1}^{N}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)}$$

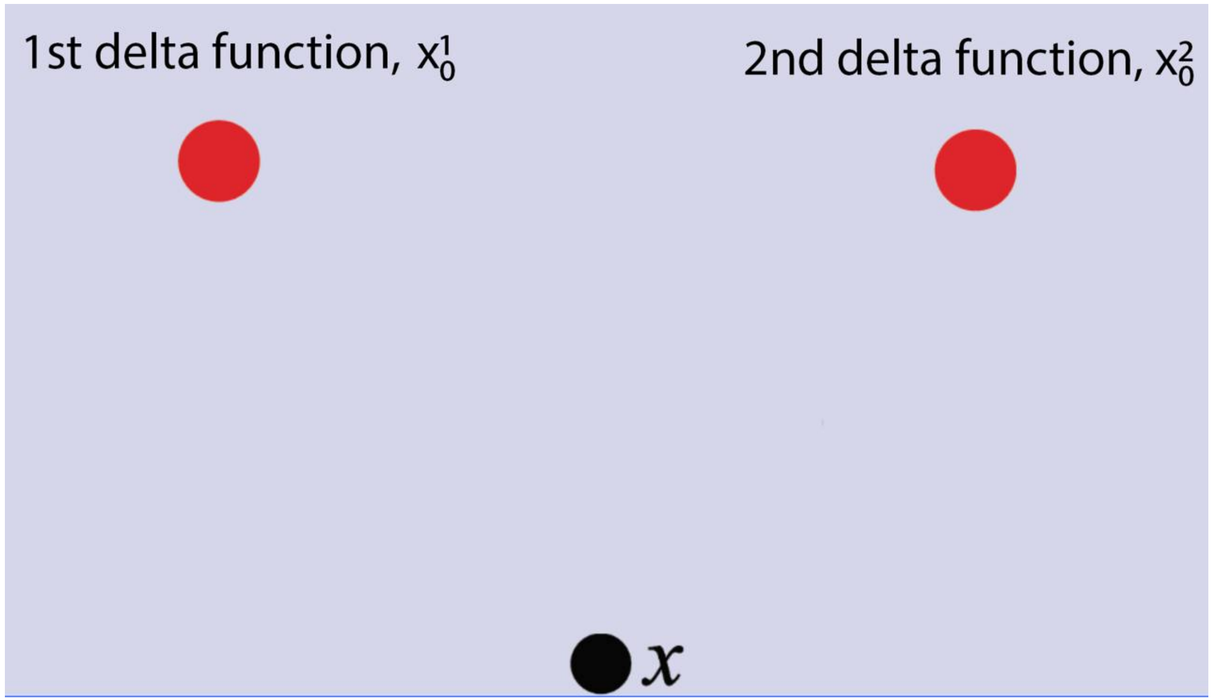
1. Ideal diffusion model

The form of the conditional expectation

$$\boldsymbol{f}_\theta(\boldsymbol{x}, t) = \frac{\sum_{j=1}^{N} \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{x}_0^j, \sigma_t^2\right) \boldsymbol{x}_0^j}{\sum_{j=1}^{N} \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{x}_0^j, \sigma_t^2\right)}$$

What if we generate data using this score? Let's consider two delta functions

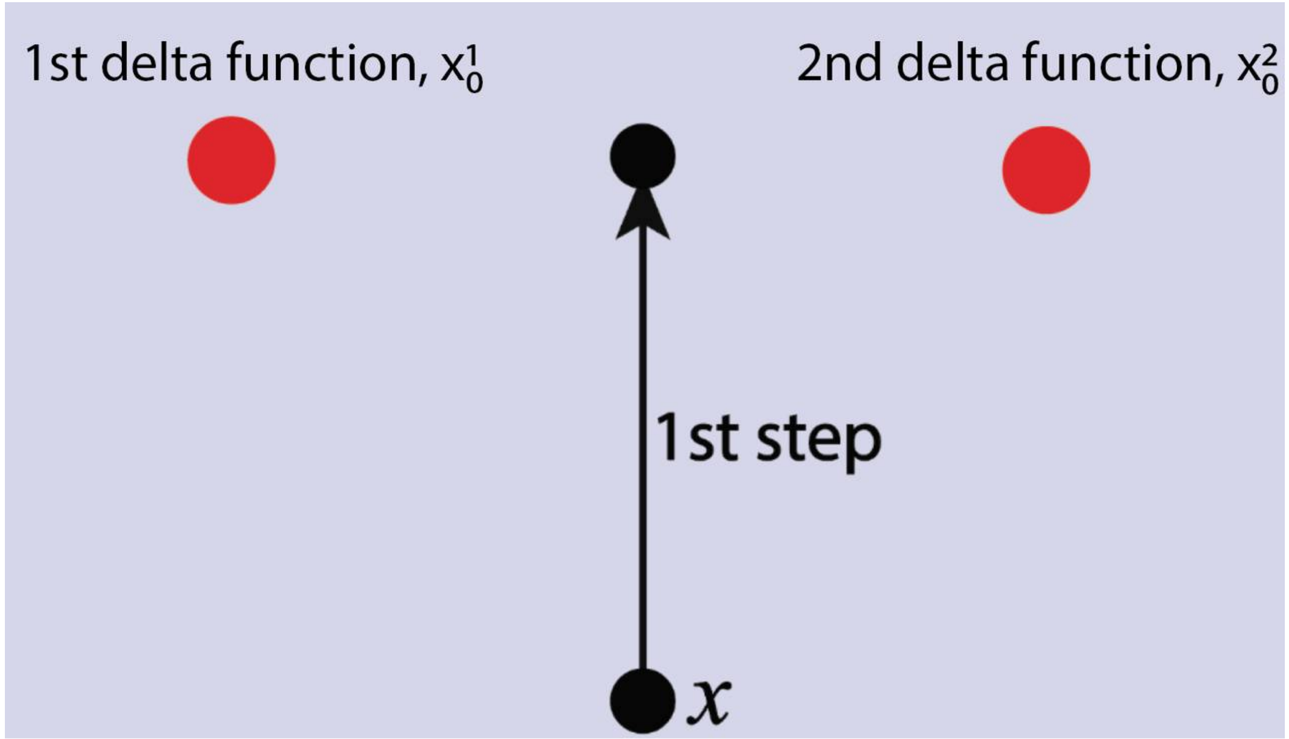
$$\boldsymbol{f}_\theta(\boldsymbol{x}, t) = \frac{\mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{x}_0^1, \sigma_t^2\right) \boldsymbol{x}_0^1 + \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{x}_0^2, \sigma_t^2\right) \boldsymbol{x}_0^2}{\mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{x}_0^1, \sigma_t^2\right) + \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{x}_0^2, \sigma_t^2\right)}$$

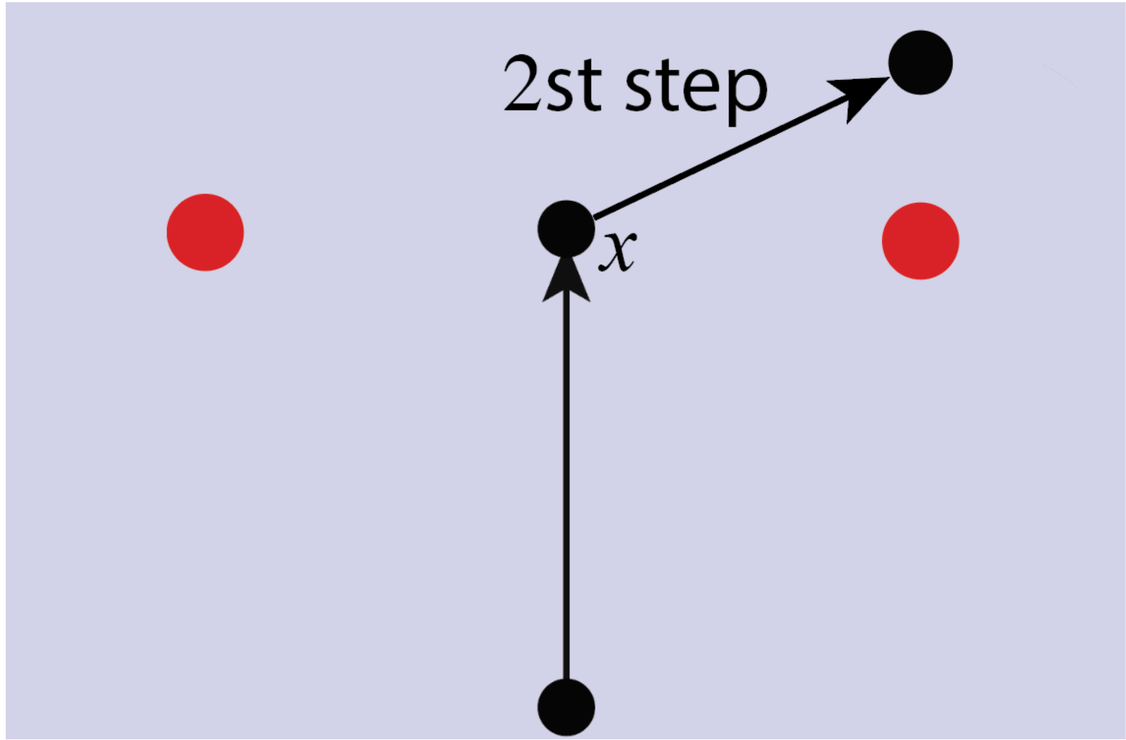


1. Sampling from ideal diffusion model

$$f_\theta(\boldsymbol{x}) = \frac{0.5\boldsymbol{x}_0^1 + 0.5\boldsymbol{x}_0^2}{0.5 + 0.5} = \frac{\boldsymbol{x}_0^1 + \boldsymbol{x}_0^2}{2}$$

$$f_\theta(\boldsymbol{x}) = 0.3\boldsymbol{x}_0^1 + 0.7\boldsymbol{x}_0^2$$

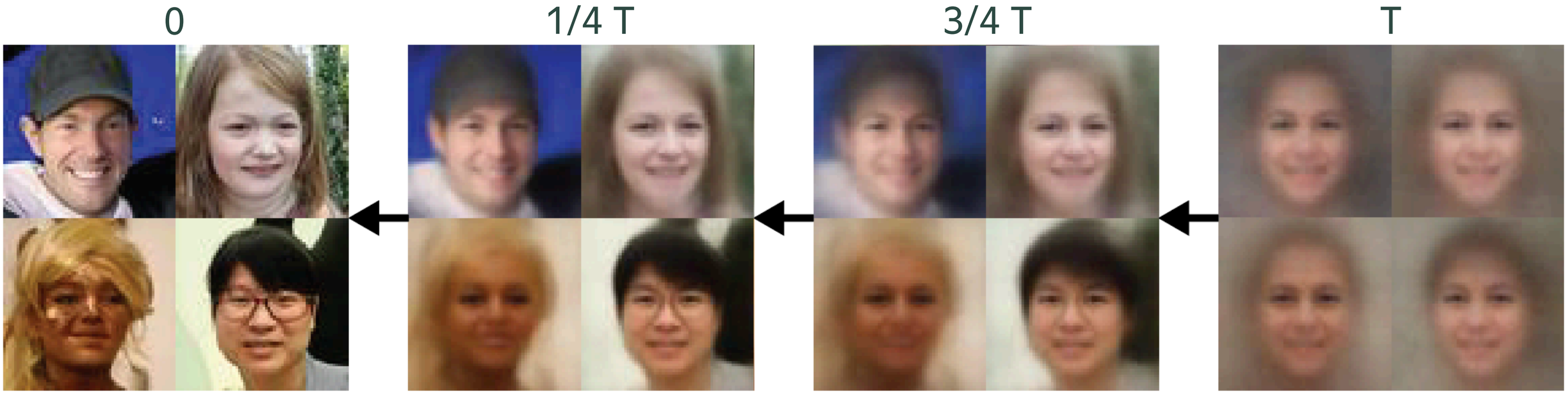




Even in the simplest case, DM is not suitable for obtaining samples in a few steps

1. Sampling from ideal diffusion model

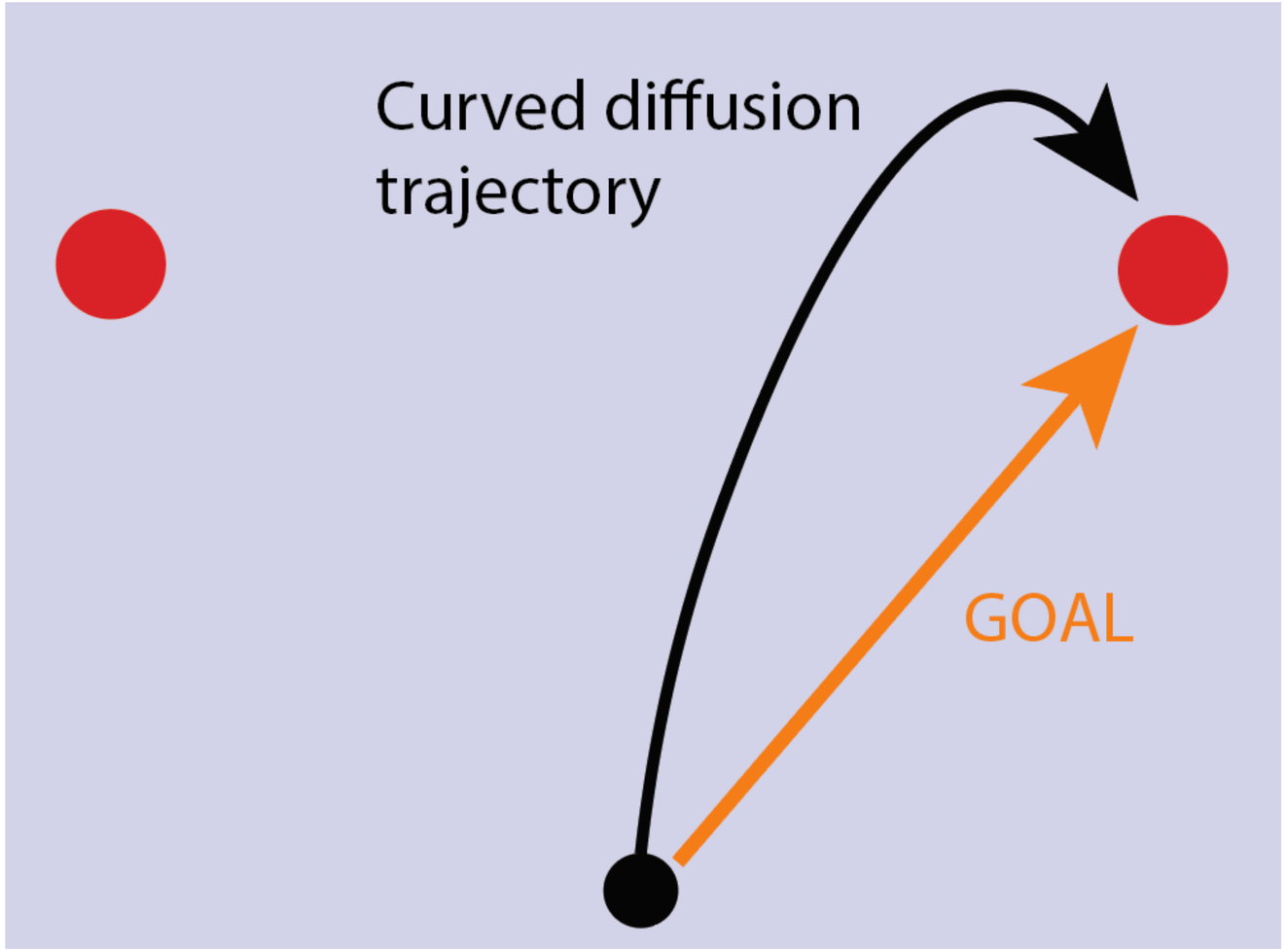
What about images?



Images become less and less averaged

1. DMs speed are limited!

Diffusion has curved trajectories



2. Moving towards fast DMs

Why DMs are slow? → The optimal solution is an averaged sample → Curved trajectories

No fast sampling with DMs :(But we want it!

1. Flow based approaches
Generalized PF-ODEs to obtain more straightened trajectories

- Flow matching
- Rectified flows
- …

2. Distillation (diffusion → few-step model)
Reformulate the problem to obtain

$$f_\phi(x, t) = x_0$$

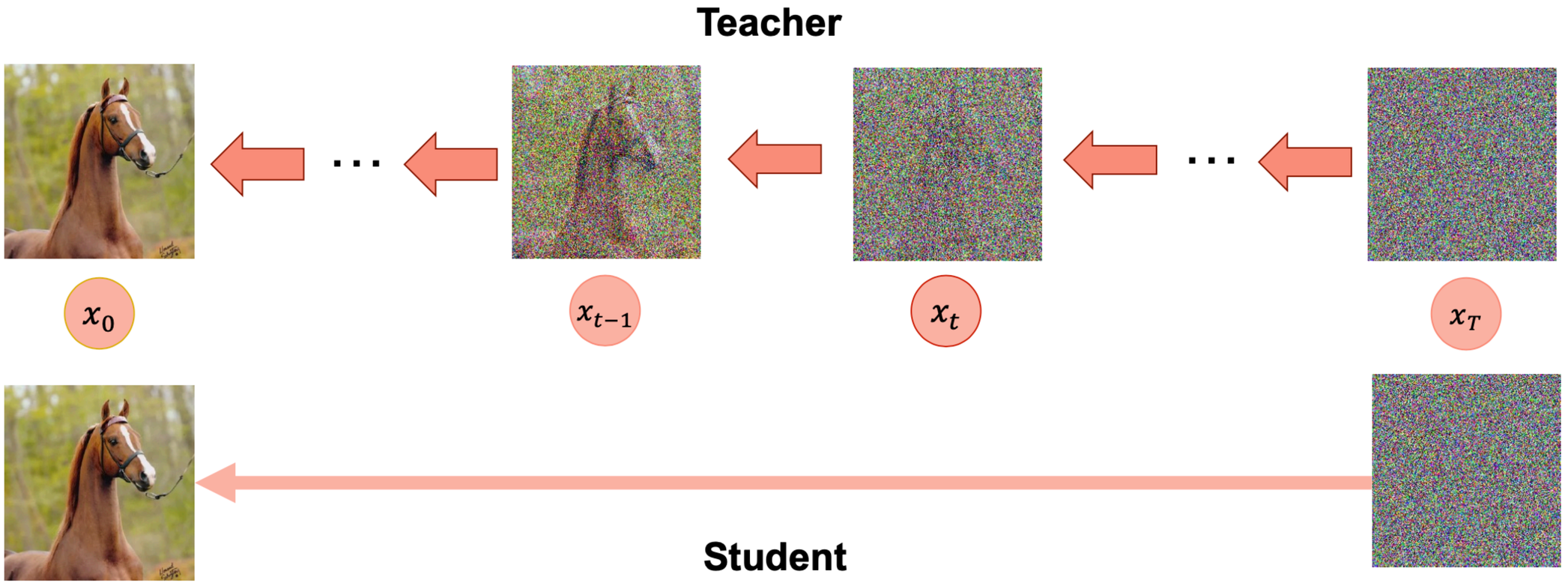
ODE-based

- Knowledge
- Consistency
- Progressive
- …

ODE-free

- DMD
- Adversarial
- …

2. Knowledge distillation



The main idea is simple:

- collect the data using diffusion, $\{(\boldsymbol{x}_T^1, \boldsymbol{x}_0^1), \ldots, (\boldsymbol{x}_T^N, \boldsymbol{x}_0^N)\}$
- fine-tune the network, $\boldsymbol{f}_\phi(\boldsymbol{x}, t)$, on these data, $\boldsymbol{f}_\phi(\boldsymbol{x}_T) \rightarrow \boldsymbol{x}_0$

2. Knowledge distillation

$$\mathrm{L}_{KD}(\boldsymbol{f}_{\phi}) = \mathbb{E}_{\mathcal{N}(\boldsymbol{x}|0,\sigma^2)}\|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_{\phi}(\boldsymbol{x},t)\|_2^2$$

$\hat{\boldsymbol{x}}_0(x,\theta)-$ sample from the pretrained DM using solver

Looks quite similar to the diffusion loss

$$\mathrm{L}_{diffusion}(\boldsymbol{f}_{\theta}) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_t(\boldsymbol{x}|\boldsymbol{x}_0)}\|\boldsymbol{x}_0 - \boldsymbol{f}_{\theta}(\boldsymbol{x},t)\|_2^2$$

But the optimum is different!

2. Knowledge distillation. Optimal solution

$$\nabla_{f_\phi} \mathrm{L}_{KD}(\boldsymbol{f}_\phi) = -\ ?$$

$$\mathrm{L}_{KD}(\boldsymbol{f}_\phi) = \mathbb{E}_{\mathcal{N}(\boldsymbol{x}|0,\sigma^2)} \|\hat{\boldsymbol{x}}_0(\boldsymbol{x}, \theta) - \boldsymbol{f}_\phi(\boldsymbol{x}, t)\|_2^2 \ =$$

2. Knowledge distillation. Optimal solution

$$\nabla_{f_\phi} \mathrm{L}_{KD}(\boldsymbol{f}_\phi) = -\,?$$

$$\mathrm{L}_{KD}(\boldsymbol{f}_\phi) = \mathbb{E}_{\mathcal{N}(\boldsymbol{x}|0,\sigma^2)} \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 \;= \int_{\mathbb{R}^n} \mathcal{N}(\boldsymbol{x}\,|\,0,\sigma^2)\; \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 \mathrm{d}\boldsymbol{x},$$

2. Knowledge distillation. Optimal solution

$$\nabla_{f_\phi} \mathrm{L}_{KD}(\boldsymbol{f}_\phi) = -\,?$$

$$\mathrm{L}_{KD}(\boldsymbol{f}_\phi) = \mathbb{E}_{\mathcal{N}(\boldsymbol{x}|0,\sigma^2)} \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 \;= \int_{\mathbb{R}^n} \mathcal{N}(\boldsymbol{x}\,|\,0,\sigma^2)\; \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 \mathrm{d}\boldsymbol{x},$$

$$\int_{\mathbb{R}^n} \mathcal{N}(\boldsymbol{x}\,|\,0,\sigma^2)\; \nabla_{f_\phi} \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 \mathrm{d}\boldsymbol{x} = 0 \rightarrow \nabla_{f_\phi} \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 = 0$$

2. Knowledge distillation. Optimal solution

$$\nabla_{f_\phi} \mathrm{L}_{KD}(\boldsymbol{f}_\phi) = -\,?$$

$$\mathrm{L}_{KD}(\boldsymbol{f}_\phi) = \mathbb{E}_{\mathcal{N}(\boldsymbol{x}|0,\sigma^2)} \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 = \int_{\mathbb{R}^n} \mathcal{N}(\boldsymbol{x}\,|\,0,\sigma^2)\ \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 \mathrm{d}\boldsymbol{x},$$

$$\int_{\mathbb{R}^n} \mathcal{N}(\boldsymbol{x}\,|\,0,\sigma^2)\ \nabla_{f_\phi} \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 \mathrm{d}\boldsymbol{x} = 0 \rightarrow \nabla_{f_\phi} \|\hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta) - \boldsymbol{f}_\phi(\boldsymbol{x},t)\|_2^2 = 0$$

$$\boldsymbol{f}_\phi(\boldsymbol{x},t) = \hat{\boldsymbol{x}}_0(\boldsymbol{x},\theta)$$

$$\boldsymbol{f}_\theta(\boldsymbol{x},t) = \frac{\sum_{j=1}^{N} \mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right) \boldsymbol{x}_0^j}{\sum_{j=1}^{N} \mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{x}_0^j, \sigma_t^2\right)}$$

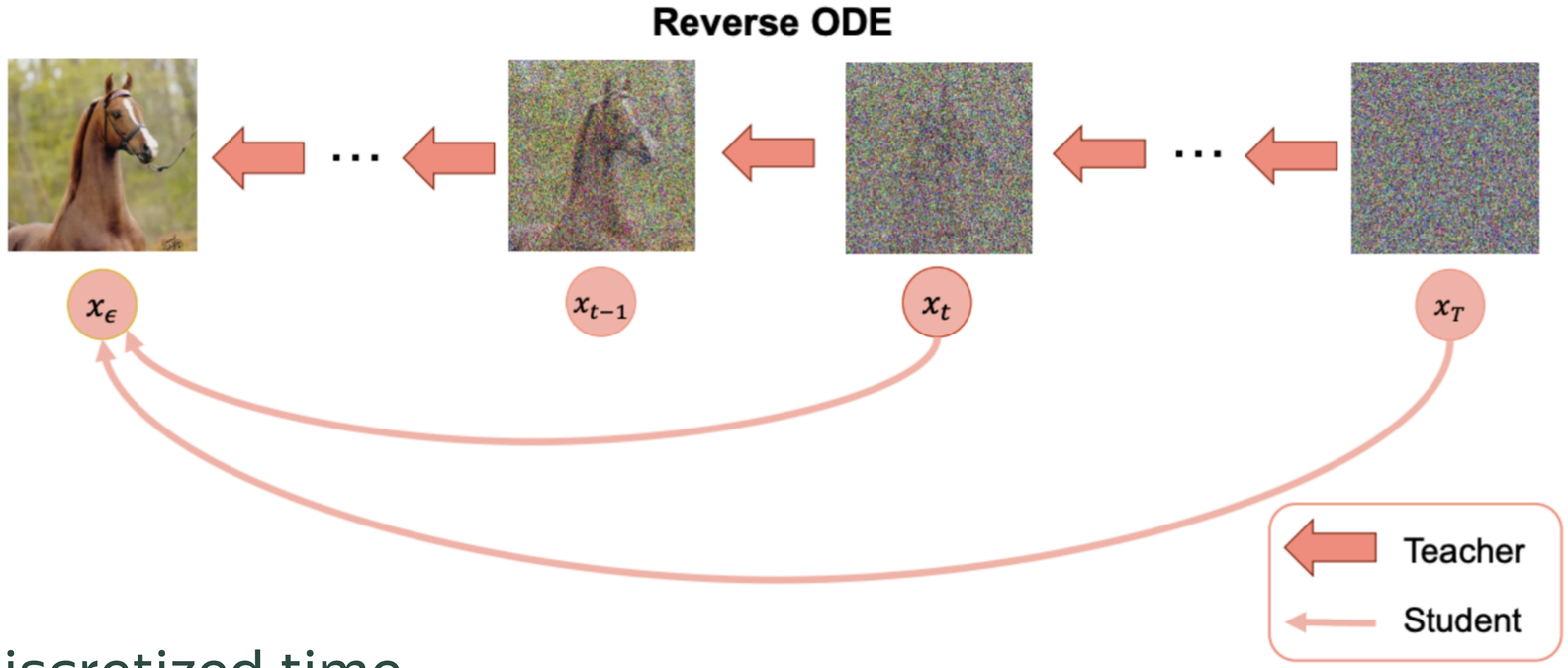
Distillation general principle

The form of different distillations may be different, but the principle remains the same.

The network should predict $\boldsymbol{x}_0$ instead of an averaged sample!

3. Consistency Distillation. Main idea

We have to sample from DM in the KD. Consistency Distillation aims to avoid this.



$\{t_1, \ldots, t_N\}$ – discretized time

$\boldsymbol{f}_\phi(\boldsymbol{x}_t, t) \rightarrow \boldsymbol{x}_0, \forall x_t$ from the same ODE trajectory

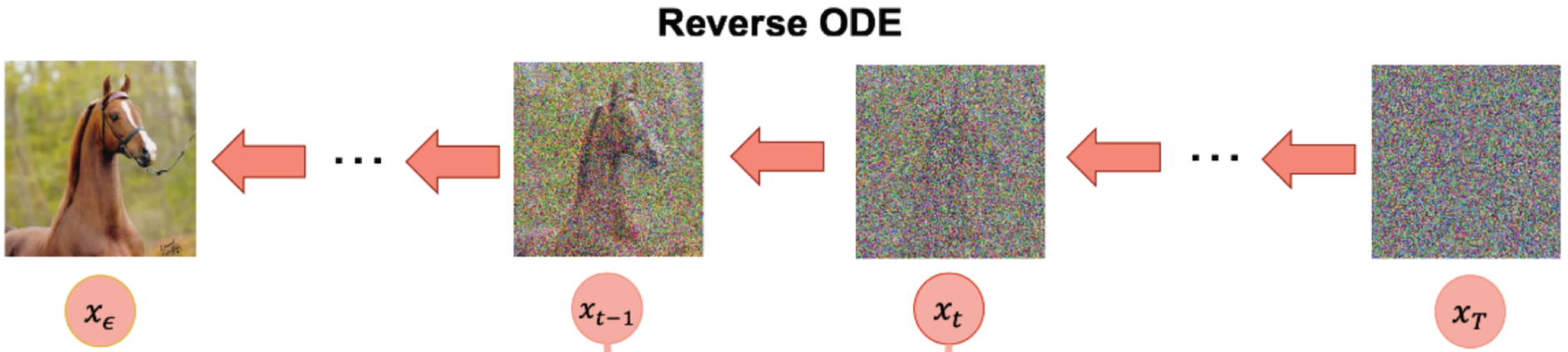
$$\boldsymbol{f}_\phi(\boldsymbol{x}_1, t_1) = \boldsymbol{f}_\phi(\boldsymbol{x}_2, t_2) = \ldots = \boldsymbol{f}_\phi(\boldsymbol{x}_N, t_N) = \boldsymbol{x}_0$$

3. Consistency Distillation. Main idea

To obtain such a model, enforce the self-consistency property

$$\|\boldsymbol{f}_\phi(\boldsymbol{x}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_\phi$$

where $\boldsymbol{x}_{n-1}, \boldsymbol{x}_n -$ adjacent points on the ODE trajectory



$$\|\boldsymbol{f}_\phi(\boldsymbol{x}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_\phi$$

3. Consistency Distillation. Loss function

How to obtain the adjacent points? $x_n \sim p_{t_n}(x_n \mid x_0);\ \ \hat{x}_{n-1} = Solver(x_n, t_n, t_{n-1} \mid \theta)$

$$\mathbb{E}_{n\sim\mathcal{U}(2,N)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_{t_n}(\boldsymbol{x}_n|\boldsymbol{x}_0)}\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_\phi$$

$\mathcal{U}(2,N)$ – Discrete uniform distribution of time points

$p_{data}(\boldsymbol{x}_0)$ – Data distribution (e.g., images)

$q_{t_n}(\boldsymbol{x} \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{x}_0, \sigma_{t_n}^2)$ – Transition kernels (how to noise the data)

3. What about optimal solution?

$$\mathbb{E}_{n\sim\mathcal{U}(2,N)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_{t_n}(\boldsymbol{x}_n|\boldsymbol{x}_0)}\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_{\boldsymbol{f}_\phi}$$

$$p_{data}(\boldsymbol{x}_0) = \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j);\ \{x_0^j\}_{j=1}^M - \text{our dataset}$$

3. What about optimal solution?

$$\mathbb{E}_{n\sim\mathcal{U}(2,N)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_{t_n}(\boldsymbol{x}_n|\boldsymbol{x}_0)}\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_{\boldsymbol{f}_\phi}$$

$$p_{data}(\boldsymbol{x}_0) = \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j);\ \{x_0^j\}_{j=1}^M - \text{our dataset}$$

$$\sum_{n=2}^{N}\frac{1}{N-2}$$

3. What about optimal solution?

$$\mathbb{E}_{n\sim\mathcal{U}(2,N)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_{t_n}(\boldsymbol{x}_n|\boldsymbol{x}_0)}\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_{\boldsymbol{f}_\phi}$$

$$p_{data}(\boldsymbol{x}_0) = \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j);\ \{x_0^j\}_{j=1}^M - \text{our dataset}$$

$$\sum_{n=2}^{N}\frac{1}{N-2}\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}_0\ \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j)$$

3. What about optimal solution?

$$\mathbb{E}_{n\sim\mathcal{U}(2,N)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_{t_n}(\boldsymbol{x}_n|\boldsymbol{x}_0)}\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_{\boldsymbol{f}_\phi}$$

$$p_{data}(\boldsymbol{x}_0) = \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j);\ \{x_0^j\}_{j=1}^M - \text{our dataset}$$

$$\sum_{n=2}^{N}\frac{1}{N-2}\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}_0\ \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j)\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}_n\ \mathcal{N}\left(\boldsymbol{x}_n\,|\,\boldsymbol{x}_0, \sigma_t^2\right)\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2$$

3. What about optimal solution?

$$\mathbb{E}_{n\sim\mathcal{U}(2,N)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_{t_n}(\boldsymbol{x}_n|\boldsymbol{x}_0)}\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_{\boldsymbol{f}_\phi}$$

$$p_{data}(\boldsymbol{x}_0) = \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j);\ \{x_0^j\}_{j=1}^M - \text{our dataset}$$

$$\sum_{n=2}^{N}\frac{1}{N-2}\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}_0\ \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j)\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}_n\ \mathcal{N}\left(\boldsymbol{x}_n\,|\,\boldsymbol{x}_0, \sigma_t^2\right)\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 =$$

$$= \sum_{n=2}^{N}\frac{1}{N-2}\int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{x}_n\ \frac{1}{M}\sum_{j=1}^{M}\mathcal{N}(\boldsymbol{x}_n\,|\,\boldsymbol{x}_0^j, \sigma_{t_n}^2)\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2$$

3. What about optimal solution?

$$\mathbb{E}_{n\sim\mathcal{U}(2,N)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_{t_n}(\boldsymbol{x}_n|\boldsymbol{x}_0)}\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_{\boldsymbol{f}_\phi}$$

$$p_{data}(\boldsymbol{x}_0) = \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j);\ \{x_0^j\}_{j=1}^M - \text{our dataset}$$

$$\sum_{n=2}^{N}\frac{1}{N-2}\int_{\mathbb{R}^n} d\boldsymbol{x}_0\ \frac{1}{M}\sum_{j=1}^{M}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^j)\int_{\mathbb{R}^n} d\boldsymbol{x}_n\ \mathcal{N}\left(\boldsymbol{x}_n\,|\,\boldsymbol{x}_0, \sigma_t^2\right)\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 =$$

$$= \sum_{n=2}^{N}\frac{1}{N-2}\int_{\mathbb{R}^n} d\boldsymbol{x}_n\ \underbrace{\frac{1}{M}\sum_{j=1}^{M}\mathcal{N}(\boldsymbol{x}_n\,|\,\boldsymbol{x}_0^j, \sigma_{t_n}^2)\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2}_{\mathrm{J}(\boldsymbol{f}_\phi, x_n, n)}$$

3. What about optimal solution?

$$\nabla_{f_\phi} \mathrm{J}(\boldsymbol{f}_\phi, x_n, n) = 0$$

$$\boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n) = \frac{\sum_{j=1}^{M} \left(\mathcal{N}(\boldsymbol{x}_n \,|\, \boldsymbol{x}_0^j, \sigma_{t_n}^2)\, \boldsymbol{f}_\phi(\hat{\mathbf{x}}_{n-1}, t_{n-1}) \right)}{\sum_{j=1}^{M} \mathcal{N}(\boldsymbol{x}_n \,|\, \boldsymbol{x}_0^j, \sigma_{t_n}^2)}$$

Hm, again a mixture? NO!

3. What about optimal solution?

$$\nabla_{\boldsymbol{f}_\phi} \mathrm{J}(\boldsymbol{f}_\phi, x_n, n) = 0$$

$$\boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n) = \frac{\sum_{j=1}^{M} \left(\mathcal{N}(\boldsymbol{x}_n \,|\, \boldsymbol{x}_0^j, \sigma_{t_n}^2)\, \boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) \right)}{\sum_{j=1}^{M} \mathcal{N}(\boldsymbol{x}_n \,|\, \boldsymbol{x}_0^j, \sigma_{t_n}^2)}$$

Hm, again a mixture? NO!

$$x_n \sim p_{t_n}(x_n \,|\, x_0^k);\;\; \hat{x}_{n-1} = Solver(x_n, t_n, t_{n-1} \,|\, \theta) \rightarrow \hat{x}_{n-1} \text{ depends only on } x_0^k$$

3. What about optimal solution?

$$\nabla_{f_\phi} \mathrm{J}(\boldsymbol{f}_\phi, x_n, n) = 0$$

$$\boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n) = \frac{\sum_{j=1}^{M} \left(\mathcal{N}(\boldsymbol{x}_n \,|\, \boldsymbol{x}_0^j, \sigma_{t_n}^2)\, \boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) \right)}{\sum_{j=1}^{M} \mathcal{N}(\boldsymbol{x}_n \,|\, \boldsymbol{x}_0^j, \sigma_{t_n}^2)}$$

Hm, again a mixture? NO!

$$x_n \sim p_{t_n}(x_n \,|\, x_0^k);\;\; \hat{x}_{n-1} = Solver(x_n, t_n, t_{n-1} \,|\, \theta) \rightarrow \hat{x}_{n-1} \text{ depends only on } x_0^k$$

$$\boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n) = \frac{\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) \sum_{j=1}^{M} \mathcal{N}(\boldsymbol{x}_n \,|\, \boldsymbol{x}_0^j, \sigma_{t_n}^2)}{\sum_{j=1}^{M} \mathcal{N}(\boldsymbol{x}_n \,|\, \boldsymbol{x}_0^j, \sigma_{t_n}^2)} = \boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1})$$

3. What about optimal solution?

$$\mathbb{E}_{n\sim\mathcal{U}(2,N)}\mathbb{E}_{p_{data}(\boldsymbol{x}_0)}\mathbb{E}_{q_{t_n}(\boldsymbol{x}_n|\boldsymbol{x}_0)}\|\boldsymbol{f}_\phi(\hat{\boldsymbol{x}}_{n-1}, t_{n-1}) - \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)\|_2^2 \to \min_\phi$$

$$\boldsymbol{f}_\phi(\boldsymbol{x}_N, t_N) = \boldsymbol{f}_\phi(\boldsymbol{x}_{N-1}, t_{N-1}) = \ldots = \boldsymbol{f}_\phi(\boldsymbol{x}_1, t_1)$$

But where $\boldsymbol{x}_0$? A possible solution:

$$\boldsymbol{f}_\phi(\boldsymbol{x}_N, t_N) = \boldsymbol{f}_\phi(\boldsymbol{x}_{N-1}, t_{N-1}) = \ldots = \boldsymbol{f}_\phi(\boldsymbol{x}_1, t_1) = 0$$

Boundary condition:

$$\boldsymbol{f}_\phi(\boldsymbol{x}_1, t_1) = \boldsymbol{x}_0 - \text{ sample from the dataset}$$

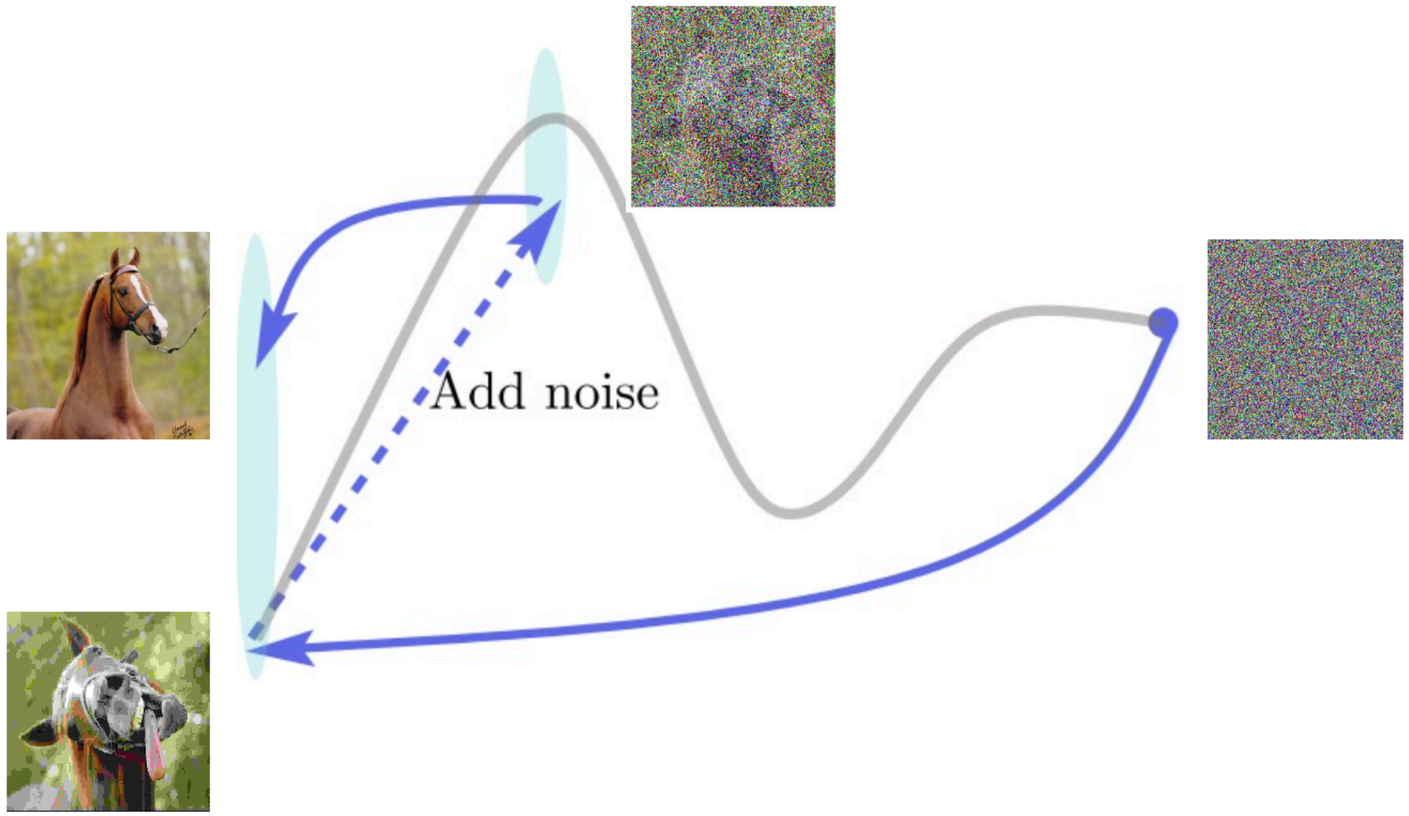
$$\boldsymbol{f}_\phi(\boldsymbol{x}_N, t_N) = \boldsymbol{f}_\phi(\boldsymbol{x}_{N-1}, t_{N-1}) = \ldots = \boldsymbol{f}_\phi(\boldsymbol{x}_1, t_1) = \boldsymbol{x}_0$$

3. Sampling from CD

$f_\phi(x_n, t_n)$ is not a score function anymore

→ we cannot use solvers to generate data

Stochastic multistep sampling:

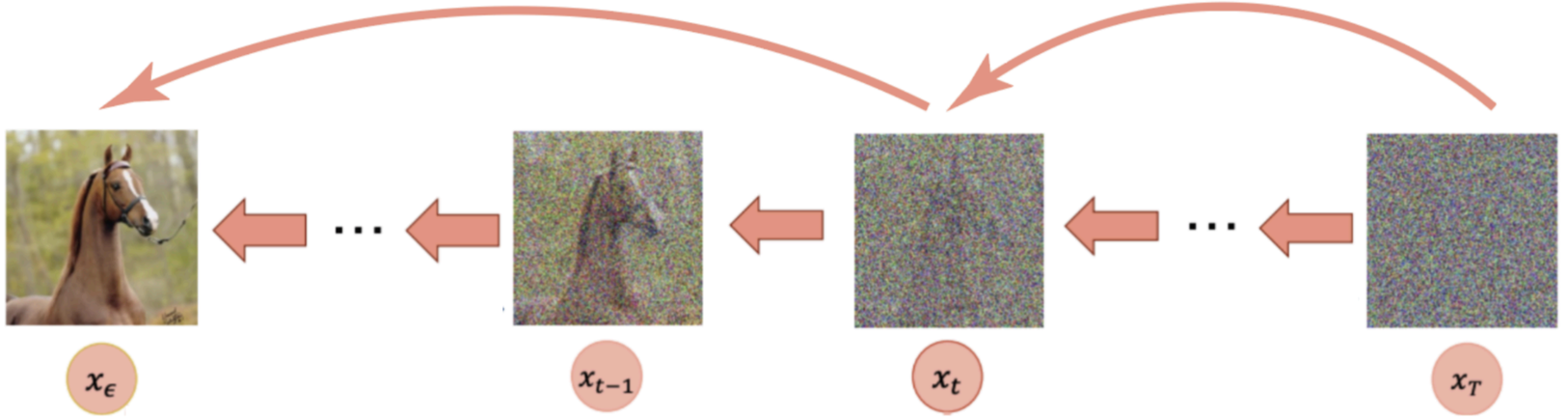


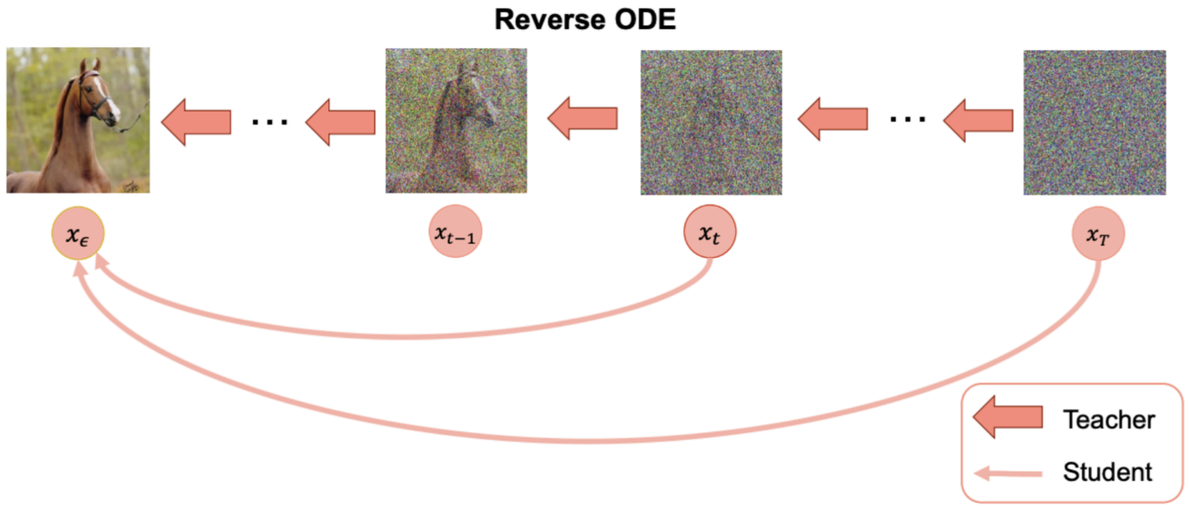
We need about 4-5 steps.

Still works bad for one step (but ODE-free approaches work better)

3. Multiboundary CD

Let’s split the solution interval



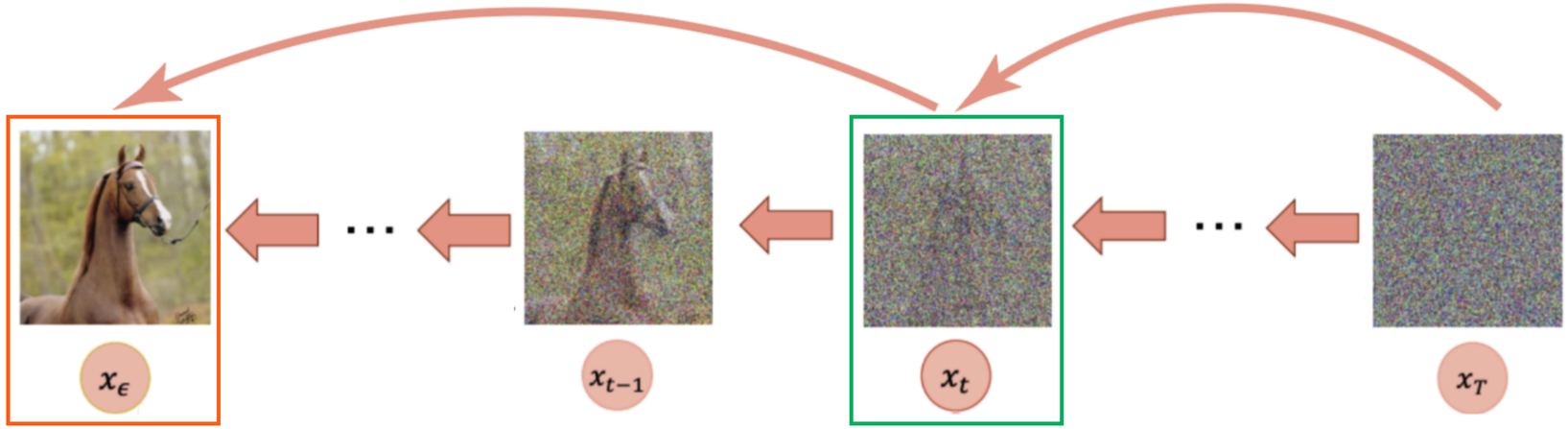


3. Multiboundary CD

For example, we can consider 2-boundary consistency distillation

$$\boldsymbol{f}_\phi(\boldsymbol{x}_1, t_1) = \ldots = \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n) = \boldsymbol{x}_0 \; - \text{first boundary}$$

$$\boldsymbol{f}_\phi(\boldsymbol{x}_{n+1}, t_{n+1}) = \ldots = \boldsymbol{f}_\phi(\boldsymbol{x}_N, t_N) = \boldsymbol{x}_{t_n} \; - \text{second boundary}$$



1. $\boldsymbol{x}_N \sim \mathcal{N}(\boldsymbol{x}_N | 0, \sigma^2)$, $\boldsymbol{x}_n = \boldsymbol{f}_\phi(\boldsymbol{x}_N, t_N)$; 2. $\boldsymbol{x}_0 = \boldsymbol{f}_\phi(\boldsymbol{x}_n, t_n)$

Summary

01

Why DMs are slow and how to accelerate them?

DM trajectories are curved

Diffusion distillation methods are needed

02

Knowledge Distillation

Requires a lot of synthetic samples and inherits discretization errors

03

Consistency Distillation

Does not require synthetic data
Does not inherit discretization errors
But the quality could be better even for 5 steps

04

Multi-boundary Consistency Distillation

Looks like a good solution, but still struggles with one step

Ours, 4 steps | Teacher, 50 steps

A donkey in a clown costume giving a lecture at the front of a lecture hall. The blackboard has mathematical equations on it.

Ours, 4 steps | Teacher, 50 steps

an airplane taxiing on a runway with the sun behind it

A castle made of tortilla chips, in a river made of salsa. There are tiny burritos walking around the castle

A raccoon wearing formal clothes, wearing a tophat and holding a cane. The raccoon is holding a garbage bag. Oil painting in the style of cubism.