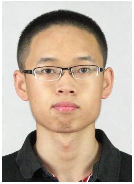
Accelerating Convergence of Score-Based Diffusion Models, Provably

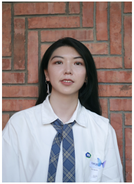
Yuxin Chen

Wharton Statistics & Data Science

*= equal contributions



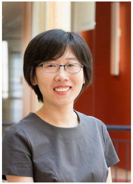
Gen Li*
CUHK



Yu Huang*
UPenn

Timofey Efimov
CMU

Yuting Wei
UPenn



Yuejie Chi
CMU

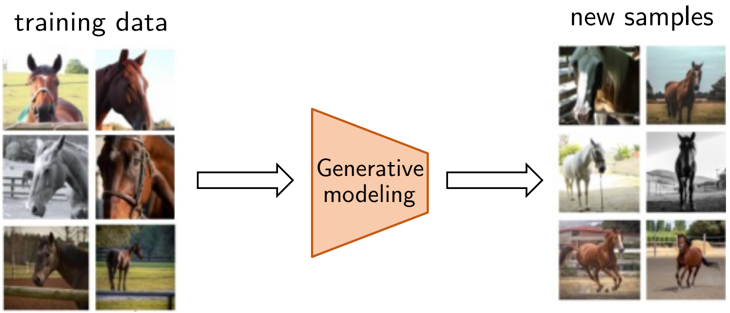
The era of generative AI

Generative models

training data

- Given training data $\underbrace{X^{\mathsf{train},i} \sim p_{\mathsf{data}}}_{\text{from a general distribution}}$ $(1 \leq i \leq N)$ in $\mathbb{R}^d$

Generative models



- Given training data $\underbrace{X^{\text{train},i} \sim p_{\text{data}}}_{\text{from a general distribution}}$ $(1 \leq i \leq N)$ in $\mathbb{R}^d$
- Generate new samples $Y \sim p_{\text{data}}$

Generative adversarial networks (GAN)

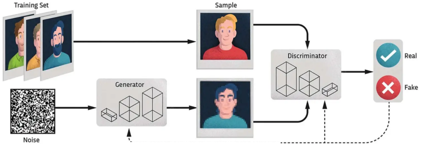


fig. credit: Science Focus

Variational autoencoder (VAE)

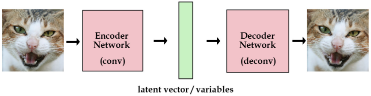


fig. credit: kevin frans blog

Diffusion models

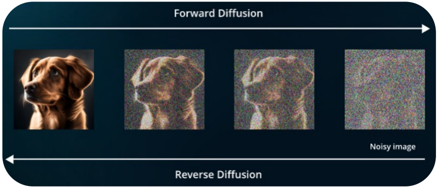


fig. credit: LeewayHertz

Inspired by nonequilibrium thermodynamics

— Sohl-Dickstein, Weiss, Maheswaranathan, Ganguli '15

Diffusion models

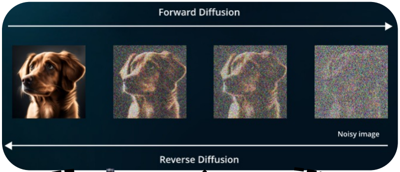
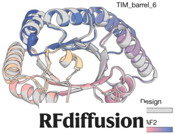


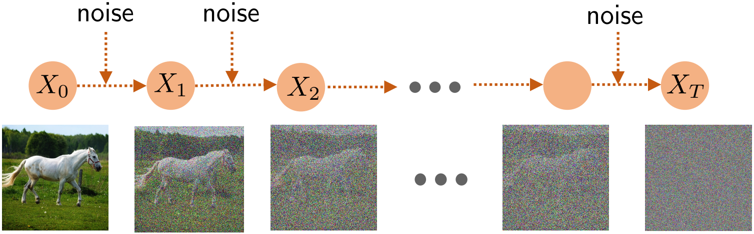
image generation

video generation

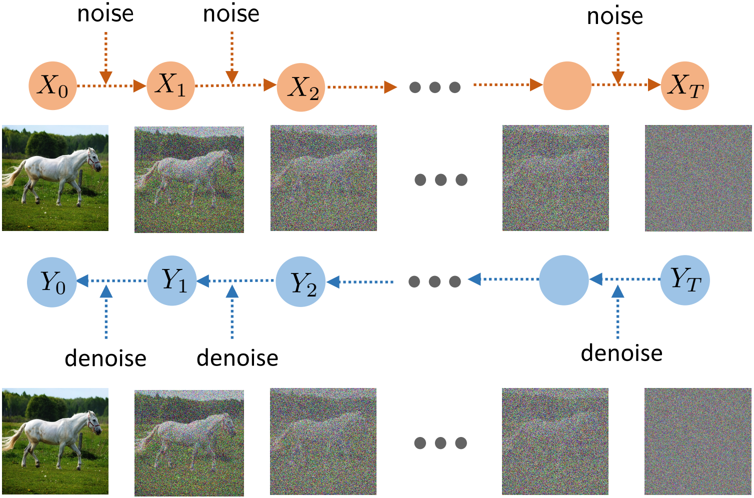


protein design

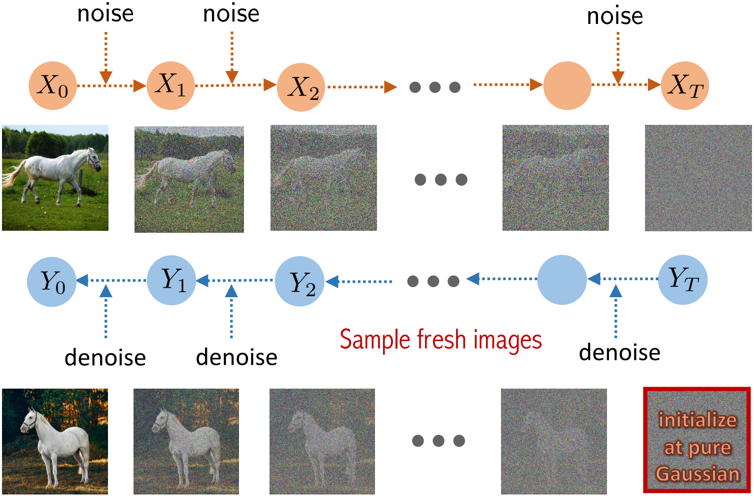
Preliminaries: score-based diffusion models



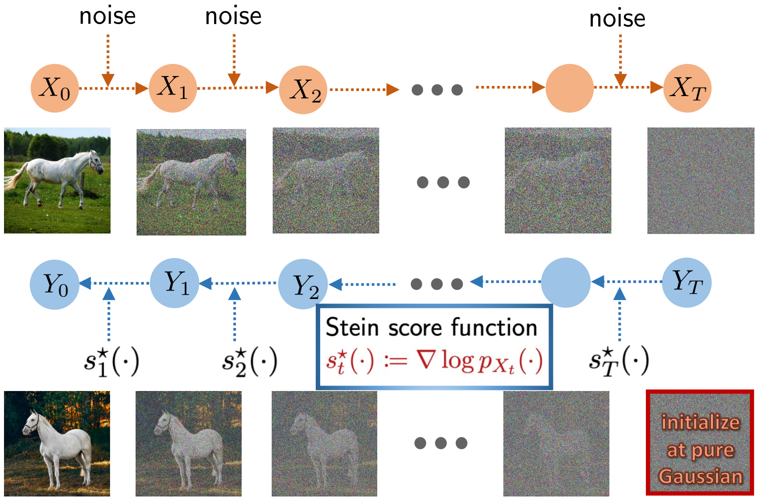
- **forward process:** (progressively) diffuse data into noise



- **forward process:** (progressively) diffuse data into noise
- **reverse process:** convert pure noise into data-like distributions



- **forward process:** (progressively) diffuse data into noise
- **reverse process:** convert pure noise into data-like distributions

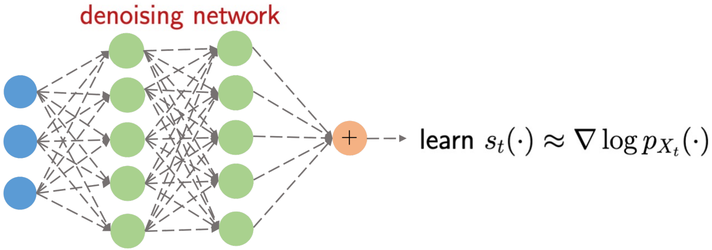


- **forward process:** (progressively) diffuse data into noise
- **reverse process:** convert pure noise into data-like distributions

Goal: $Y_t \overset{\mathrm{d}}{\approx} X_t, \quad t = T, \cdots, 1$

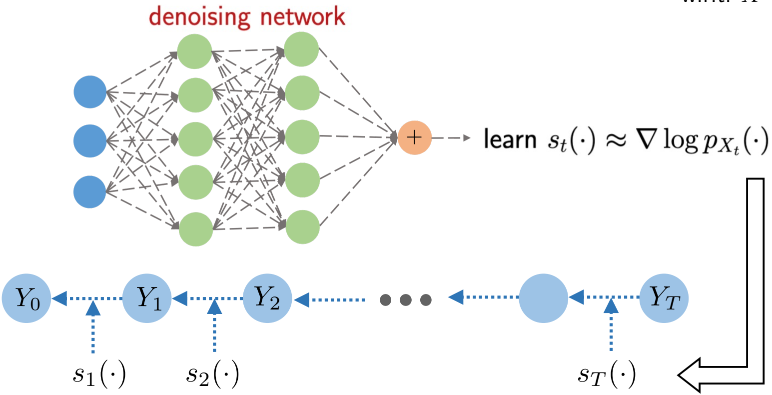
key component: **score functions** of forward process: $\underbrace{\nabla \log p_{X_t}(X)}_{\text{w.r.t. } X}$

key component: **score functions** of forward process: $\underbrace{\nabla \log p_{X_t}(X)}_{\text{w.r.t. } X}$



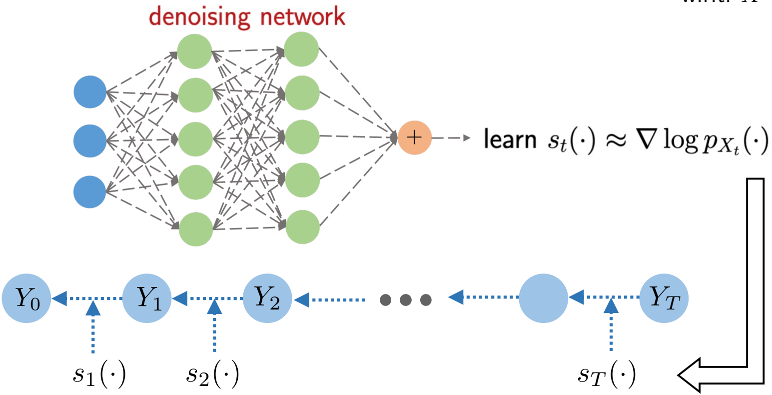
1. **score learning/matching:** learn estimates $s_t(\cdot)$ for $\nabla \log p_{X_t}(\cdot)$

key component: **score functions** of forward process: $\underbrace{\nabla \log p_{X_t}(X)}_{\text{w.r.t. } X}$



1. **score learning/matching:** learn estimates $s_t(\cdot)$ for $\nabla \log p_{X_t}(\cdot)$
2. **data generation:** sampling w/ the aid of score estimates $\{s_t(\cdot)\}$

key component: **score functions** of forward process: $\underbrace{\nabla \log p_{X_t}(X)}_{\text{w.r.t. } X}$



1. **score learning/matching:** learn estimates $s_t(\cdot)$ for $\nabla \log p_{X_t}(\cdot)$
2. **data generation:** sampling w/ the aid of score estimates $\{s_t(\cdot)\}$

Two mainstream approaches

— Ho, Jain, Abbeel '20

$$X_0 \sim p_{\mathsf{data}}, \quad X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d), \quad 1 \le t \le T$$

1. A stochastic sampler: $\underbrace{\textbf{denoising diffusion probabilistic models}}_{\text{DDPM}}$

Two mainstream approaches

— Ho, Jain, Abbeel '20

$$X_0 \sim p_{\mathsf{data}}, \quad X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d), \quad 1 \le t \le T$$

1. A stochastic sampler: $\underbrace{\textbf{denoising diffusion probabilistic models}}_{\text{DDPM}}$

$$Y_T \sim \mathcal{N}(0, I_d)$$

$$Y_{t-1} = \Psi_t(Y_t, \mathsf{noise}), \quad t = T, \cdots, 1$$

Two mainstream approaches

— *Ho, Jain, Abbeel '20*

$$X_0 \sim p_{\mathsf{data}}, \quad X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d), \quad 1 \le t \le T$$

1. A <u>stochastic</u> sampler: $\underbrace{\textbf{denoising diffusion probabilistic models}}_{\text{DDPM}}$

$$Y_T \sim \mathcal{N}(0, I_d)$$

$$Y_{t-1} = \underbrace{\frac{1}{\sqrt{1-\beta_t}}\Big(Y_t + \beta_t s_t(Y_t)\Big)}_{\text{deterministic component}} + \underbrace{\sqrt{\frac{\beta_t}{1-\beta_t}}\mathcal{N}(0, I_d)}_{\text{random component}}, \quad t = T, \cdots, 1$$

Two mainstream approaches

— Song, Sohl-Dickstein, Kingma, Kumar, Ermon, Poole '20
— Song, Meng, Ermon '20

$$X_0 \sim p_{\mathsf{data}}, \quad X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d), \quad 1 \leq t \leq T$$

2. A <u>deterministic</u> sampler: based on $\underbrace{\textbf{probability flow ODE}}_{\text{or DDIM}}$

Two mainstream approaches

— Song, Sohl-Dickstein, Kingma, Kumar, Ermon, Poole '20
— Song, Meng, Ermon '20

$$X_0 \sim p_{\text{data}}, \quad X_t = \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d), \quad 1 \le t \le T$$

2. A <u>deterministic</u> sampler: based on $\underbrace{\textbf{probability flow ODE}}_{\text{or DDIM}}$

$$Y_T \sim \mathcal{N}(0, I_d)$$

$$Y_{t-1} = \Phi_t(Y_t), \quad t = T, \cdots, 1$$

Two mainstream approaches

— Song, Sohl-Dickstein, Kingma, Kumar, Ermon, Poole '20

— Song, Meng, Ermon '20

$$X_0 \sim p_{\mathsf{data}}, \quad X_t = \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t} \mathcal{N}(0, I_d), \quad 1 \le t \le T$$

2. A deterministic sampler: based on $\underbrace{\textbf{probability flow ODE}}_{\text{or DDIM}}$

$$Y_T \sim \mathcal{N}(0, I_d)$$

$$Y_{t-1} = \underbrace{\frac{1}{\sqrt{1-\beta_t}} \left(Y_t + \frac{\beta_t}{2} s_t(Y_t) \right)}_{\text{purely deterministic}}, \qquad t = T, \cdots, 1$$

Interpretations: continuous-time limits

forward process
(marginal: $q_t := p_{X_t}$)

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d)$$

$$\Longrightarrow \quad \mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}W_t \quad \text{(SDE)}$$

Interpretations: continuous-time limits

forward process
(marginal: $q_t := p_{X_t}$)

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d)$$

$$\Longrightarrow \quad \mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}W_t \quad \text{(SDE)}$$

|| marginals

DDPM-type
stochastic sampler
(time-reversed SDE, Anderson '82)

$$Y_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\Big(Y_t + \beta_t \nabla \log q_t(Y_t)\Big) + \sqrt{\frac{\beta_t}{1-\beta_t}}\mathcal{N}(0, I_d)$$

$$\Longrightarrow \quad \mathrm{d}Y_t = \Big(-\frac{1}{2}\beta(t)Y_t - \beta(t)\nabla \log q_t(Y_t)\Big)\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\widetilde{W}_t \quad \text{(reversed)}$$

Interpretations: continuous-time limits

forward process
(marginal: $q_t := p_{X_t}$)

$$X_t = \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d)$$
$$\Longrightarrow \quad \mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}W_t \quad \text{(SDE)}$$

|| marginals

DDPM-type
stochastic sampler
(time-reversed SDE, Anderson '82)

$$Y_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\Big(Y_t + \beta_t \nabla \log q_t(Y_t)\Big) + \sqrt{\frac{\beta_t}{1-\beta_t}}\mathcal{N}(0, I_d)$$
$$\Longrightarrow \quad \mathrm{d}Y_t = \Big(-\frac{1}{2}\beta(t)Y_t - \beta(t)\nabla \log q_t(Y_t)\Big)\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\widetilde{W}_t \quad \text{(reversed)}$$

|| marginals

deterministic sampler
(probability flow ODE)

$$Y_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(Y_t + \frac{\beta_t}{2}\nabla \log q_t(Y_t)\right)$$
$$\Longrightarrow \quad \mathrm{d}Y_t = \left(-\frac{1}{2}\beta(t)Y_t - \frac{1}{2}\beta(t)\nabla \log q_t(Y_t)\right)\mathrm{d}t \quad \text{(reversed)}$$

Key takeaway: in continuous-time limits, sampling is feasible once perfect score functions are available

— *almost no restriction on target data distributions*

Key takeaway: in continuous-time limits, sampling is feasible once perfect score functions are available

— *almost no restriction on target data distributions*

Questions:

- what happens in discrete time? — effect of discretization error
- what if we only have imperfect scores? — effect of score error

Towards mathematical theory for diffusion models

- hard to develop full-fledged **end-to-end** theory

Towards mathematical theory for diffusion models

- hard to develop full-fledged **end-to-end** theory

  ✄ decouple
- score learning $\leftarrow$ ✗ $\rightarrow$ generative sampling

Towards mathematical theory for diffusion models

- hard to develop full-fledged **end-to-end** theory

$\quad\quad\quad\quad$ ✄ decouple

- score learning $\leftarrow$ ✗ $\rightarrow$ generative sampling

This talk:

1. non-asymptotic convergence theory in discrete time
2. acceleration?

Part 1: non-asymptotic convergence theory for probability flow ODE

"Towards non-asymptotic convergence for diffusion-based generative models,"
G. Li, Y. Wei, Y. Chen, Y. Chi, `arXiv:2306.09251`, ICLR 2024

Prior analyses for DDIM & DDPM

— Li, Lu, Tan '22
— Chen, Lee, Lu '22
— Chen, Chewi, Li, Li, Salim, Zhang '22
— Chen, Daras, Dimakis '23
— Chen, Chewi, Lee, Li, Lu, Salim '23
— Benton, De Bortoli, Doucet, Deligiannidis '23

discrete-time
diffusion process

continuous-time limits via
SDE/ODE toolbox (e.g., Girsanov thm)

Prior analyses for DDIM & DDPM

— Li, Lu, Tan '22
— Chen, Lee, Lu '22
— Chen, Chewi, Li, Li, Salim, Zhang '22
— Chen, Daras, Dimakis '23
— Chen, Chewi, Lee, Li, Lu, Salim '23
— Benton, De Bortoli, Doucet, Deligiannidis '23

discrete-time
diffusion process

continuous-time limits via
SDE/ODE toolbox (e.g., Girsanov thm)

control discretization error

Prior analyses for DDIM & DDPM

— Li, Lu, Tan '22
— Chen, Lee, Lu '22
— Chen, Chewi, Li, Li, Salim, Zhang '22
— Chen, Daras, Dimakis '23
— Chen, Chewi, Lee, Li, Lu, Salim '23
— Benton, De Bortoli, Doucet, Deligiannidis '23

discrete-time diffusion process

continuous-time limits via SDE/ODE toolbox (e.g., Girsanov thm)

control discretization error

Analogy: (stochastic) gradient descent vs. gradient flow, TD learning via ODE

Prior analyses for DDIM & DDPM

— *Li, Lu, Tan '22*
— *Chen, Lee, Lu '22*
— *Chen, Chewi, Li, Li, Salim, Zhang '22*
— *Chen, Daras, Dimakis '23*
— *Chen, Chewi, Lee, Li, Lu, Salim '23*
— *Benton, De Bortoli, Doucet, Deligiannidis '23*

discrete-time diffusion process

continuous-time limits via SDE/ODE toolbox (e.g., Girsanov thm)

control discretization error

- Built upon toolboxes from SDE/ODE
- Existing analyses **highly inadequate** for deterministic samplers

Can we develop a versatile non-asymptotic framework that

- *analyzes discrete-time processes directly*
- *accommodates both deterministic & stochastic samplers?*

Assumptions: target data distribution

$\mathbb{P}(\|X_0\|_2 \leq T^{c_R}) = 1$ for arbitrarily large const $c_R > 0$

Assumptions: target data distribution

$$\mathbb{P}(\|X_0\|_2 \leq T^{c_R}) = 1 \text{ for arbitrarily large const } c_R > 0$$

- support size can be very large

Assumptions: target data distribution

$$\mathbb{P}(\|X_0\|_2 \le T^{c_R}) = 1 \text{ for arbitrarily large const } c_R > 0$$

- support size can be very large
- very general: *no need of assumptions like log-concavity, smoothness, etc*

Assumptions: score estimates $\{s_t(\cdot)\}$

- *ℓ_2 score estimation error:* $s_t^\star(X) := \nabla \log p_{X_t}(X)$,

$$\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{X\sim p_{X_t}}\left[\|s_t(X) - s_t^\star(X)\|_2^2\right] \le \varepsilon_{\text{score}}^2$$

Assumptions: score estimates $\{s_t(\cdot)\}$

- ℓ_2 *score estimation error:* $s_t^\star(X) := \nabla \log p_{X_t}(X)$,

$$\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{X\sim p_{X_t}}\left[\|s_t(X) - s_t^\star(X)\|_2^2\right] \le \varepsilon_{\text{score}}^2$$

  - much weaker than *pointwise* score error assumption

Assumptions: score estimates $\{s_t(\cdot)\}$

- ℓ_2 *score estimation error:* $s_t^\star(X) := \nabla \log p_{X_t}(X)$,

$$\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{X\sim p_{X_t}} \left[\|s_t(X) - s_t^\star(X)\|_2^2\right] \le \varepsilon_{\mathsf{score}}^2$$

  - much weaker than *pointwise* score error assumption
  - this assumption alone is sufficient for DDPM
    but insufficient for DDIM (i.e., counterexamples exist)

Assumptions: score estimates $\{s_t(\cdot)\}$

- *ℓ_2 score estimation error:* $s_t^\star(X) := \nabla \log p_{X_t}(X)$,

$$\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{X\sim p_{X_t}} \left[\| s_t(X) - s_t^\star(X) \|_2^2 \right] \le \varepsilon_{\mathsf{score}}^2$$

  - much weaker than *pointwise* score error assumption
  - this assumption alone is sufficient for DDPM but insufficient for DDIM (i.e., counterexamples exist)

- *Jacobian estimation error:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{X\sim p_{X_t}} \left[\left\| \frac{\partial s_t}{\partial X}(X) - \frac{\partial s_t^\star}{\partial X}(X) \right\| \right] \le \varepsilon_{\mathsf{Jacobi}}$$

Learning rates

$$X_0 \sim p_{\mathsf{data}}, \quad X_t = \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d)$$

For some large constants $c_0, c_1 > 0$,

$$\begin{aligned} \beta_1 &= \frac{1}{T^{c_0}} \\ \beta_t &= \frac{c_1 \log T}{T} \min\left\{\beta_1\Big(1 + \frac{c_1 \log T}{T}\Big)^t, 1\right\} \end{aligned}$$

- 2 phases: (i) exponentially growing; (ii) flat
- common choice in diffusion model theory (e.g., Benton et al. '23)

Main result: probability flow ODE sampler

$$X_t = \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t} \mathcal{N}(0, I_d), \qquad t = 1, \cdots, T$$
$$Y_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left(Y_t + \frac{\beta_t}{2} s_t(Y_t) \right), \qquad t = T, \cdots, 1 \tag{1}$$

Theorem 1 (Li, Wei, Chen, Chi '23)

The probability flow ODE sampler (1) *obeys (up to log factor)*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{d^2}{T} + \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$$

Main result: probability flow ODE sampler

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d), \qquad t = 1, \cdots, T$$
$$Y_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(Y_t + \frac{\beta_t}{2}s_t(Y_t)\right), \qquad t = T, \cdots, 1 \tag{1}$$

Theorem 1 (Li, Wei, Chen, Chi '23)

The probability flow ODE sampler (1) *obeys (up to log factor)*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{d^2}{T} + \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$$

- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield TV dist} \leq \varepsilon}$ d^2/ε for small enough ε

Main result: probability flow ODE sampler

$$
\begin{aligned}
X_t &= \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d), \qquad & t = 1, \cdots, T \\
Y_{t-1} &= \frac{1}{\sqrt{1-\beta_t}}\left(Y_t + \frac{\beta_t}{2}s_t(Y_t)\right), \qquad & t = T, \cdots, 1
\end{aligned}
\tag{1}
$$

Theorem 1 (Li, Wei, Chen, Chi '23)

The probability flow ODE sampler (1) *obeys (up to log factor)*

$$
\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{d^2}{T} + \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}
$$

- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield TV dist} \leq \varepsilon}$ d^2/ε for small enough ε
- **stability:** $\mathsf{TV}(p_{X_1}, p_{Y_1}) \propto$ error measures $\varepsilon_{\mathsf{score}}$ and $\varepsilon_{\mathsf{Jacobi}}$

Main result: probability flow ODE sampler

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d), \qquad t = 1, \cdots, T$$
$$Y_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(Y_t + \frac{\beta_t}{2}s_t(Y_t)\right), \qquad t = T, \cdots, 1 \tag{1}$$

Theorem 1 (Li, Wei, Chen, Chi '23)

The probability flow ODE sampler (1) *obeys (up to log factor)*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{d^2}{T} + \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$$

- **general data distribution:** don't need smoothness, log-concavity

Main result: probability flow ODE sampler

$$
\begin{aligned}
X_t &= \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d), \qquad && t = 1, \cdots, T \\
Y_{t-1} &= \frac{1}{\sqrt{1-\beta_t}}\left(Y_t + \frac{\beta_t}{2}s_t(Y_t)\right), \qquad && t = T, \cdots, 1
\end{aligned} \tag{1}
$$

Theorem 1 (Li, Wei, Chen, Chi '23)

The probability flow ODE sampler (1) *obeys (up to log factor)*

$$
\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{d^2}{T} + \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}
$$

- **general data distribution:** don't need smoothness, log-concavity
- d**-dependency:** might be improvable to $\frac{d}{T} + \cdots$ (ongoing work)

Comparison w/ prior probability flow ODE theory

(our theory) $\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{d^2}{T} + \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$

- *Chen, Daras, Dimakis '23:* $\underbrace{\text{no concrete poly dependency}}_{\text{ours: } d^2/\varepsilon}$
  $\underbrace{\text{exponential in smoothness parameter}}_{\text{ours: independent of smoothness pars}}$
  $\underbrace{\text{needs exact score functions}}_{\text{ours: allow score errors}}$

Comparison w/ prior probability flow ODE theory

$$\text{(our theory)} \quad \mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim \frac{d^2}{T} + \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$$

- *Chen, Daras, Dimakis '23:* $\underbrace{\text{no concrete poly dependency}}_{\text{ours: } d^2/\varepsilon}$ $\underbrace{\text{exponential in smoothness parameter}}_{\text{ours: independent of smoothness pars}}$ $\underbrace{\text{needs exact score functions}}_{\text{ours: allow score errors}}$
- *Chen, Chewi, Lee, Li, Lu, Salim '23:* requires additional $\underbrace{\text{stochastic correction steps}}_{\text{different from probability flow ODE}}$ & smoothness

Proof strategy

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I), \quad Y_{t-1} = \underbrace{\frac{1}{\sqrt{1-\beta_t}}Y_t + \frac{\beta_t}{2\sqrt{1-\beta_t}}s_t(Y_t)}_{=: \Phi_t(Y_t)}$$

$\mathsf{TV}\big(p_{X_t}, p_{Y_t}\big) \approx 0$

Proof strategy

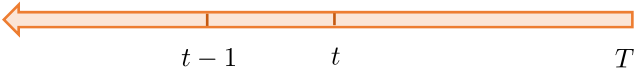
$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0,I), \quad Y_{t-1} = \underbrace{\frac{1}{\sqrt{1-\beta_t}}Y_t + \frac{\beta_t}{2\sqrt{1-\beta_t}}s_t(Y_t)}_{=:\,\Phi_t(Y_t)}$$

$$\mathsf{TV}\big(p_{X_t}, p_{Y_t}\big) \approx 0 \qquad \Longleftarrow \qquad \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)} \approx 1 \quad \forall y_t \in \mathcal{E}_t \text{ (some “typical” set)}$$

Proof strategy

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I), \quad Y_{t-1} = \underbrace{\frac{1}{\sqrt{1-\beta_t}}Y_t + \frac{\beta_t}{2\sqrt{1-\beta_t}}s_t(Y_t)}_{=:\, \Phi_t(Y_t)}$$

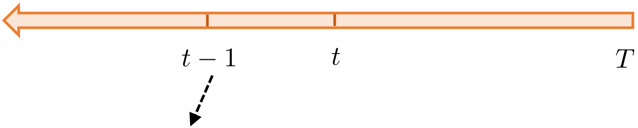
$$\mathsf{TV}\big(p_{X_t}, p_{Y_t}\big) \approx 0 \qquad \Longleftarrow \qquad \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)} \approx 1 \quad \forall y_t \in \mathcal{E}_t \text{ (some "typical" set)}$$



Proof strategy

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0,I), \quad Y_{t-1} = \underbrace{\frac{1}{\sqrt{1-\beta_t}}Y_t + \frac{\beta_t}{2\sqrt{1-\beta_t}}s_t(Y_t)}_{=:\,\Phi_t(Y_t)}$$

$$\mathsf{TV}\big(p_{X_t}, p_{Y_t}\big) \approx 0 \qquad \Longleftarrow \qquad \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)} \approx 1 \quad \forall y_t \in \mathcal{E}_t \text{ (some “typical” set)}$$

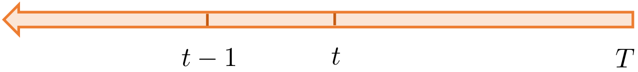


$$\frac{p_{Y_{t-1}}(y_{t-1})}{p_{X_{t-1}}(y_{t-1})} = \underbrace{\frac{p_{Y_{t-1}}(y_{t-1})}{p_{Y_t}(y_t)}}_{\text{relation btw } Y_t \ \& \ Y_{t-1}} \Bigg(\underbrace{\frac{p_{X_{t-1}}(y_{t-1})}{p_{X_t}(y_t)}}_{\text{relation btw } X_t \ \& \ X_{t-1}} \Bigg)^{-1} \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)}$$

Proof strategy

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0,I), \quad Y_{t-1} = \underbrace{\frac{1}{\sqrt{1-\beta_t}}Y_t + \frac{\beta_t}{2\sqrt{1-\beta_t}}s_t(Y_t)}_{=:\,\Phi_t(Y_t)}$$

$$\mathsf{TV}\big(p_{X_t}, p_{Y_t}\big) \approx 0 \qquad \Longleftarrow \qquad \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)} \approx 1 \quad \forall y_t \in \mathcal{E}_t \text{ (some “typical” set)}$$

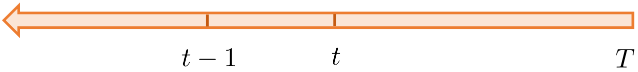


$$\frac{p_{Y_{t-1}}(\Phi_t(y_t))}{p_{X_{t-1}}(\Phi_t(y_t))} = \underbrace{\frac{p_{Y_{t-1}}(\Phi_t(y_t))}{p_{Y_t}(y_t)}}_{\text{relation btw } Y_t \text{ \& } Y_{t-1}} \Bigg(\underbrace{\frac{p_{X_{t-1}}(\Phi_t(y_t))}{p_{X_t}(y_t)}}_{\text{relation btw } X_t \text{ \& } X_{t-1}} \Bigg)^{-1} \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)}$$

Proof strategy

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0,I), \quad Y_{t-1} = \underbrace{\frac{1}{\sqrt{1-\beta_t}}Y_t + \frac{\beta_t}{2\sqrt{1-\beta_t}}s_t(Y_t)}_{=:\,\Phi_t(Y_t)}$$

$$\mathsf{TV}\big(p_{X_t}, p_{Y_t}\big) \approx 0 \qquad \Longleftarrow \qquad \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)} \approx 1 \quad \forall y_t \in \mathcal{E}_t \text{ (some “typical” set)}$$



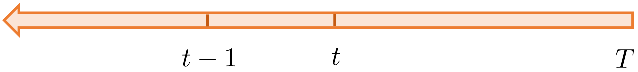
$$\frac{p_{Y_{t-1}}(\Phi_t(y_t))}{p_{X_{t-1}}(\Phi_t(y_t))} = \underbrace{\frac{p_{Y_{t-1}}(\Phi_t(y_t))}{p_{Y_t}(y_t)}}_{\text{relation btw } Y_t \,\&\, Y_{t-1}} \Bigg(\underbrace{\frac{p_{X_{t-1}}(\Phi_t(y_t))}{p_{X_t}(y_t)}}_{\text{relation btw } X_t \,\&\, X_{t-1}} \Bigg)^{-1} \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)}$$

$$\frac{p_{\Phi_t(Y_t)}(\Phi_t(y_t))}{p_{Y_t}(y_t)} = \det\left(\frac{\partial \Phi_t}{\partial y_t}\right)^{-1}$$

Proof strategy

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0,I), \quad Y_{t-1} = \underbrace{\frac{1}{\sqrt{1-\beta_t}}Y_t + \frac{\beta_t}{2\sqrt{1-\beta_t}}s_t(Y_t)}_{=:\, \Phi_t(Y_t)}$$

$$\mathsf{TV}\big(p_{X_t}, p_{Y_t}\big) \approx 0 \quad \Longleftarrow \quad \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)} \approx 1 \quad \forall y_t \in \mathcal{E}_t \text{ (some “typical” set)}$$



$$\frac{p_{Y_{t-1}}(\Phi_t(y_t))}{p_{X_{t-1}}(\Phi_t(y_t))} = \underbrace{\frac{p_{Y_{t-1}}(\Phi_t(y_t))}{p_{Y_t}(y_t)}}_{\text{relation btw } Y_t \,\&\, Y_{t-1}} \Bigg(\underbrace{\frac{p_{X_{t-1}}(\Phi_t(y_t))}{p_{X_t}(y_t)}}_{\text{relation btw } X_t \,\&\, X_{t-1}} \Bigg)^{-1} \frac{p_{Y_t}(y_t)}{p_{X_t}(y_t)}$$

$$\frac{p_{\Phi_t(Y_t)}\big(\Phi_t(y_t)\big)}{p_{Y_t}(y_t)} = \det\left(\frac{\partial \Phi_t}{\partial y_t}\right)^{-1}$$

some concentration bounds

Part 2: acceleration

"Accelerating convergence of score-based diffusion models, provably," G. Li*, Y. Huang*, T. Efimov, Y. Wei, Y. Chi, Y. Chen, `arXiv:2403.03852`, 2024

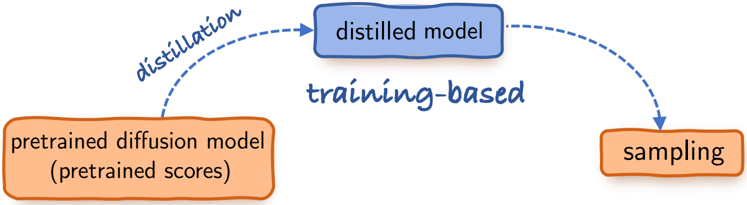
Diffusion-based sampling is often slow

— *Song, Meng, Ermon '20*

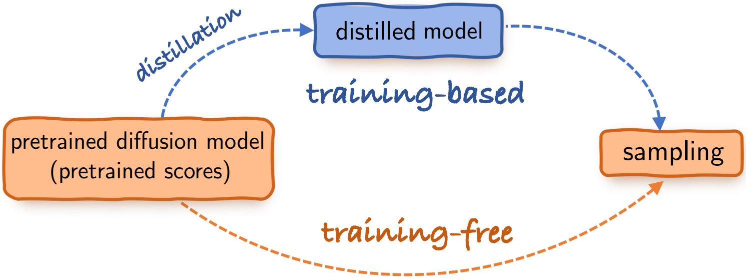
Diffusion-based sampling is often slow

50K 32$\times$32 images: DDPM (20h) *vs.* single-step GANs ($<$ 1min)

— *Song, Meng, Ermon '20*



- **Training-based:** distill pre-trained diffusion model (requires additional training) into another model that can be executed rapidly
  - e.g., progressive distillation (Salimans et al. '22), consistency model (Song et al. '23), . . .



- **Training-free:** directly invoke pre-trained diffusion models (particularly score estimates) for sampling w/o additional training
  - e.g., DPM-Solver/++ (Lu et al. '22), UniPC (Zhao et al. '23), ...

Can we design a training-free deterministic sampler that converges provably faster than probability flow ODE?

Proposed accelerated deterministic sampler

$$Y_t^{-} = \Phi_t(Y_t), \qquad Y_{t-1} = \Psi_t(Y_t, Y_t^{-}) \qquad \text{for } t = T, \cdots, 1 \qquad (2)$$

- compute $\underbrace{\text{a midpoint } Y_t^{-}}_{\text{estimate of } Y_{t+1} \text{ using } Y_t}$; update based on $\underbrace{\text{both } Y_t \text{ and } Y_t^{-}}_{\text{provide 2nd-order info}}$

Proposed accelerated deterministic sampler

$$Y_t^- = \Phi_t(Y_t), \qquad Y_{t-1} = \Psi_t(Y_t, Y_t^-) \qquad \text{for } t = T, \cdots, 1 \qquad (2)$$

$$\Psi_t(Y_t, Y_t^-) = \frac{1}{\sqrt{\alpha_t}} \Bigg(\underbrace{Y_t + \frac{1-\alpha_t}{2} s_t(Y_t)}_{\text{original DDIM}} \Bigg)$$

- compute $\underbrace{\text{a midpoint } Y_t^-}_{\text{estimate of } Y_{t+1} \text{ using } Y_t}$; update based on $\underbrace{\text{both } Y_t \text{ and } Y_t^-}_{\text{provide 2nd-order info}}$

Proposed accelerated deterministic sampler

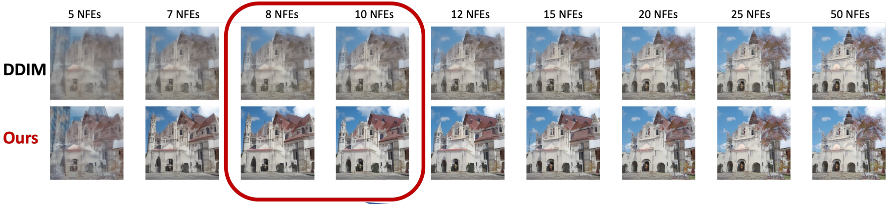
$$Y_t^- = \Phi_t(Y_t), \qquad Y_{t-1} = \Psi_t(Y_t, Y_t^-) \qquad \text{for } t = T, \cdots, 1 \qquad (2)$$

$$\Phi_t(Y_t) = \sqrt{\alpha_{t+1}}\left(Y_t - \frac{1-\alpha_{t+1}}{2} s_t(Y_t)\right)$$

$$\Psi_t(Y_t, Y_t^-) = \frac{1}{\sqrt{\alpha_t}}\Bigg(\underbrace{Y_t + \frac{1-\alpha_t}{2} s_t(Y_t)}_{\text{original DDIM}} + \underbrace{\frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\big(s_t(Y_t) - \sqrt{\alpha_{t+1}} s_{t+1}(Y_t^-)\big)}_{\text{“momentum”}}\Bigg)$$

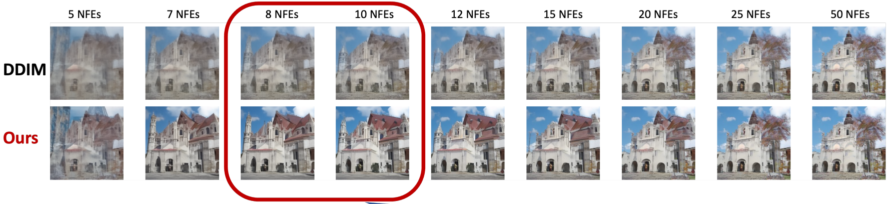
- compute $\underbrace{\text{a midpoint } Y_t^-}_{\text{estimate of } Y_{t+1} \text{ using } Y_t}$; update based on $\underbrace{\text{both } Y_t \text{ and } Y_t^-}_{\text{provide 2nd-order info}}$
- 2 score function evaluations per iteration

Numbers of function evaluation (NFE) 4 → 50



high-quality samples within **10** NFEs

Numbers of function evaluation (NFE) 4 → 50



high-quality samples within **10** NFEs

sampled images with 5 NFEs: crisper and less noisy

Recap: our assumptions

- ℓ_2 **score estimation error:**

$$\frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{X\sim p_{X_t}}\left[\|s_t(X)-s_t^{\star}(X)\|_2^2\right]\leq\varepsilon_{\mathsf{score}}^2$$

- **Jacobian estimation error:**

$$\frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{X\sim p_{X_t}}\left[\left\|\frac{\partial s_t}{\partial X}(X)-\frac{\partial s_t^{\star}}{\partial X}(X)\right\|\right]\leq\varepsilon_{\mathsf{Jacobi}}$$

- $\mathbb{P}(\|X_0\|_2\leq T^{c_R})=1$ for arbitrarily large const $c_R>0$

Main result: accelerated deterministic sampler

Theorem 2 (Li, Huang, Efimov, Wei, Chi, Chen '24)

Our accelerated deterministic sampler (2) *obeys (up to log factor)*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$$

Main result: accelerated deterministic sampler

Theorem 2 (Li, Huang, Efimov, Wei, Chi, Chen '24)

Our accelerated deterministic sampler (2) *obeys (up to log factor)*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$$

- $\underbrace{\textbf{iteration complexity}}_{\text{to yield TV dist} \leq \varepsilon}$: $\frac{\mathrm{poly}(d)}{\sqrt{\varepsilon}}$
    - outperforms vanilla DDIM (iteration complexity: $\mathrm{poly}(d)/\varepsilon$)

Main result: accelerated deterministic sampler

Theorem 2 (Li, Huang, Efimov, Wei, Chi, Chen '24)

Our accelerated deterministic sampler (2) *obeys (up to log factor)*

$$\mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$$

- $\underbrace{\textbf{iteration complexity}}_{\text{to yield TV dist} \leq \varepsilon}$: $\frac{\mathrm{poly}(d)}{\sqrt{\varepsilon}}$
  - outperforms vanilla DDIM (iteration complexity: $\mathrm{poly}(d)/\varepsilon$)
- **stability:** TV distance propotional to $\varepsilon_{\mathsf{score}} + \varepsilon_{\mathsf{Jacobi}}$
- **minimal assumptions** on data distributions

Main result: accelerated deterministic sampler

Theorem 2 (Li, Huang, Efimov, Wei, Chi, Chen '24)

Our accelerated deterministic sampler (2) *obeys (up to log factor)*

$$\mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim \frac{d^6}{T^2} + \sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$$

- $\underbrace{\textbf{iteration complexity}}_{\text{to yield TV dist} \leq \varepsilon}$: $\frac{\mathrm{poly}(d)}{\sqrt{\varepsilon}}$
  - outperforms vanilla DDIM (iteration complexity: $\mathrm{poly}(d)/\varepsilon$)
- **stability:** TV distance propotional to $\varepsilon_{\mathsf{score}} + \varepsilon_{\mathsf{Jacobi}}$
- **minimal assumptions** on data distributions
- d**-dependency:** might be improvable to $\frac{d^4}{T^2}$ (ongoing work)

Interpretation via high-order discretization

$$X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t}\,\mathcal{N}(0, I_d) \qquad \text{with } \overline{\alpha}_t := \prod_{k=1}^{t}(1-\beta_k)$$

Interpretation via high-order discretization

$$X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t}\,\mathcal{N}(0, I_d) \qquad \text{with } \overline{\alpha}_t := \prod_{k=1}^{t} (1-\beta_k)$$

General form for $0 < \gamma < 1$:

$$X(\gamma) := \sqrt{\gamma} X_0 + \sqrt{1-\gamma}\,\mathcal{N}(0, I_d)$$
$$s^{\star}_{\gamma}(x) := \nabla \log p_{X(\gamma)}(x)$$

Interpretation via high-order discretization

$$X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t}\,\mathcal{N}(0, I_d) \qquad \text{with } \overline{\alpha}_t := \prod_{k=1}^{t}(1-\beta_k)$$

General form for $0 < \gamma < 1$:

$$X(\gamma) := \sqrt{\gamma} X_0 + \sqrt{1-\gamma}\,\mathcal{N}(0, I_d)$$
$$s_\gamma^\star(x) := \nabla \log p_{X(\gamma)}(x)$$

$$\Longrightarrow \quad X(\overline{\alpha_t}) \overset{\mathrm{d}}{=} X_t, \; s^\star_{\overline{\alpha_t}} = s^\star_t$$

Interpretation via high-order discretization

$$X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t}\,\mathcal{N}(0, I_d) \qquad \text{with } \overline{\alpha}_t := \prod_{k=1}^{t}(1-\beta_k)$$

General form for $0 < \gamma < 1$:

$$X(\gamma) := \sqrt{\gamma} X_0 + \sqrt{1-\gamma}\,\mathcal{N}(0, I_d)$$
$$s^\star_\gamma(x) := \nabla \log p_{X(\gamma)}(x)$$

$$\Longrightarrow \quad X(\overline{\alpha_t}) \overset{\mathrm{d}}{=} X_t, \; s^\star_{\overline{\alpha_t}} = s^\star_t$$

A key reversed relation $X(\overline{\alpha}_{t-1}) \to X(\overline{\alpha}_t)$:

$$X(\overline{\alpha}_{t-1}) = \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}}}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} s^\star_\gamma\big(X(\gamma)\big)\mathrm{d}\gamma$$

Interpretation via high-order discretization

$$X_t \overset{\mathrm{d}}{=} \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1 - \overline{\alpha}_t} \mathcal{N}(0, I_d) \qquad \text{with } \overline{\alpha}_t := \prod_{k=1}^{t} (1 - \beta_k)$$

General form for $0 < \gamma < 1$:

$$X(\gamma) := \sqrt{\gamma} X_0 + \sqrt{1-\gamma} \mathcal{N}(0, I_d)$$
$$s^\star_\gamma(x) := \nabla \log p_{X(\gamma)}(x)$$

$$\Longrightarrow \quad X(\overline{\alpha_t}) \overset{\mathrm{d}}{=} X_t, \;\; s^\star_{\overline{\alpha_t}} = s^\star_t$$

A key reversed relation $X(\overline{\alpha}_{t-1}) \to X(\overline{\alpha}_t)$:

$$X(\overline{\alpha}_{t-1}) = \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}}}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} s^\star_\gamma\big(X(\gamma)\big) \mathrm{d}\gamma$$

"solution" to probability flow ODE

Interpretation via high-order discretization

$$X(\overline{\alpha}_{t-1}) = \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}}}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s_\gamma^\star\big(X(\gamma)\big)}_{\text{approximated by?}} \mathrm{d}\gamma$$

Interpretation via high-order discretization

$$X(\overline{\alpha}_{t-1}) = \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}}}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\big(X(\gamma)\big)}_{\text{approximated by?}} \, \mathrm{d}\gamma$$

Scheme 1: $s^\star_\gamma\big(X(\gamma)\big) \approx s^\star_{\overline{\alpha}_t}\big(X(\overline{\alpha}_t)\big) \approx s_t(X_t)$

$$\Longrightarrow \quad X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}} \left(X(\overline{\alpha}_t) + \frac{1-\alpha_t}{2} s_t(X_t) \right) \quad \text{original DDIM}$$

Interpretation via high-order discretization

$$X(\overline{\alpha}_{t-1}) = \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}}}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s_\gamma^\star\big(X(\gamma)\big)}_{\text{approximated by?}} \, \mathrm{d}\gamma$$

Scheme 1: $s_\gamma^\star\big(X(\gamma)\big) \approx s_{\overline{\alpha}_t}^\star\big(X(\overline{\alpha}_t)\big) \approx s_t(X_t)$

$$\Longrightarrow \quad X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}} \left(X(\overline{\alpha}_t) + \frac{1-\alpha_t}{2} s_t(X_t) \right) \quad \text{original DDIM}$$

refined approximation?

Interpretation via high-order discretization

$$X(\overline{\alpha}_{t-1}) = \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}}}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\big(X(\gamma)\big)}_{\text{approximated by?}} \mathrm{d}\gamma$$

Scheme 1: $s^\star_\gamma\big(X(\gamma)\big) \approx s^\star_{\overline{\alpha}_t}\big(X(\overline{\alpha}_t)\big) \approx s_t(X_t)$

$$\implies \quad X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}} \left(X(\overline{\alpha}_t) + \frac{1-\alpha_t}{2} s_t(X_t) \right) \quad \text{original DDIM}$$

refined approximation?

$$\begin{aligned} s^\star_\gamma\big(X(\gamma)\big) &\approx s^\star_{\overline{\alpha}_t}\big(X(\overline{\alpha}_t)\big) + \frac{\mathrm{d}s^\star_\gamma\big(X(\gamma)\big)}{\mathrm{d}\gamma} (\gamma - \overline{\alpha}_t) \\ &\approx s_t(X_t) + \frac{\gamma - \overline{\alpha}_t}{\overline{\alpha}_t - \overline{\alpha}_{t+1}} \Big(s_t(X_t) - s_{t+1}(X_{t+1}) \Big) \end{aligned}$$

Interpretation via high-order discretization

$$X(\overline{\alpha}_{t-1}) = \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}}}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\big(X(\gamma)\big)}_{\text{approximated by?}} \, \mathrm{d}\gamma$$

Scheme 1: $s^\star_\gamma\big(X(\gamma)\big) \approx s^\star_{\overline{\alpha}_t}\big(X(\overline{\alpha}_t)\big) \approx s_t(X_t)$

$$\implies \quad X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}} \left(X(\overline{\alpha}_t) + \frac{1-\alpha_t}{2} s_t(X_t) \right) \quad \text{original DDIM}$$

Scheme 2: $s^\star_\gamma\big(X(\gamma)\big) \approx s_t(X_t) + \frac{\gamma - \overline{\alpha}_t}{\overline{\alpha}_t - \overline{\alpha}_{t+1}} \big(s_t(X_t) - s_{t+1}(X_{t+1})\big)$

$$\implies \quad X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}} \left(X(\overline{\alpha}_t) + \frac{1-\alpha_t}{2} s_t(X_t) \right) + \frac{1}{\sqrt{\alpha_t}} \left(\frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})} \Big(s_t\big(X_t\big) - \sqrt{\alpha_{t+1}} s_{t+1}\big(X_{t+1}\big) \Big) \right) \quad \textbf{Ours}$$

Interpretation via high-order discretization

$$X(\overline{\alpha}_{t-1}) = \frac{1}{\sqrt{\alpha_t}} X(\overline{\alpha}_t) + \frac{\sqrt{\overline{\alpha}_{t-1}}}{2} \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s_\gamma^\star\big(X(\gamma)\big)}_{\text{approximated by?}} \mathrm{d}\gamma$$

Scheme 1: $s_\gamma^\star\big(X(\gamma)\big) \approx s_{\overline{\alpha}_t}^\star\big(X(\overline{\alpha}_t)\big) \approx s_t(X_t)$

$$\Longrightarrow \quad X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}} \left(X(\overline{\alpha}_t) + \frac{1-\alpha_t}{2} s_t(X_t) \right) \quad \text{original DDIM}$$

Scheme 2: $s_\gamma^\star\big(X(\gamma)\big) \approx s_t(X_t) + \frac{\gamma - \overline{\alpha}_t}{\overline{\alpha}_t - \overline{\alpha}_{t+1}} \big(s_t(X_t) - s_{t+1}(X_{t+1})\big)$

$$\Longrightarrow \quad X(\overline{\alpha}_{t-1}) \approx \frac{1}{\sqrt{\alpha_t}} \left(X(\overline{\alpha}_t) + \frac{1-\alpha_t}{2} s_t(X_t) \right) + \frac{1}{\sqrt{\alpha_t}} \left(\frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})} \Big(s_t\big(X_t\big) - \sqrt{\alpha_{t+1}} s_{t+1}\big(X_{t+1}\big) \Big) \right) \quad \textbf{Ours}$$

— *similar in spirit to DPM-Solver-2 (Lu et al '22)*

Can we design a training-free stochastic sampler that converges provably faster than DDPM?

Proposed accelerated stochastic sampler

$$Y_t^+ = \Phi_t(Y_t, Z_t), \quad Y_{t-1} = \Psi_t(Y_t^+, Z_t^+) \quad \text{with } Z_t, Z_t^+ \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d) \quad (3)$$

- compute a midpoint Y_t^+; then compute Y_{t-1} using Y_t^+ (similar to extragradient method)

Proposed accelerated stochastic sampler

$$Y_t^+ = \Phi_t(Y_t, Z_t), \quad Y_{t-1} = \Psi_t(Y_t^+, Z_t^+) \quad \text{with } Z_t, Z_t^+ \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d) \quad (3)$$

$$\Phi_t(x, z) = x + \sqrt{\frac{1-\alpha_t}{2}} z \quad \text{injecting additional noise}$$

$$\Psi_t(y, z) = \frac{1}{\sqrt{\alpha_t}} \left(y + (1-\alpha_t) s_t(y) + \sqrt{\frac{1-\alpha_t}{2}} z \right) \quad \text{same as DDPM}$$

- compute a midpoint Y_t^+; then compute Y_{t-1} using Y_t^+ (similar to extragradient method)

Proposed accelerated stochastic sampler

$$Y_t^+ = \Phi_t(Y_t, Z_t), \quad Y_{t-1} = \Psi_t(Y_t^+, Z_t^+) \quad \text{with } Z_t, Z_t^+ \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d) \quad (3)$$

$$\Phi_t(x, z) = x + \sqrt{\frac{1-\alpha_t}{2}} z \quad \text{injecting additional noise}$$

$$\Psi_t(y, z) = \frac{1}{\sqrt{\alpha_t}} \left(y + (1-\alpha_t) s_t(y) + \sqrt{\frac{1-\alpha_t}{2}} z \right) \quad \text{same as DDPM}$$

- compute a midpoint Y_t^+; then compute Y_{t-1} using Y_t^+ (similar to extragradient method)
- 1 score function evaluation per iteration

Main result: accelerated stochastic sampler

Theorem 3 (Li, Huang, Efimov, Wei, Chi, Chen ’24)

The accelerated stochastic sampler (3) *obeys (up to log factor)*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \sqrt{\mathsf{KL}(p_{X_1} \parallel p_{Y_1})} \lesssim \frac{d^3}{T} + \sqrt{d}\varepsilon_{\mathsf{score}}$$

Main result: accelerated stochastic sampler

Theorem 3 (Li, Huang, Efimov, Wei, Chi, Chen '24)

The accelerated stochastic sampler (3) *obeys (up to log factor)*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \sqrt{\mathsf{KL}(p_{X_1} \parallel p_{Y_1})} \lesssim \frac{d^3}{T} + \sqrt{d}\varepsilon_{\mathsf{score}}$$

- **iteration complexity**: $\frac{\mathrm{poly}(d)}{\varepsilon}$
  - outperforms vanilla DDPM (iteration complexity: d/ε^2)

    — *Chen et al. '23, Li et el. 23, Benton et al. '23*
- **general data distributions**
- ℓ_2 score error assumption suffices (no need of Jacobians)

Main result: accelerated stochastic sampler

Theorem 3 (Li, Huang, Efimov, Wei, Chi, Chen '24)

The accelerated stochastic sampler (3) *obeys (up to log factor)*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \sqrt{\mathsf{KL}(p_{X_1} \parallel p_{Y_1})} \lesssim \frac{d^3}{T} + \sqrt{d}\varepsilon_{\mathsf{score}}$$

- **iteration complexity**: $\frac{\mathrm{poly}(d)}{\varepsilon}$
  - outperforms vanilla DDPM (iteration complexity: d/ε^2)
    — *Chen et al. '23, Li et el. 23, Benton et al. '23*
- **general data distributions**
- ℓ_2 score error assumption suffices (no need of Jacobians)
- **intuition:** higher-order approx. of $p_{X_{t-1}|X_t}$ via simply adding noise

Interpretation via higher-order approximation

characterizing $p_{X_{t-1}|X_t=x_t} \approx \mathcal{N}\big(\mu_t^{\star}(x_t), \Sigma_t^{\star}(x_t)\big)$

Interpretation via higher-order approximation

characterizing $p_{X_{t-1}|X_t=x_t} \approx \mathcal{N}\big(\mu_t^\star(x_t), \Sigma_t^\star(x_t)\big)$

- $\mu_t^\star(x_t) = \frac{1}{\sqrt{\alpha_t}}\big(x_t + (1-\alpha_t)\, s_t^\star(x_t)\big)$

Interpretation via higher-order approximation

characterizing $p_{X_{t-1}|X_t=x_t} \approx \mathcal{N}\left(\mu_t^\star(x_t), \Sigma_t^\star(x_t)\right)$

- $\mu_t^\star\left(x_t\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t + \left(1-\alpha_t\right) s_t^\star\left(x_t\right)\right)$
- $\Sigma_t^\star(x_t) = \frac{1-\alpha_t}{\alpha_t} \underbrace{\left(I + \frac{1-\alpha_t}{2}\frac{\partial s_t^\star}{\partial X}\left(x_t\right)\right)\left(I + \frac{1-\alpha_t}{2}\frac{\partial s_t^\star}{\partial X}\left(x_t\right)\right)^\top}_{\text{DDPM analysis uses simpler approx } I \text{ (Li et al., 2023)}}$

Interpretation via higher-order approximation

characterizing $p_{X_{t-1}|X_t=x_t} \approx \mathcal{N}\big(\mu_t^\star(x_t), \Sigma_t^\star(x_t)\big)$

- $\mu_t^\star(x_t) = \frac{1}{\sqrt{\alpha_t}}\big(x_t + (1-\alpha_t)\, s_t^\star(x_t)\big)$
- $\Sigma_t^\star(x_t) = \frac{1-\alpha_t}{\alpha_t} \underbrace{\left(I + \frac{1-\alpha_t}{2}\frac{\partial s_t^\star}{\partial X}(x_t)\right)\left(I + \frac{1-\alpha_t}{2}\frac{\partial s_t^\star}{\partial X}(x_t)\right)^\top}_{\text{DDPM analysis uses simpler approx } I \text{ (Li et al., 2023)}}$

constructing $p_{Y_{t-1}|Y_t=x_t} \approx \mathcal{N}\big(\mu_t^\star(x_t), \Sigma_t^\star(x_t)\big)$**:**

Interpretation via higher-order approximation

characterizing $p_{X_{t-1}|X_t=x_t} \approx \mathcal{N}(\mu_t^\star(x_t), \Sigma_t^\star(x_t))$

- $\mu_t^\star(x_t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t + (1-\alpha_t)\, s_t^\star(x_t)\right)$
- $\Sigma_t^\star(x_t) = \frac{1-\alpha_t}{\alpha_t} \underbrace{\left(I + \frac{1-\alpha_t}{2}\frac{\partial s_t^\star}{\partial X}(x_t)\right)\left(I + \frac{1-\alpha_t}{2}\frac{\partial s_t^\star}{\partial X}(x_t)\right)^\top}_{\text{DDPM analysis uses simpler approx } I \text{ (Li et al., 2023)}}$

constructing $p_{Y_{t-1}|Y_t=x_t} \approx \mathcal{N}(\mu_t^\star(x_t), \Sigma_t^\star(x_t))$**:**

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Bigg(\underbrace{Y_t + \sqrt{\tfrac{1-\alpha_t}{2}}Z_t}_{\Phi(Y_t, Z_t)} + \sqrt{\tfrac{1-\alpha_t}{2}}Z_t^+ + (1-\alpha_t)\underbrace{\left(s_t^\star(Y_t) - \sqrt{\tfrac{1-\alpha_t}{2}}\tfrac{\partial s_t^\star}{\partial X}(Y_t)Z_t\right)}_{\text{first-order } \approx s_t^\star\left(\Phi(Y_t, Z_t)\right)}\Bigg)$$

Interpretation via higher-order approximation

characterizing $p_{X_{t-1}|X_t=x_t} \approx \mathcal{N}\big(\mu_t^\star(x_t), \Sigma_t^\star(x_t)\big)$

- $\mu_t^\star(x_t) = \frac{1}{\sqrt{\alpha_t}}\big(x_t + (1-\alpha_t)\, s_t^\star(x_t)\big)$
- $\Sigma_t^\star(x_t) = \frac{1-\alpha_t}{\alpha_t} \underbrace{\left(I + \frac{1-\alpha_t}{2}\frac{\partial s_t^\star}{\partial X}(x_t)\right)\left(I + \frac{1-\alpha_t}{2}\frac{\partial s_t^\star}{\partial X}(x_t)\right)^\top}_{\text{DDPM analysis uses simpler approx } I \text{ (Li et al., 2023)}}$

constructing $p_{Y_{t-1}|Y_t=x_t} \approx \mathcal{N}\big(\mu_t^\star(x_t), \Sigma_t^\star(x_t)\big)$**:**

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Bigg(\underbrace{Y_t + \sqrt{\tfrac{1-\alpha_t}{2}}Z_t}_{\Phi(Y_t,Z_t)} + \sqrt{\tfrac{1-\alpha_t}{2}}Z_t^+ + (1-\alpha_t)\underbrace{\Big(s_t^\star(Y_t) - \sqrt{\tfrac{1-\alpha_t}{2}}\tfrac{\partial s_t^\star}{\partial X}(Y_t)Z_t\Big)}_{\text{first-order } \approx s_t^\star(\Phi(Y_t,Z_t))}\Bigg)$$

$$\approx \Psi_t\big(\Phi_t(Y_t,Z_t), Z_t^+\big) \qquad \text{(Ours)}$$

Concluding remarks

- Non-asymptotic theory for probability flow ODE
- New schemes via higher-order approximation to achieve provable acceleration in score-based diffusion models

Concluding remarks

- Non-asymptotic theory for probability flow ODE
- New schemes via higher-order approximation to achieve provable acceleration in score-based diffusion models

Future directions:

- better dependency on problem dimension d?
- acceleration via higher-order ODE/SDE?
  DPM-Solver-3 (third-order ODE)
- end-to-end theory to account for score learning + sampling?

Papers:

"Towards non-asymptotic convergence for diffusion-based generative models," G. Li, Y. Wei, Y. Chen, Y. Chi, arXiv:2306.09251, ICLR 2024

"Accelerating convergence of score-based diffusion models, provably," G. Li*, Y. Huang*, T. Efimov, Y. Wei, Y. Chi, Y. Chen, arXiv:2403.03852, 2024 (*=equal contributions)