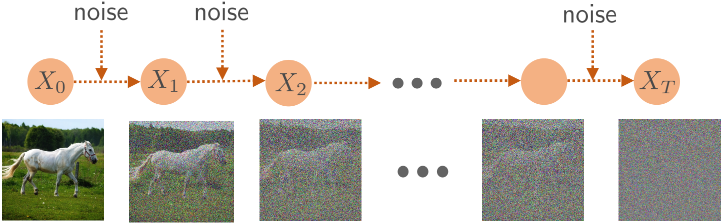
Harnessing Low Dimensionality in Diffusion Models: From Theory to Practice

Lecture II: Sampling Theory for Diffusion Models

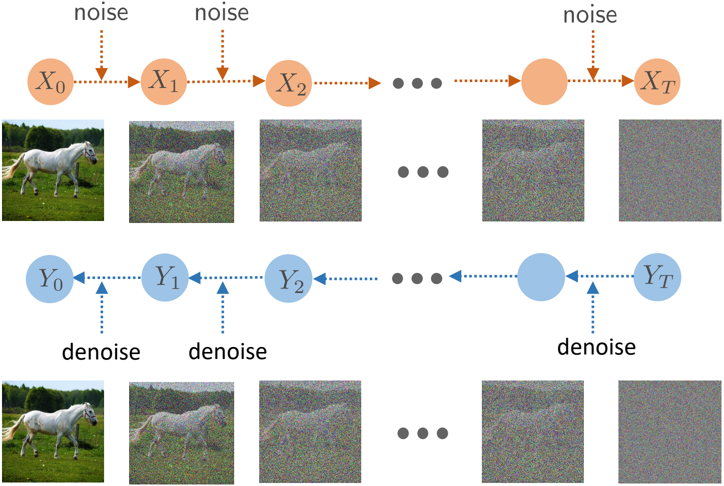
Yuxin Chen, Qing Qu, Liyue Shen

International Conference on Machine Learning (ICML) 2025

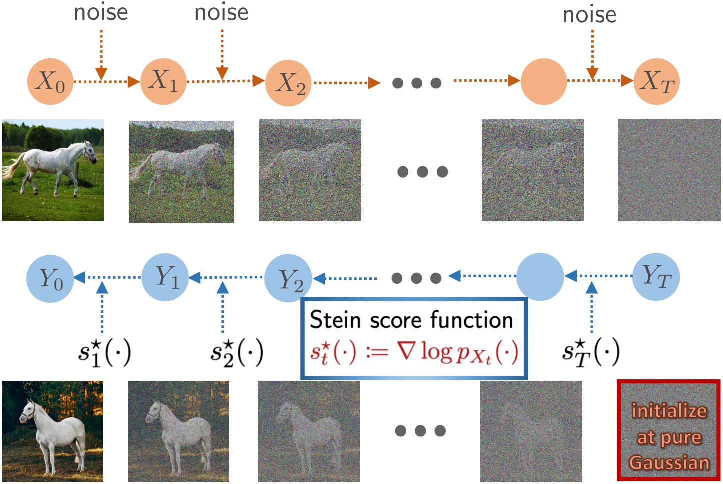
Warton Statistics and Data Science, University of Pennsylvania



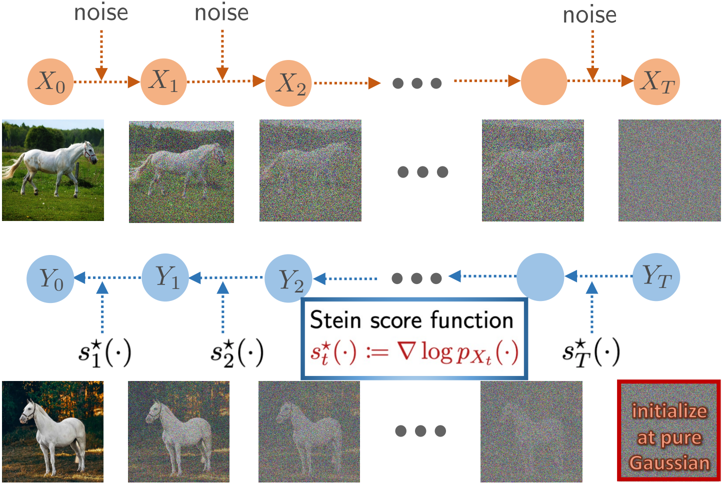
- **forward process:** diffuse data into noise



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



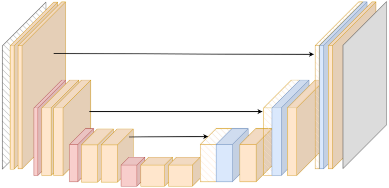
- **forward process:** diffuse data into noise
- **reverse process:** convert pure noise into data-like distributions



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

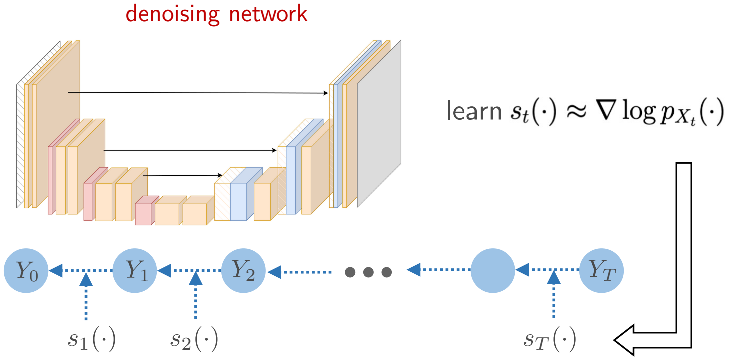
Goal: $Y_t \overset{\mathrm{d}}{\approx} X_t, \quad t = T, \cdots, 1$

denoising network



learn $s_t(\cdot) \approx \nabla \log p_{X_t}(\cdot)$

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



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 $\{s_t(\cdot)\}$

Towards mathematical theory for diffusion models

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

Towards mathematical theory for diffusion models

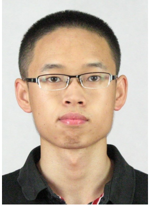
- hard to develop full-fledged **end-to-end** theory

- "divide-and-conquer": score learning $\overset{\text{✄ decouple}}{\longleftarrow \text{✗} \longrightarrow}$ sampling

Outline of Lecture II

1. non-asymptotic convergence theory
2. adaptation to (unknown) low dimensionality
3. acceleration via higher-order approximation

Part 1: nonasymptotic convergence theory

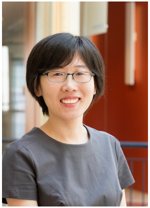


Gen Li

CUHK

Yuting Wei

UPenn



Yuejie Chi

Yale

Two mainstream approaches

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

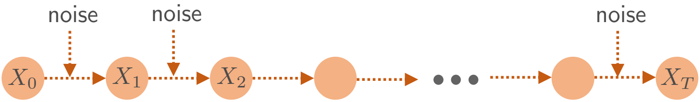
Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

Denoising Diffusion Implicit Models

Jiaming Song, Chenlin Meng & Stefano Ermon
Stanford University
{tsong,chenlin,ermon}@cs.stanford.edu

DDPM vs. DDIM

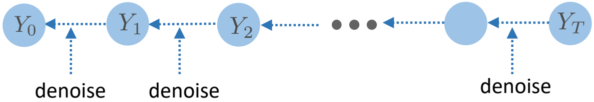


forward process: $X_0 \sim p_{\text{data}}$ (target distribution)

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

- $\beta_t := 1 - \alpha_t$ controls variance of injected noise

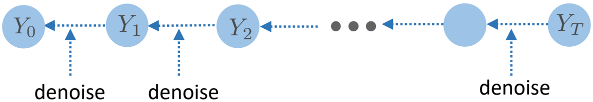
DDPM vs. DDIM



— *Ho, Jain, Abbeel '20*

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

DDPM vs. DDIM



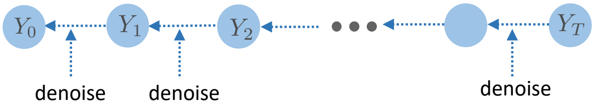
— *Ho, Jain, Abbeel '20*

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

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

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(\underbrace{Y_t + \eta_t^{\mathsf{ddpm}} s_t(Y_t)}_{\text{deterministic}} + \underbrace{\sigma_t^{\mathsf{ddpm}} \mathcal{N}(0, I_d)}_{\text{stochastic}}\Big), \quad t = T, \cdots, 1$$

DDPM vs. DDIM

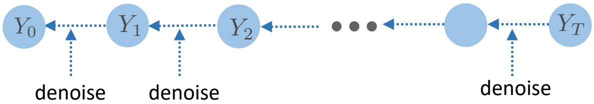


— *Song, Meng, Ermon '20*

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

2. A deterministic sampler: $\underbrace{\textbf{denoising diffusion implicit models}}_{\text{DDIM; or probability flow ODE}}$

DDPM vs. DDIM



— *Song, Meng, Ermon '20*

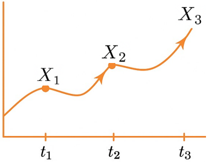
— *Song, Sohl-Dickstein, Kingma, Kumar, Ermon, Poole '20*

2. A deterministic sampler: $\underbrace{\textbf{denoising diffusion implicit models}}_{\text{DDIM; or probability flow ODE}}$

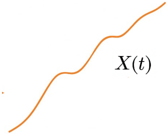
$$Y_T \sim \mathcal{N}(0, I_d)$$

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(\underbrace{Y_t + \eta_t^{\text{ddim}} s_t(Y_t)}_{\text{deterministic}}\Big), \quad t = T, \cdots, 1$$

Interpretations from lens of SDE/ODE

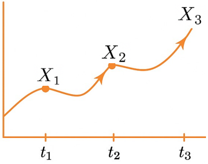


discrete-time

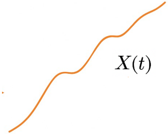


continuous-time

Interpretations from lens of SDE/ODE



discrete-time



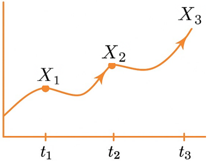
continuous-time

forward process

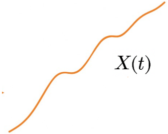
$$X_t = \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d)$$

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

Interpretations from lens of SDE/ODE



discrete-time



continuous-time

forward process

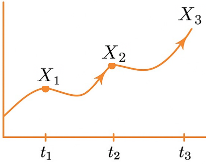
$$X_t = \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t}\mathcal{N}(0, I_d)$$

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

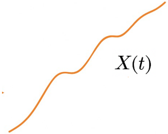
- $\exists$ reverse-time SDE w/ same path distribution (Anderson '82)

$$\mathrm{d}Y_t = \big(Y_t + 2 s^{\star}_{T-t}(Y_t)\big)\,\beta(T-t)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}W_t$$

Interpretations from lens of SDE/ODE



discrete-time



continuous-time

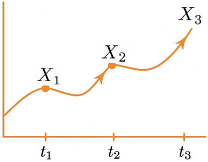
forward process

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

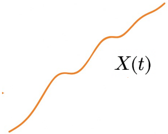
- $\exists$ reverse-time SDE w/ same path distribution (Anderson '82)

$\xrightarrow{\text{time discretization}}$ DDPM

Interpretations from lens of SDE/ODE



discrete-time



continuous-time

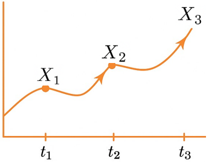
forward process

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

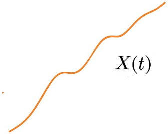
- $\exists$ reverse-time ODE w/ same *marginal* dist (Song et al. '20)

$$
\mathrm{d}Y_t = \big(Y_t + s^{\star}_{T-t}(Y_t)\big)\,\beta(T-t)\mathrm{d}t
$$

Interpretations from lens of SDE/ODE



discrete-time



continuous-time

forward process

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

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

- $\exists$ reverse-time ODE w/ same *marginal* dist (Song et al. '20)

$$\xrightarrow{\text{time discretization}} \text{DDIM}$$

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

— *almost no restriction on target data distributions*

Key takeaway: in continuous-time limits, sampling is feasible once exact 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

A small sample of convergence theory

- Lee, Lu, Tan '22
- Chen, Chewi, Li, Li, Salim, Zhang '22
- Chen, Lee, Lu '22
- Lee, Lu, Tan '23
- Chen, Daras, Dimakis '23
- Chen, Chewi, Lee, Li, Lu, Salim '23
- Benton, De Bortoli, Doucet, Deligiannidis '23
- Li, Wei, Chen, Chi '23
- Benton, Deligiannidis, Doucet '23
- Cheng, Lu, Tan, Xie '23
- Tang '23
- Li, Wei, Chi, Chen '24
- Li, Yan '24a, '24b
- Azangulov, Deligiannidis, Rousseau '24
- Potaptchik, Azangulov, Deligiannidis '24
- Huang, Wei, Chen '24
- Gao, Zhu '24
- Huang, Huang, Lin '24
- Li, Jiao '24
- Li, Di, Gu '24
- Liang, Ju, Liang, Shroff '24
- Tang, Zhao '24
- Liang, Huang, Chen '25
- Li, Cai, Wei '25
- Tang, Yan '25
- Yu, Yu '25
- Gentiloni-Silveri, Ocello '25
- . . .

Assumptions: target data distribution p_{data}

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

Assumptions: target data distribution p_{data}

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

- support size can be very large

Assumptions: target data distribution p_{data}

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

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

Assumptions: target data distribution p_{data}

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

- support size can be very large
- very general: *no need of log-concavity, smoothness, etc*
- can also be replaced by $\mathbb{E}[\|X_0\|_2] \le T^{c_M}$ for large const c_M

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

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

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

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

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

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

  - 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
  - suffices for DDPM but not DDIM (counterexample in Li et al. '24)

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
  - suffices for DDPM but not DDIM (counterexample in Li et al. '24)

- *Jacobian estimation error (for DDIM only):*

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

Assumptions: learning rates

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

- **learning rates:** for some consts $c_0, c_1 > 0$,

$$\begin{aligned} 1-\alpha_1 &= \frac{1}{T^{c_0}} \\ 1-\alpha_t &= \underbrace{\frac{c_1 \log T}{T} \min\left\{(1-\alpha_1)\Big(1+\frac{c_1\log T}{T}\Big)^t,\ 1\right\}}_{\text{2 phases: exp growth} \rightarrow \text{flat}} \end{aligned}$$

A glimpse of convergence theory (up to log factor)

DDPM:

$$\mathsf{KL}\big(p_{X_1} \,\|\, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}}^2 \quad \text{(Benton et al. '23)}$$

$$\mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}} \quad \text{(Li, Yan '24)}$$

A glimpse of convergence theory (up to log factor)

DDPM:

$$\mathsf{KL}\big(p_{X_1} \,\|\, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}}^2$$ (Benton et al. '23)

$$\mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}}$$ (Li, Yan '24)

- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield KL} \leq \varepsilon^2}$ d/ε^2 (assuming accurate scores)
- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield TV} \leq \varepsilon}$ d/ε (assuming accurate scores)

A glimpse of convergence theory (up to log factor)

DDPM:
$$\mathsf{KL}\big(p_{X_1} \,\|\, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}}^2 \quad \text{(Benton et al. '23)}$$
$$\mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}} \quad \text{(Li, Yan '24)}$$

- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield KL} \leq \varepsilon^2}$ d/ε^2 (assuming accurate scores)
- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield TV} \leq \varepsilon}$ d/ε (assuming accurate scores)
    - Pinsker inequality $(\mathsf{TV} \leq \sqrt{\frac{1}{2}\mathsf{KL}})$ is loose when bounding TV

A glimpse of convergence theory (up to log factor)

DDPM: $\mathsf{KL}\big(p_{X_1} \,\|\, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}}^2$ (Benton et al. '23)

$\mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}}$ (Li, Yan '24)

- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield KL} \leq \varepsilon^2}$ d/ε^2 (assuming accurate scores)
- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield TV} \leq \varepsilon}$ d/ε (assuming accurate scores)
- **stability:** degrades gracefully as $\varepsilon_{\mathsf{score}} \uparrow$

A glimpse of convergence theory (up to log factor)

DDPM:
$$\mathsf{KL}\big(p_{X_1} \,\|\, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}}^2$$ (Benton et al. '23)
$$\mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\mathsf{score}}$$ (Li, Yan '24)

- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield KL} \leq \varepsilon^2}$ d/ε^2 (assuming accurate scores)
- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield TV} \leq \varepsilon}$ d/ε (assuming accurate scores)
- **stability:** degrades gracefully as $\varepsilon_{\mathsf{score}} \uparrow$ and $\varepsilon_{\mathsf{Jacobi}} \uparrow$
- **general data distribution:** no need of smoothness, log-concavity

A glimpse of convergence theory (up to log factor)

DDPM: $\mathsf{KL}\big(p_{X_1} \,\|\, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\text{score}}^2$ (Benton et al. '23)

$\mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim d/T + \varepsilon_{\text{score}}$ (Li, Yan '24)

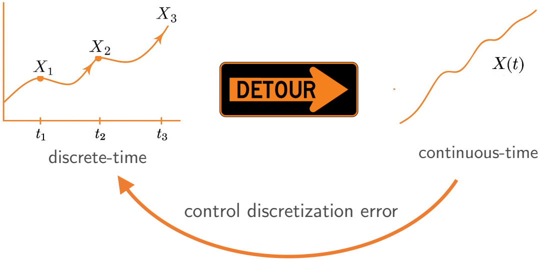
DDIM: $\mathsf{TV}\big(p_{X_1}, p_{Y_1}\big) \lesssim d/T + \sqrt{d}\,\varepsilon_{\text{score}} + d\varepsilon_{\text{Jacobi}}$ (Li et al. '24)

- ~~$\underbrace{\textbf{iteration complexity:}}_{\text{to yield KL} \le \varepsilon^2}$ d/ε^2 (assuming accurate scores)~~
- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield TV} \le \varepsilon}$ d/ε (assuming accurate scores)
- **stability:** degrades gracefully as $\varepsilon_{\text{score}} \uparrow$ and $\varepsilon_{\text{Jacobi}} \uparrow$
- **general data distribution:** no need of smoothness, log-concavity

Analysis strategy # 1 (for DDPM)

— Chen, Chewi, Li, Li, Salim, Zhang '22, Chen, Lee, Lu '22, Tang, Zhao '24

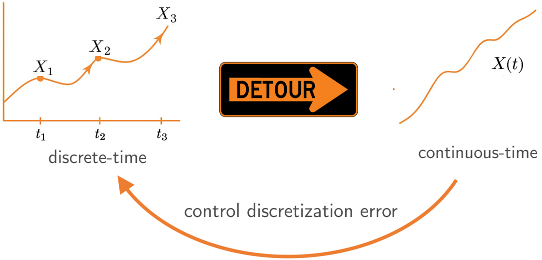
— Benton, De Bortoli, Doucet, Deligiannidis '23, Huang, Wei, Chen '24



Analysis strategy # 1 (for DDPM)

— Chen, Chewi, Li, Li, Salim, Zhang '22, Chen, Lee, Lu '22, Tang, Zhao '24

— Benton, De Bortoli, Doucet, Deligiannidis '23, Huang, Wei, Chen '24



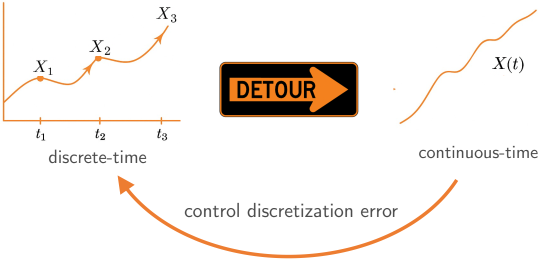
Analogy: (stochastic) gradient descent vs. gradient flow, TD learning via ODE

yields state-of-the-art KL-based theory for DDPM!

Analysis strategy # 1 (for DDPM)

— Chen, Chewi, Li, Li, Salim, Zhang '22,   Chen, Lee, Lu '22,   Tang, Zhao '24

— Benton, De Bortoli, Doucet, Deligiannidis '23, Huang, Wei, Chen '24



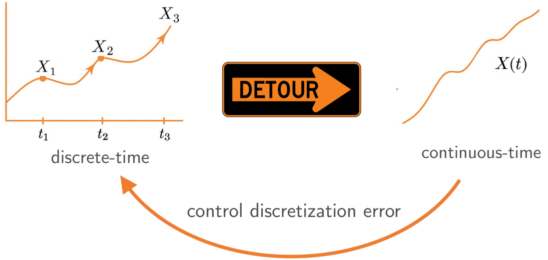
2 key steps:

- apply change of measure (e.g. Girsanov thm) to show

$$\mathsf{KL}\big(P^{\mathsf{true}} \,\|\, P^{\mathsf{ddpm}}\big) \le \int w(t) \underbrace{\mathbb{E}\Big[\big\|\mathsf{drift}^{\mathsf{true}}(t) - \mathsf{drift}^{\mathsf{ddpm}}(t)\big\|\Big]^2}_{\text{score error + discretization error}} \mathrm{d}t + \mathsf{small\text{-}term}$$

Analysis strategy # 1 (for DDPM)

— Chen, Chewi, Li, Li, Salim, Zhang '22, Chen, Lee, Lu '22, Tang, Zhao '24
— Benton, De Bortoli, Doucet, Deligiannidis '23, Huang, Wei, Chen '24



2 key steps:

- leverage stochastic localization to chacracterize

$$\text{discretization error} \overset{\text{link}}{\longleftrightarrow} \mathbb{E}\big[\mathrm{Cov}\big(X_0 \mid X_t\big)\big]$$

Analysis strategy # 2 (for DDIM & DDPM)

— Li, Wei, Chen, Chi '24, Li, Wei, Chi, Chen '24

— Li, Yan '24, Liang, Huang, Chen '25

Tackle discrete-time process directly & track changes of TV distance

Analysis strategy # 2 (for DDIM & DDPM)

— Li, Wei, Chen, Chi '24, Li, Wei, Chi, Chen '24
— Li, Yan '24, Liang, Huang, Chen '25

Tackle discrete-time process directly & track changes of TV distance

yields state-of-the-art TV-based theory for DDIM & DDPM!

Analysis strategy # 2 (for DDIM & DDPM)

— Li, Wei, Chen, Chi '24, Li, Wei, Chi, Chen '24
— Li, Yan '24, Liang, Huang, Chen '25

Tackle discrete-time process directly & track changes of TV distance

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

Analysis strategy # 2 (for DDIM & DDPM)

— Li, Wei, Chen, Chi '24, Li, Wei, Chi, Chen '24
— Li, Yan '24, Liang, Huang, Chen '25

Tackle discrete-time process directly & track changes of TV distance

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

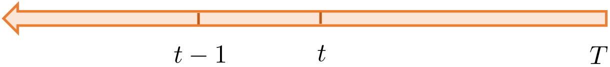
Analysis strategy # 2 (for DDIM & DDPM)

— Li, Wei, Chen, Chi '24, Li, Wei, Chi, Chen '24

— Li, Yan '24, Liang, Huang, Chen '25

Tackle discrete-time process directly & track changes of TV distance

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

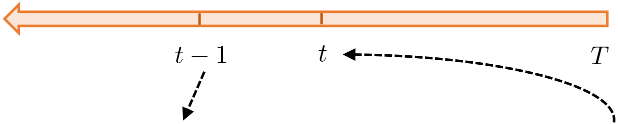


Analysis strategy # 2 (for DDIM & DDPM)

— Li, Wei, Chen, Chi '24, Li, Wei, Chi, Chen '24
— Li, Yan '24, Liang, Huang, Chen '25

Tackle discrete-time process directly & track changes of TV distance

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



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

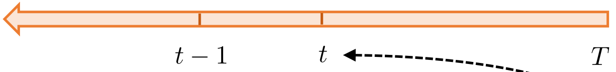
Analysis strategy # 2 (for DDIM & DDPM)

— Li, Wei, Chen, Chi '24, Li, Wei, Chi, Chen '24

— Li, Yan '24, Liang, Huang, Chen '25

Tackle discrete-time process directly & track changes of TV distance

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



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

change of variables
(Jacobian transformation)

concentration bounds
$+$ stochastic localization

Part 2: adaptation to (unknown) low dimensionality

Jiadong Liang

UPenn

Zhihan Huang

UPenn

Yuting Wei

UPenn

Recap: theory for mainstream diffusion models

Denoising Diffusion Probabilistic Models

Jonathan Ho UC Berkeley jonathanho@berkeley.edu
Ajay Jain UC Berkeley ajayj@berkeley.edu
Pieter Abbeel UC Berkeley pabbeel@cs.berkeley.edu

DENOISING DIFFUSION IMPLICIT MODELS

Jiaming Song, Chenlin Meng & Stefano Ermon
Stanford University
{tsong,chenlin,ermon}@cs.stanford.edu

Theorem (Li, Wei, Chi, Chen '24, Li, Yan '24)

With perfect scores, both DDIM & DDPM yield TV$(p_{X_1}, p_{Y_1}) \leq \varepsilon$ in

$$\widetilde{O}(d/\varepsilon) \text{ iterations}$$

- d: ambient dimension

d/ε iterations are too slow . . .

ImageNet: $d = 150,528$ pixels per image

d/ε iterations are too slow . . .

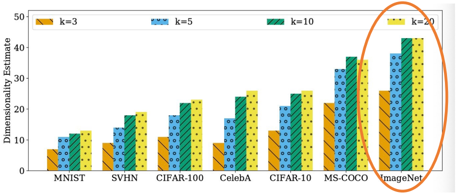
ImageNet: $d = 150,528$ pixels per image (so $\frac{d}{\varepsilon} > 10^6$ for moderate ε)

d/ε iterations are too slow . . .

ImageNet: $d = 150,528$ pixels per image (so $\frac{d}{\varepsilon} > 10^6$ for moderate ε)

In practice, DDIM/DDPM yield good samples in <u>hundreds (or tens)</u> of iterations . . .

d/ε iterations are too slow ...



ImageNet: $d = 150,528$ pixels per image (so $\frac{d}{\varepsilon} > 10^6$ for moderate ε)

$k = 43$ intrinsic dimension (Pope et al. '21)

In practice, DDIM/DDPM yield good samples in hundreds (or tens) of iterations ...

Can diffusion models adapt to intrinsic low dimensionality?

Intrinsic dimension

The target distribution p_{data} is said to have intrinsic dimension k if

$$\log \underbrace{N^{\mathsf{cover}}(\mathsf{support}(p_{\mathsf{data}}), \|\cdot\|_2, \varepsilon_0)}_{\text{covering number of support of } p_{\mathsf{data}}} \lesssim k \log \frac{1}{\varepsilon_0}$$

Intrinsic dimension

The target distribution p_{data} is said to have intrinsic dimension k if

$$\log \underbrace{N^{\mathsf{cover}}(\mathsf{support}(p_{\mathsf{data}}), \|\cdot\|_2, \varepsilon_0)}_{\text{covering number of support of } p_{\mathsf{data}}} \lesssim k \log \frac{1}{\varepsilon_0}$$

- k-dimensional linear subspace
- low-dimensional manifold
- doubling dimension, Minkowski dimension
- ...

— *see Huang, Wei, Chen '24*

Assumptions

- **minimal data assumptions:**

$$\mathbb{P}\Big(\|X_0\|_2 \le \underbrace{T^{c_R}}_{\text{polynomially large diameter}}\Big) = 1$$

  for arbitrarily large constant $c_R > 0$

- **perfect score estimates:** $s_t(\cdot) = \nabla \log p_{X_t}(\cdot)$
  $\longrightarrow$ not needed; only to simplify presentation

Convergence theory in total variation

Theorem (Liang, Huang, Chen '24)

Both DDPM $\&$ DDIM (in original form) yield $\mathsf{TV}(p_{X_1}, p_{Y_1}) \le \varepsilon$ in

$\widetilde{O}(k/\varepsilon)$ iterations

— *concurrent work Li, Yan '25, Tang, Yan '25*

DDIM: $$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t^{\mathsf{ddim}} s_t(Y_t)\big)$$

DDPM: $$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t^{\mathsf{ddpm}} s_t(Y_t) + \sigma_t^{\mathsf{ddpm}} \mathcal{N}(0, I_d)\big)$$

Convergence theory in total variation

Theorem (Liang, Huang, Chen ’24)

Both DDPM & DDIM (in original form) yield $\mathsf{TV}(p_{X_1}, p_{Y_1}) \leq \varepsilon$ in

$\widetilde{O}(k/\varepsilon)$ iterations

— *concurrent work Li, Yan ’25, Tang, Yan ’25*

$$\text{DDIM:} \quad Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(Y_t + \frac{1-\alpha_t}{1+\sqrt{\frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}}} s_t(Y_t)\right)$$

$$\text{DDPM:} \quad Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(Y_t + (1-\alpha_t)s_t(Y_t) + \sqrt{\frac{(1-\alpha_t)(\alpha_t-\overline{\alpha}_t)}{1-\overline{\alpha}_t}}\mathcal{N}(0, I_d)\right)$$

where $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$

Convergence theory in total variation

Theorem (Liang, Huang, Chen '24)

Both DDPM & DDIM (in original form) yield $\mathsf{TV}(p_{X_1}, p_{Y_1}) \le \varepsilon$ in

$\widetilde{O}(k/\varepsilon)$ iterations

— *concurrent work Li, Yan '25, Tang, Yan '25*

DDIM: $$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(Y_t + \frac{1-\alpha_t}{1+\sqrt{\frac{\alpha_t - \overline{\alpha}_t}{1-\overline{\alpha}_t}}} s_t(Y_t)\right)$$

DDPM: $$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(Y_t + (1-\alpha_t)s_t(Y_t) + \sqrt{\frac{(1-\alpha_t)(\alpha_t - \overline{\alpha}_t)}{1-\overline{\alpha}_t}}\mathcal{N}(0, I_d)\right)$$

where $\overline{\alpha}_t = \prod_{i=1}^t \alpha_i$

- *originally derived to optimize variational lower bounds!*

Convergence theory in KL divergence

Theorem (Huang, Wei, Chen '24)

DDPM sampler (its original form) yields $\mathsf{KL}(p_{X_1} \parallel p_{Y_1}) \le \varepsilon$ in

$$\widetilde{O}(k/\varepsilon) \quad \text{iterations}$$

— prior work Li and Yan '24, Azangulov et al. '24

— concurrent work Potaptchik et al.'24

Convergence theory in KL divergence

Theorem (Huang, Wei, Chen '24)

DDPM sampler (its original form) yields KL$(p_{X_1} \parallel p_{Y_1}) \leq \varepsilon$ in

$\widetilde{O}(k/\varepsilon)$ iterations

— prior work Li and Yan '24, Azangulov et al. '24

— concurrent work Potaptchik et al.'24

- optimal scaling in k

Interpretation from lens of SDE/ODE

reverse-time SDE (same distribution as X_t):

$$\mathrm{d}Y_t = \big(Y_t + 2s^{\star}_{T-t}(Y_t)\big)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

Interpretation from lens of SDE/ODE

reverse-time SDE (same distribution as X_t):

$$\mathrm{d}Y_t = \big(Y_t + 2s^{\star}_{T-t}(Y_t)\big)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

Tweedie's formula $\Downarrow$

$$\mathrm{d}Y_t = \big(\underbrace{c_{1,t}Y_t}_{\text{linear drift}} + c_{2,t}\underbrace{\mathbb{E}[X_0 \mid X_{T-t} = Y_t]}_{\text{cond. mean of } X_0}\big)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

- **key enabler:** $\mathbb{E}[X_0 \mid X_t]$ is "projection" onto low-dimensional structure

Interpretation from lens of SDE/ODE

reverse-time SDE (same distribution as X_t):

$$\mathrm{d}Y_t = \big(Y_t + 2s^{\star}_{T-t}(Y_t)\big)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

Tweedie's formula $\Downarrow$

$$\mathrm{d}Y_t = \Big(\underbrace{c_{1,t}Y_t}_{\text{linear drift}} + c_{2,t}\underbrace{\mathbb{E}[X_0 \mid X_{T-t} = Y_t]}_{\text{cond. mean of } X_0}\Big)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

time-discretize $\Downarrow$

DDIM or DDPM

- **key enabler:** $\mathbb{E}[X_0 \mid X_t]$ is "projection" onto low-dimensional structure
- **discretization scheme matters:** time-discretize carefully to retain low-dimensional adaptation

Crucial choices of coefficients: a lower bound

DDPM $\&$ DDIM updates take the form

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t s_t(Y_t) + \sigma_t \mathcal{N}(0, I_d)\big)$$

Crucial choices of coefficients: a lower bound

DDPM $\&$ DDIM updates take the form

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t s_t(Y_t) + \sigma_t \mathcal{N}(0, I_d)\big)$$

Theorem (Liang, Huang, Chen '24)

Even when starting from $X_t = Y_t$, one can have

$$\mathsf{TV}(X_{t-1}, Y_{t-1}) \gtrsim \sqrt{d} \cdot \left| \frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 + \frac{\sigma_t^2}{\alpha_t - \overline{\alpha}_t} - 1 \right|$$

Crucial choices of coefficients: a lower bound

DDPM $\&$ DDIM updates take the form

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t s_t(Y_t) + \sigma_t \mathcal{N}(0, I_d)\big)$$

Theorem (Liang, Huang, Chen '24)

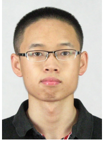
Even when starting from $X_t = Y_t$, one can have

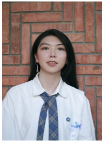
$$\mathsf{TV}(X_{t-1}, Y_{t-1}) \gtrsim \sqrt{d} \cdot \left| \frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 + \frac{\sigma_t^2}{\alpha_t - \overline{\alpha}_t} - 1 \right|$$

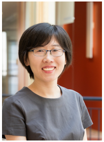
To avoid scaling in d, one needs orange term ≈ 0

- both DDIM and DDPM satisfy this!

Part 3: acceleration via higher-order approximation







Gen Li* **CUHK**

Yuchen Zhou* **UIUC**

Yu Huang* **UPenn**

Timofey Efimov CMU

Yuting Wei UPenn

Yuejie Chi Yale

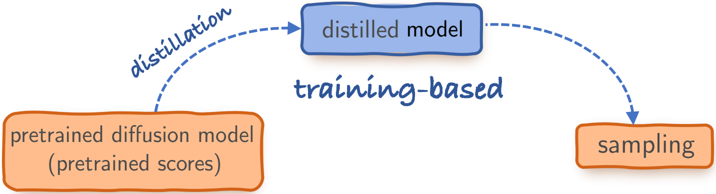
DDPM and DDIM are still slow . . .

— *Song, Meng, Ermon '20*

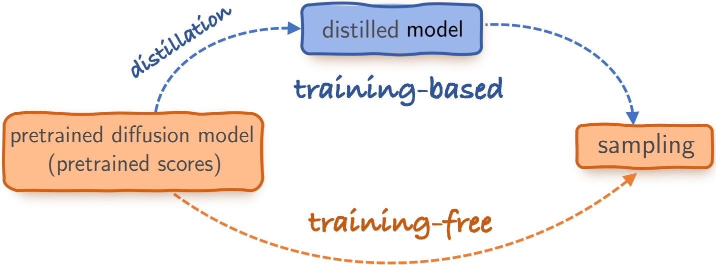
DDPM and DDIM are still slow . . .

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

— *Song, Meng, Ermon '20*



- **training-based:** $\underbrace{\text{distill pre-trained diffusion model}}_{\text{requires additional training}}$ into another model that can be executed rapidly
  - e.g., progressive distillation, consistency model



- **training-free:** directly invoke pre-trained 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 sampler that is <u>provably faster</u> than DDIM/DDPM?

Discretization $\longleftrightarrow$ approximation

A starting point: equiv solution to probability flow ODE

$$\underbrace{Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}}}_{\text{represent } Y_{t-1}} = \frac{1}{\sqrt{\alpha_t}} \underbrace{Y^{\mathsf{ode}}_{\overline{\alpha}_t}}_{\text{represent } Y_t} + \underbrace{\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} s^\star_\gamma\big(Y^{\mathsf{ode}}_\gamma\big) \mathrm{d}\gamma}$$

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

Discretization $\longleftrightarrow$ approximation

A starting point: equiv solution to probability flow ODE

$$\underbrace{Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}}}_{\text{represent } Y_{t-1}} = \frac{1}{\sqrt{\alpha_t}} \underbrace{Y^{\mathsf{ode}}_{\overline{\alpha}_t}}_{\text{represent } Y_t} + \underbrace{\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} s^\star_\gamma\big(Y^{\mathsf{ode}}_\gamma\big)\mathrm{d}\gamma}_{\text{infinitely many score evaluations!}}$$

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

Discretization $\longleftrightarrow$ approximation

A starting point: equiv solution to probability flow ODE

$$\underbrace{Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}}}_{\text{represent } Y_{t-1}} = \frac{1}{\sqrt{\alpha_t}} \underbrace{Y^{\mathsf{ode}}_{\overline{\alpha}_t}}_{\text{represent } Y_t} + \underbrace{\int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} s^\star_\gamma\big(Y^{\mathsf{ode}}_\gamma\big)\mathrm{d}\gamma}_{\text{infinitely many score evaluations!}}$$

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

- can we approximate the integral by a few score evals?

$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^{\star}_{\gamma}\left(Y^{\mathsf{ode}}_{\gamma}\right)}_{\text{approximate by?}} \, \mathrm{d}\gamma$$

$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right)}_{\text{approximate by?}} \,\mathrm{d}\gamma$$

1st order approx: $s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right) \approx s^\star_{\overline{\alpha}_t}\left(Y^{\mathsf{ode}}_{\overline{\alpha}_t}\right) \approx s_t(Y_t)$

$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^{\star}_{\gamma}\big(Y^{\mathsf{ode}}_{\gamma}\big)}_{\text{approximate by?}} \, \mathrm{d}\gamma$$

1st order approx: $s^{\star}_{\gamma}\big(Y^{\mathsf{ode}}_{\gamma}\big) \approx s^{\star}_{\overline{\alpha}_t}\big(Y^{\mathsf{ode}}_{\overline{\alpha}_t}\big) \approx s_t(Y_t)$

$$\Longrightarrow \quad Y_{t-1} \approx \frac{1}{\sqrt{\alpha_t}} \left(Y_t + \frac{1-\alpha_t}{2} s_t(Y_t) \right) \qquad \text{original DDIM}$$

$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right)}_{\text{approximate by?}} \mathrm{d}\gamma$$

1st order approx: $s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right) \approx s^\star_{\overline{\alpha}_t}\left(Y^{\mathsf{ode}}_{\overline{\alpha}_t}\right) \approx s_t(Y_t)$

$\frac{d}{\varepsilon}$ iterations; 1 score eval per iteration (DDIM)

$$Y^{\text{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\text{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\left(Y^{\text{ode}}_\gamma\right)}_{\text{approximate by?}} \mathrm{d}\gamma$$

1st order approx: $s^\star_\gamma\left(Y^{\text{ode}}_\gamma\right) \approx s^\star_{\overline{\alpha}_t}\left(Y^{\text{ode}}_{\overline{\alpha}_t}\right) \approx s_t(Y_t)$

$\frac{d}{\varepsilon}$ iterations; 1 score eval per iteration (DDIM)

refined approximation?

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

1st order approx: $s^\star_\gamma\big(Y^{\mathsf{ode}}_\gamma\big) \approx s^\star_{\overline{\alpha}_t}\big(Y^{\mathsf{ode}}_{\overline{\alpha}_t}\big) \approx s_t(Y_t)$

$\frac{d}{\varepsilon}$ iterations; 1 score eval per iteration (DDIM)

refined approximation?

$$s^\star_\gamma\big(Y^{\mathsf{ode}}_\gamma\big) \approx s_t(Y_t) + \frac{\gamma - \overline{\alpha}_t}{\overline{\alpha}_t - \overline{\alpha}_{t+1}} \Big(s_t(Y_t) - s_{t+1}(Y_{t+1})\Big)$$

$$Y^{\text{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\text{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\left(Y^{\text{ode}}_\gamma\right)}_{\text{approximate by?}} \, \mathrm{d}\gamma$$

1st order approx: $s^\star_\gamma\left(Y^{\text{ode}}_\gamma\right) \approx s^\star_{\overline{\alpha}_t}\left(Y^{\text{ode}}_{\overline{\alpha}_t}\right) \approx s_t(Y_t)$

$\frac{d}{\varepsilon}$ iterations; 1 score eval per iteration (DDIM)

2nd order approx: (Li, Huang, Efimov, Wei, Chi, Chen '24)

$$\sqrt{\alpha_t} Y_{t-1} \approx Y_t + \frac{1-\alpha_t}{2} s_t(Y_t) + \frac{(1-\alpha_t)^2}{4(1-\alpha_{t+1})}\left(s_t(Y_t) - \sqrt{\alpha_{t+1}} s_{t+1}(Y_{t+1})\right)$$

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

$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right)}_{\text{approximate by?}} \mathrm{d}\gamma$$

1st order approx: $s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right) \approx s^\star_{\overline{\alpha}_t}\left(Y^{\mathsf{ode}}_{\overline{\alpha}_t}\right) \approx s_t(Y_t)$

$\frac{d}{\varepsilon}$ iterations; 1 score eval per iteration (DDIM)

2nd order approx: (Li, Huang, Efimov, Wei, Chi, Chen '24)

$\frac{\mathrm{poly}(d)}{\sqrt{\varepsilon}}$ iterations; 2 score evals per iteration

$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right)}_{\text{approximate by?}} \mathrm{d}\gamma$$

1st order approx: $s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right) \approx s^\star_{\overline{\alpha}_t}\left(Y^{\mathsf{ode}}_{\overline{\alpha}_t}\right) \approx s_t(Y_t)$

$\frac{d}{\varepsilon}$ iterations; 1 score eval per iteration (DDIM)

even higher-order approximation?

$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right)}_{\text{approximate by?}} \mathrm{d}\gamma$$

1st order approx: $s^\star_\gamma\left(Y^{\mathsf{ode}}_\gamma\right) \approx s^\star_{\overline{\alpha}_t}\left(Y^{\mathsf{ode}}_{\overline{\alpha}_t}\right) \approx s_t(Y_t)$

$\frac{d}{\varepsilon}$ iterations; 1 score eval per iteration (DDIM)

even higher-order approximation? for order K:

$$\frac{1}{\gamma^{3/2}} s^\star_\gamma(Y^{\mathsf{ode}}_\gamma) \approx \sum_{0 \le i < K} \psi_i(\gamma) \frac{s_{\gamma_{t,i}}\left(Y^{\mathsf{ode}}_{\gamma_{t,i}}\right)}{(\gamma_{t,i})^{3/2}}$$

- K anchor points: $\gamma_{t,0}, \ldots, \gamma_{t,K-1}$

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

1st order approx: $s^\star_\gamma\big(Y^{\text{ode}}_\gamma\big) \approx s^\star_{\overline{\alpha}_t}\big(Y^{\text{ode}}_{\overline{\alpha}_t}\big) \approx s_t(Y_t)$

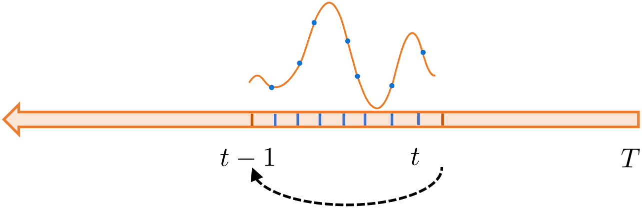
$\frac{d}{\varepsilon}$ iterations; 1 score eval per iteration (DDIM)

even higher-order approximation? for order K:

$$\frac{1}{\gamma^{3/2}} s^\star_\gamma(Y^{\text{ode}}_\gamma) \approx \sum\nolimits_{0 \le i < K} \psi_i(\gamma) \frac{s_{\gamma_{t,i}}(Y^{\text{ode}}_{\gamma_{t,i}})}{(\gamma_{t,i})^{3/2}}$$

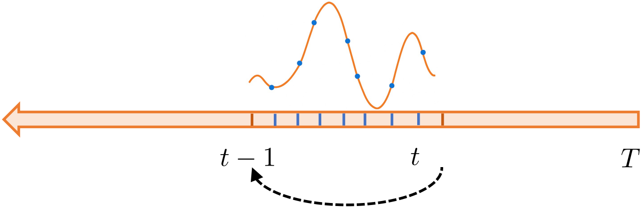
- K anchor points: $\gamma_{t,0}, \dots, \gamma_{t,K-1}$
- Lagrange basis polynomial: $\psi_i(\gamma) := \frac{\prod_{i' : i' \neq i}(\gamma - \gamma_{t,i'})}{\prod_{i' : i' \neq i}(\gamma_{t,i} - \gamma_{t,i'})}$

Proposed K-th order sampler (Li, Zhou, Wei, Chen '25)



$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^{\star}_{\gamma}\big(Y^{\mathsf{ode}}_{\gamma}\big)}_{\text{approx by deg-}(K-1)\text{ Lagrange polynomials}} \mathrm{d}\gamma$$

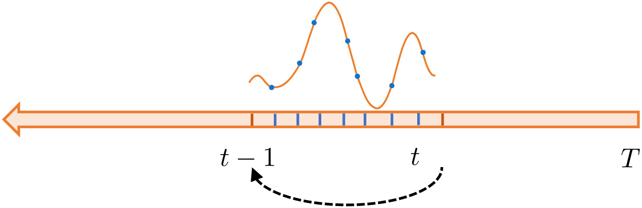
Proposed K-th order sampler (Li, Zhou, Wei, Chen ’25)



$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^{\star}_{\gamma}\left(Y^{\mathsf{ode}}_{\gamma}\right)}_{\text{approx by deg-}(K-1)\text{ Lagrange polynomials}} \mathrm{d}\gamma$$

- successively, alternately refine $Y^{\mathsf{ode}}_{\gamma_{t,i}}$ and $s_{\gamma_{t,i}}\left(Y^{\mathsf{ode}}_{\gamma_{t,i}}\right)$

Proposed K-th order sampler (Li, Zhou, Wei, Chen '25)



$$Y^{\mathsf{ode}}_{\overline{\alpha}_{t-1}} = \frac{1}{\sqrt{\alpha_t}} Y^{\mathsf{ode}}_{\overline{\alpha}_t} + \int_{\overline{\alpha}_t}^{\overline{\alpha}_{t-1}} \frac{1}{\sqrt{\gamma^3}} \underbrace{s^{\star}_{\gamma}\left(Y^{\mathsf{ode}}_{\gamma}\right)}_{\text{approx by deg-}(K-1)\text{ Lagrange polynomials}} \mathrm{d}\gamma$$

- successively, alternately refine $Y^{\mathsf{ode}}_{\gamma_{t,i}}$ and $s_{\gamma_{t,i}}\left(Y^{\mathsf{ode}}_{\gamma_{t,i}}\right)$

K score evals per iteration; $\widetilde{O}(1)$ rounds of refinements

Convergence theory for our accelerated sampler

Theorem (Li, Zhou, Wei, Chen '25)

Consider any $K = O(1)$. With perfect scores, our accelerated deterministic sampler yields $\mathsf{TV}(p_{X_1}, p_{Y_1}) \le \varepsilon$ in

$$\widetilde{O}\big(d^{1+2/K}/\varepsilon^{1/K}\big) \text{ iterations}$$

Convergence theory for our accelerated sampler

Theorem (Li, Zhou, Wei, Chen '25)

Consider any $K = O(1)$. With perfect scores, our accelerated deterministic sampler yields $\mathsf{TV}(p_{X_1}, p_{Y_1}) \le \varepsilon$ in

$$\widetilde{O}\big(d^{1+2/K}/\varepsilon^{1/K}\big) \text{ iterations}$$

- **# score function evaluations**: $\frac{d^{1+o(1)}}{\varepsilon^{1/K}}$

Convergence theory for our accelerated sampler

Theorem (Li, Zhou, Wei, Chen '25)

Consider any $K = O(1)$. With perfect scores, our accelerated deterministic sampler yields $\mathsf{TV}(p_{X_1}, p_{Y_1}) \le \varepsilon$ in

$$\widetilde{O}\big(d^{1+2/K}/\varepsilon^{1/K}\big) \text{ iterations}$$

- **# score function evaluations**: $\frac{d^{1+o(1)}}{\varepsilon^{1/K}}$
- outperforms vanilla DDIM (d/ε)
  - substantially improved ε-dependency
  - almost no loss in d-dependency;

Convergence theory for our accelerated sampler

Theorem (Li, Zhou, Wei, Chen '25)

Consider any $K = O(1)$. With perfect scores, our accelerated deterministic sampler yields $\mathsf{TV}(p_{X_1}, p_{Y_1}) \le \varepsilon$ in

$$\widetilde{O}\big(d^{1+2/K}/\varepsilon^{1/K}\big) \text{ iterations}$$

- **# score function evaluations**: $\frac{d^{1+o(1)}}{\varepsilon^{1/K}}$
- outperforms vanilla DDIM (d/ε)
    - substantially improved ε-dependency
    - almost no loss in d-dependency;
- **minimal assumptions** on data distributions
    - see also Huang et al. '24, '25 (Runge-Kutta; stronger assumptions)

Conclusion of Lecture II

- nonasymptotic convergence theory for diffusion models
- adaptation to unknown low dimensionality
- provable training-free acceleration

Conclusion of Lecture II

- nonasymptotic convergence theory for diffusion models
- adaptation to unknown low dimensionality
- provable training-free acceleration

Future directions:

- end-to-end theory that accounts for score learning $+$ sampling?
- adaptive improvement under stylized statistical models
- design of high-order stochastic samplers
- parallelization
- discrete-valued/structured problems

Papers

"A sharp convergence theory for the probability flow ODEs of diffusion models," G. Li, Y. Wei, Y. Chi, Y. Chen, arXiv:2408.02320, 2024

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

"Low-dimensional adaptation of diffusion models: convergence in total variation," J. Liang, Z. Huang, Y. Chen, arXiv:2501.12982, COLT 2025

"Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality," Z. Huang, Y. Wei, Y. Chen, arXiv:2410.18784, 2024

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

"Stochastic Runge-Kutta methods: Provable acceleration of diffusion models," Y. Wu, Y. Chen, Y. Wei, arXiv:2410.04760, 2024

"Faster diffusion models via higher-order approximation," G. Li*, Y. Zhou*, Y. Wei, Y. Chen, arXiv:2506.24042, 2025

(*=equal contributions)