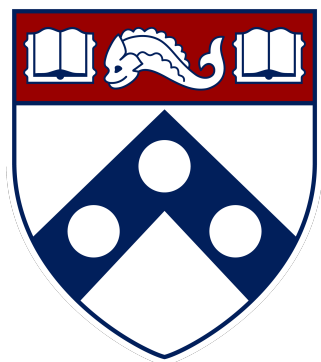
Low-dimensional adaptation of diffusion generative models



Yuxin Chen

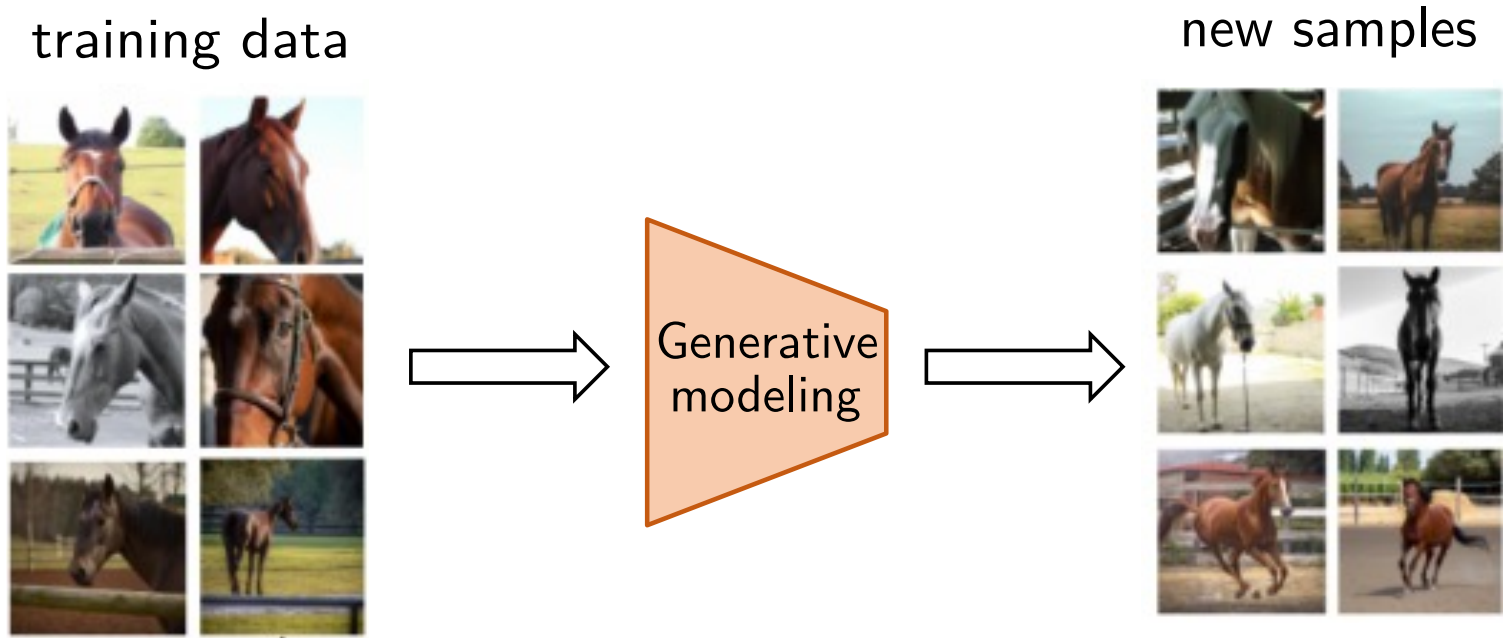
Wharton Statistics & Data Science

Generative modeling

training data

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

Generative modeling



- 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 $Y \sim p_{\text{data}}$

Inspired by nonequilibrium thermodynamics

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

Diffusion models

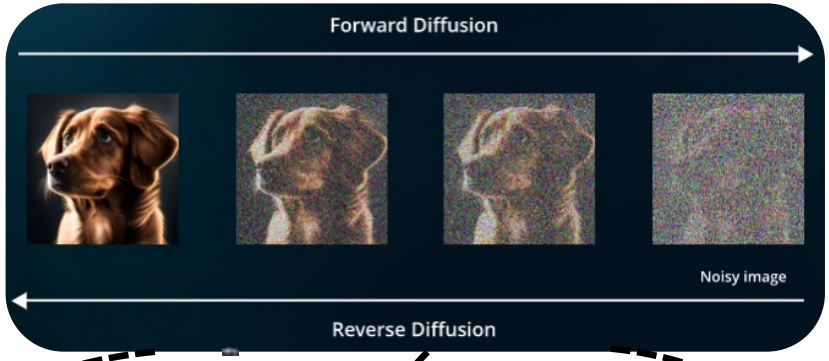
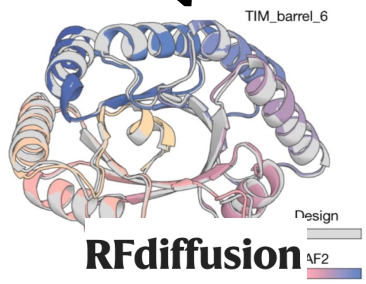
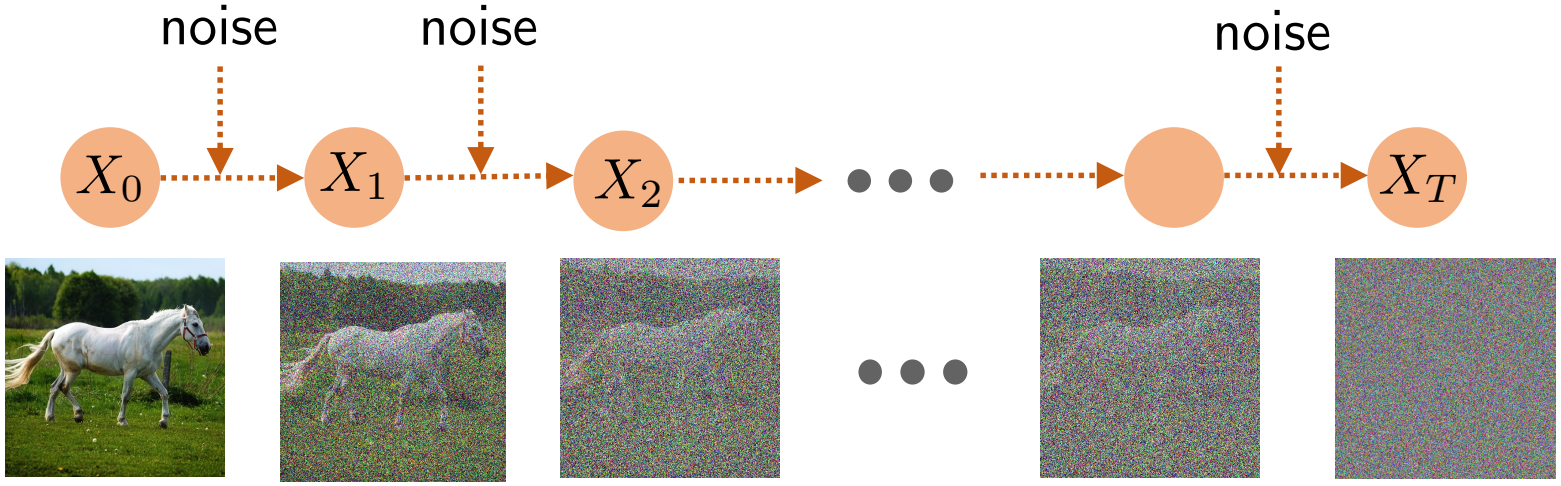


image generation

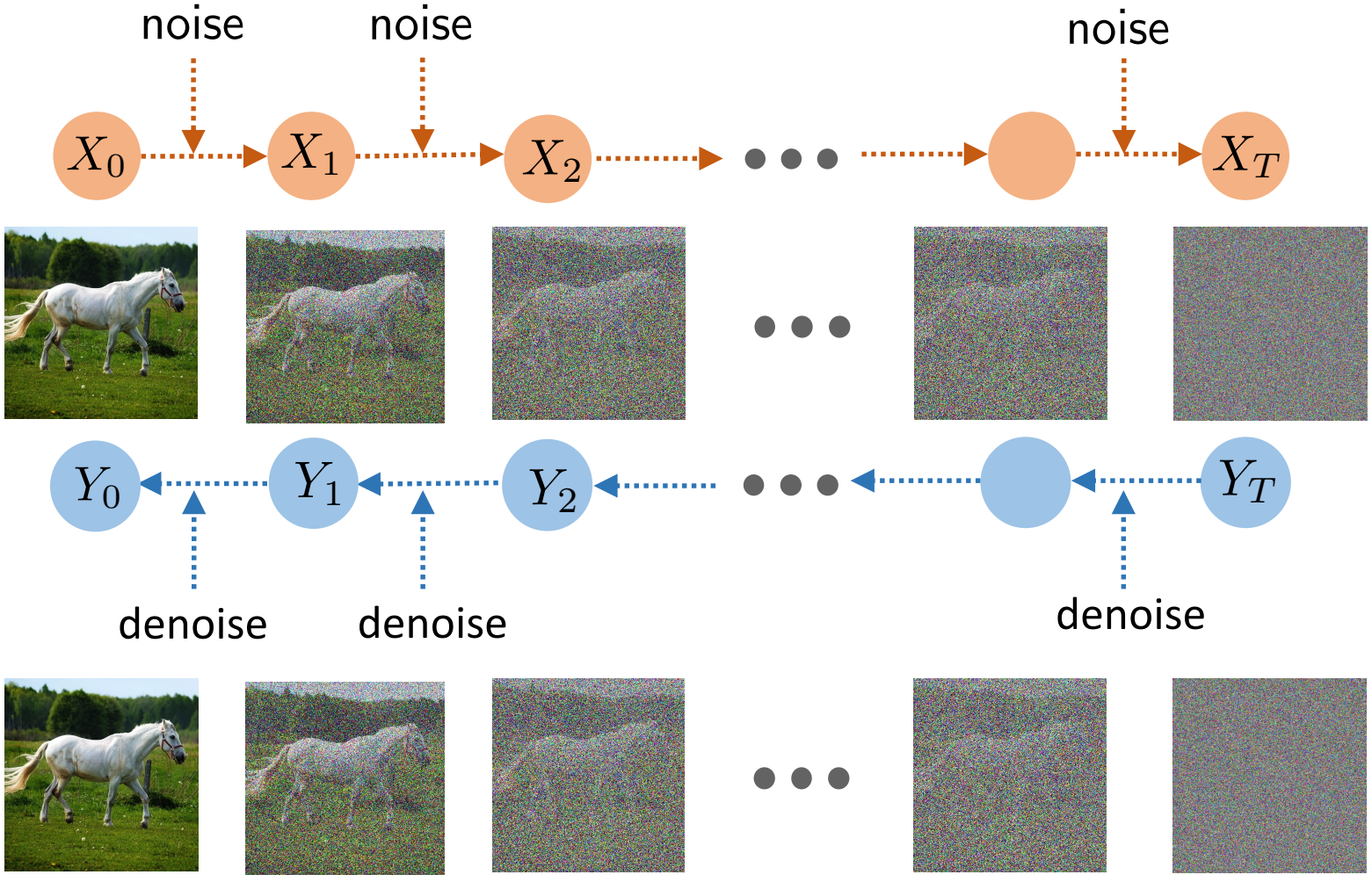
video generation



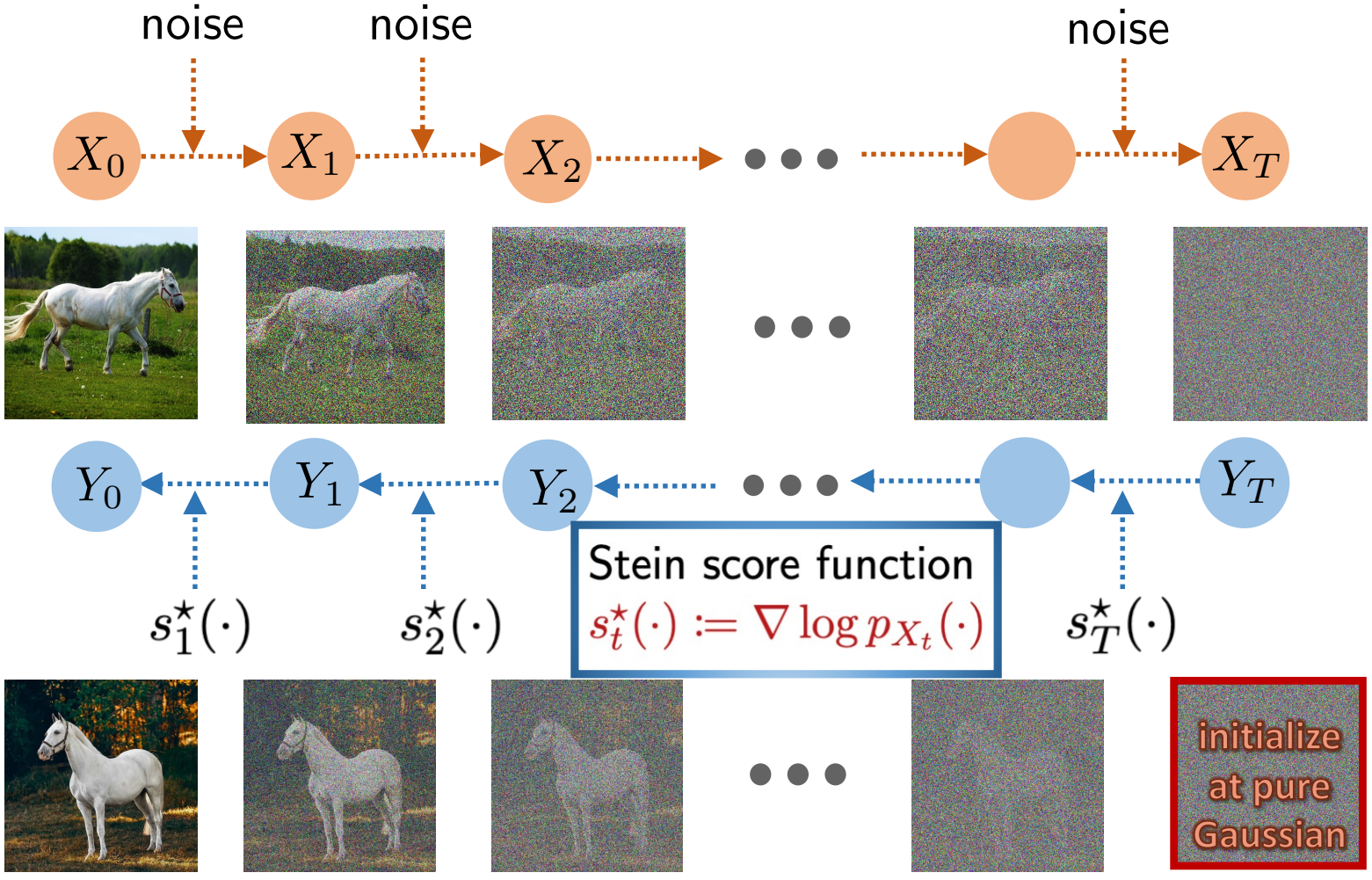
protein design



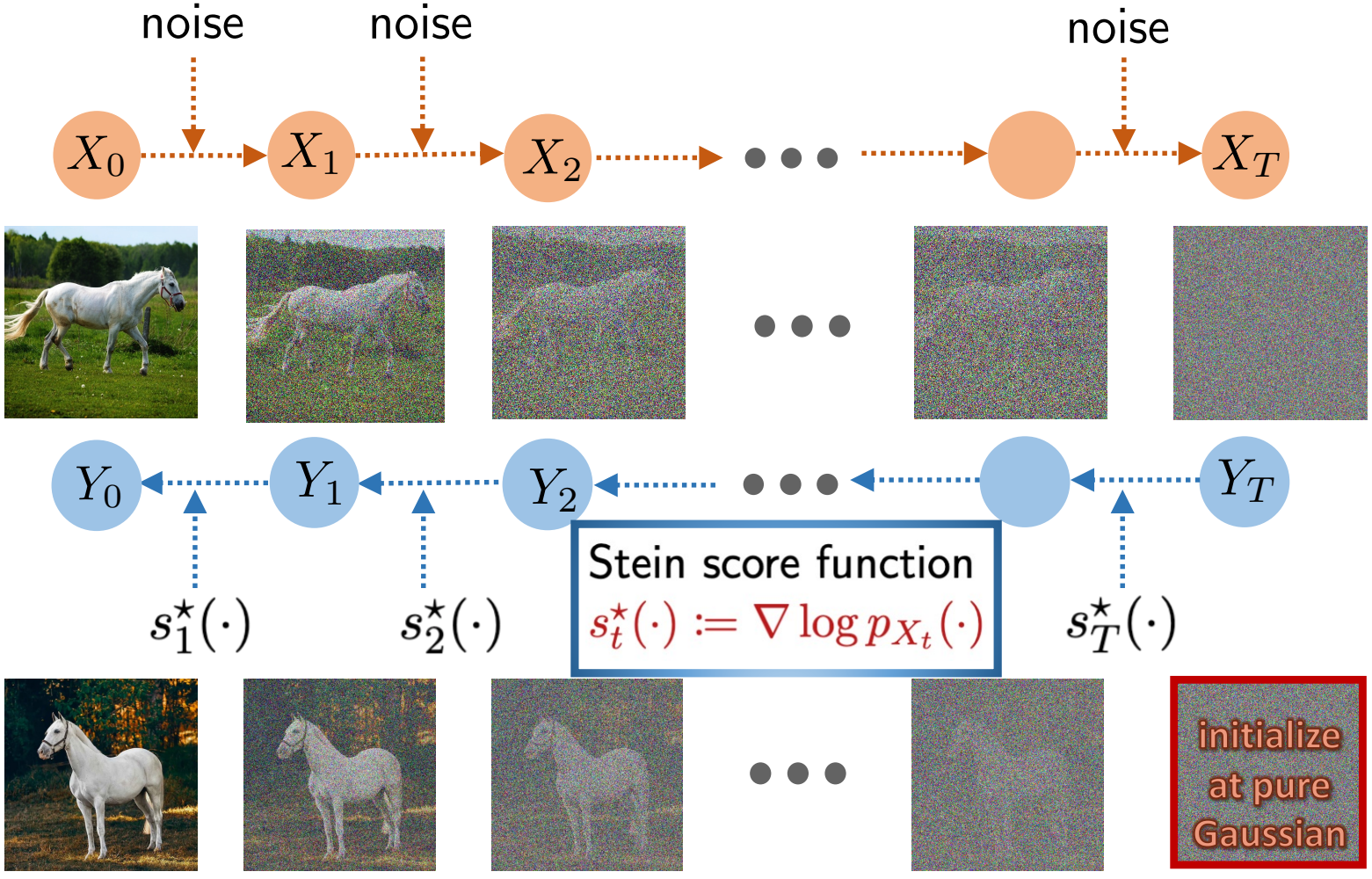
- **forward process:** diffuse data into noise



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

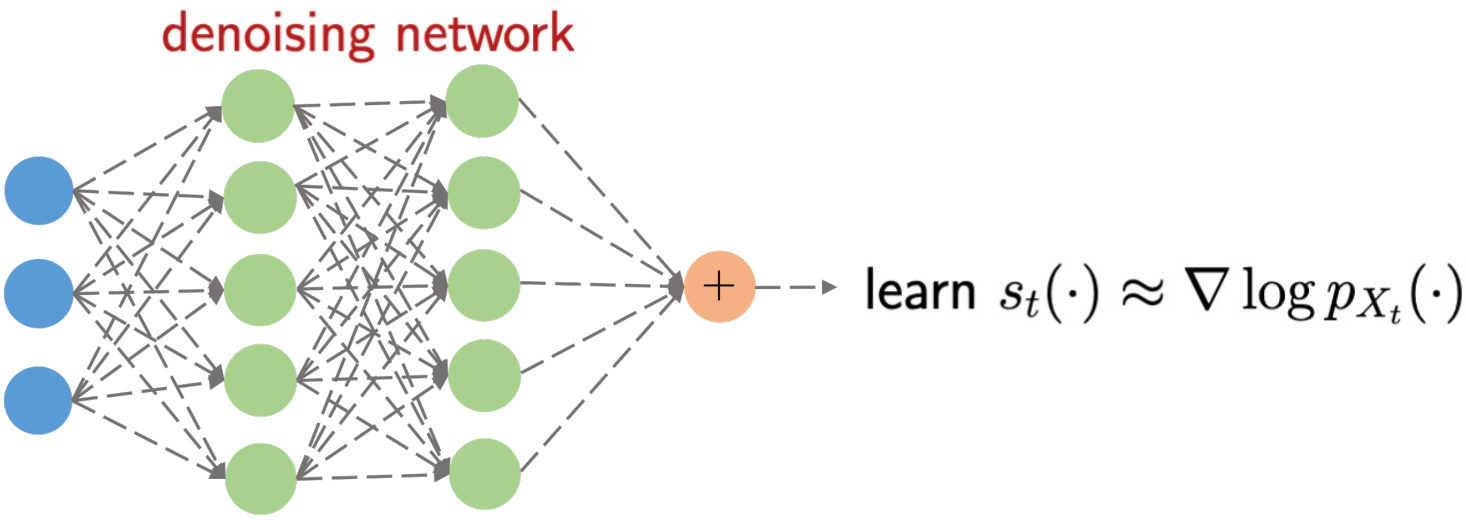


- **forward process:** diffuse data into noise
- **reverse process:** convert pure noise into data-like distributions

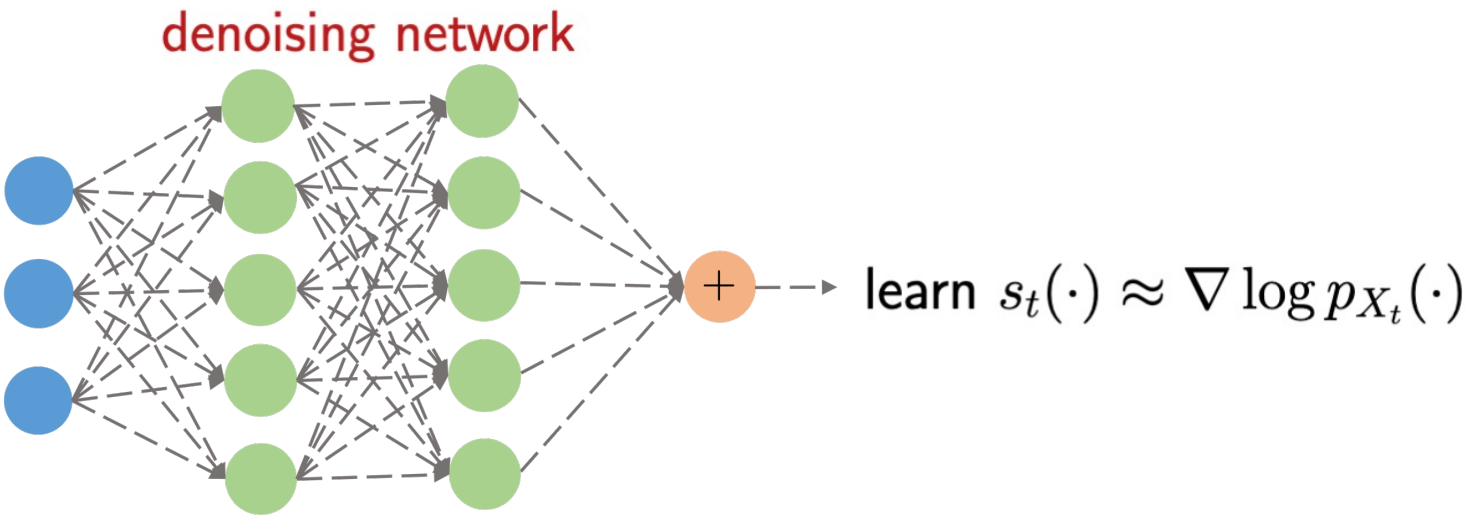


- **forward process:** diffuse data into noise
- **reverse process:** convert pure noise into data-like distributions

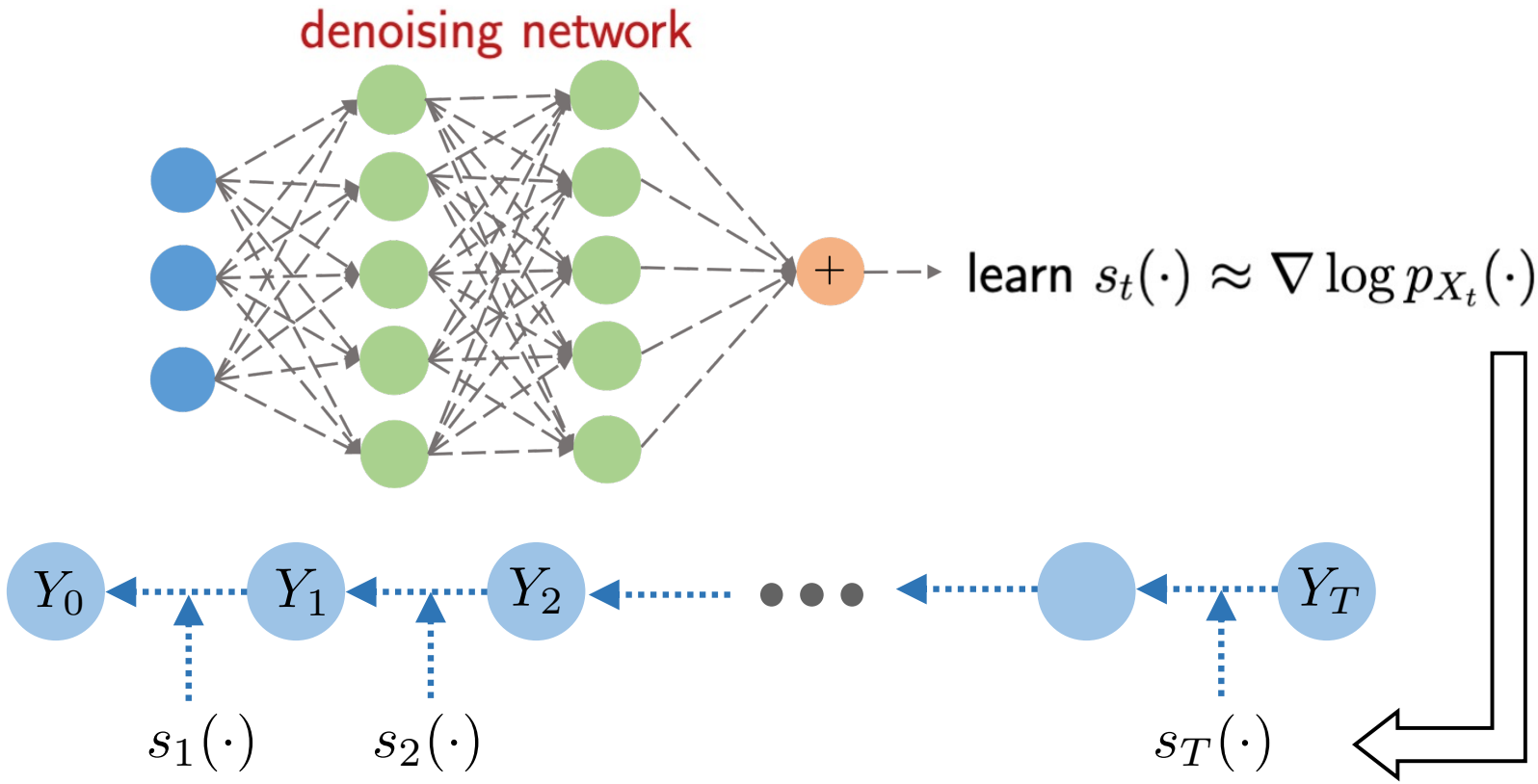
Goal: $Y_t \overset{\mathrm{d}}{\approx} X_t, \quad t = T, \cdots, 1$



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 $\{s_t(\cdot)\}$



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 $\overset{\text{✂ decouple}}{\leftarrow \not\rightarrow}$ sampling

Two mainstream approaches

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

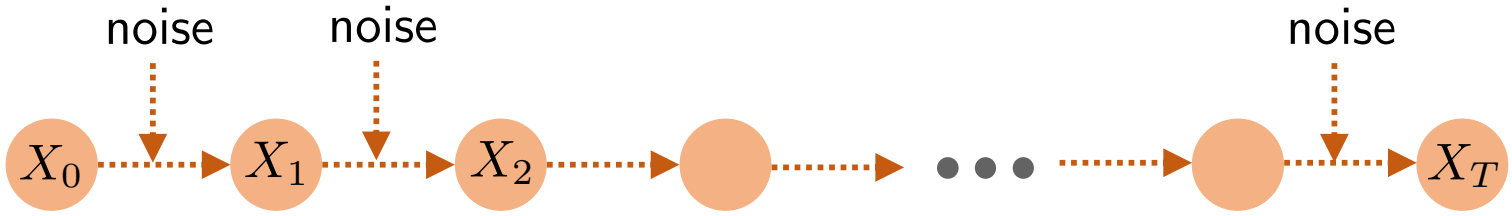
Ajay Jain
UC Berkeley
ajayj@berkeley.edu

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

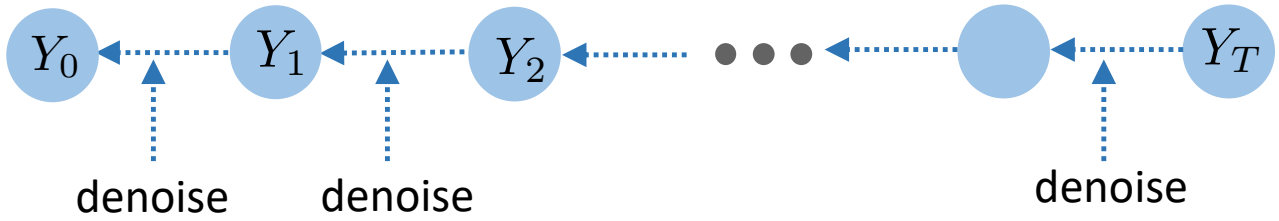
DDPM vs. DDIM



forward process: $X_0 \sim p_{\text{data}}$,

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

DDPM vs. DDIM

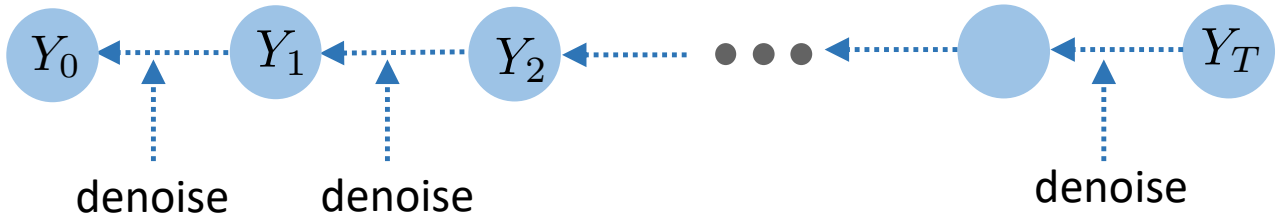


— *Ho, Jain, Abbeel '20*

1. A stochastic sampler: **denoising diffusion probabilistic models**

DDPM

DDPM vs. DDIM



— Ho, Jain, Abbeel '20

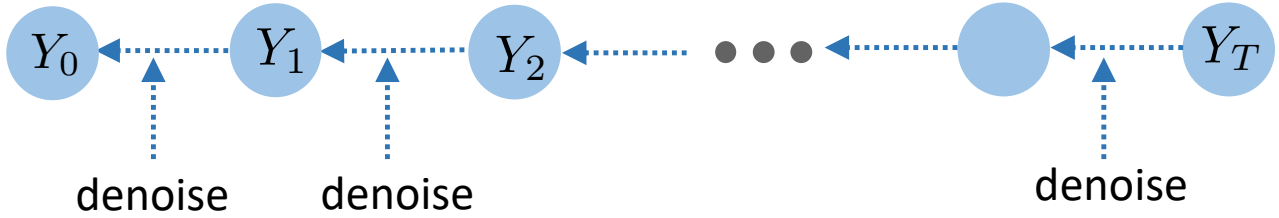
1. A stochastic sampler: **denoising diffusion probabilistic models**

DDPM

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

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

DDPM vs. DDIM

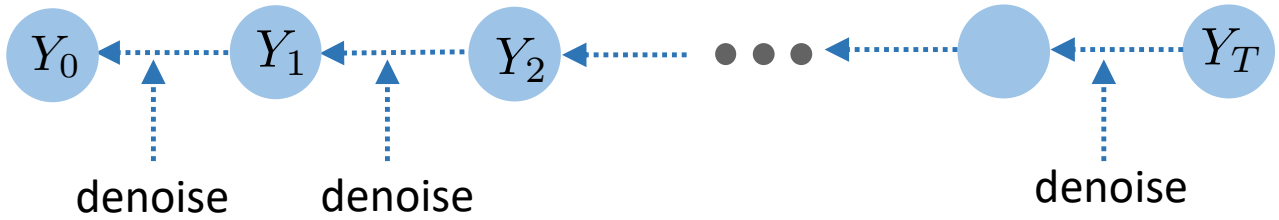


— *Song, Meng, Ermon '20*

2. A <u>deterministic</u> sampler: **denoising diffusion implicit models**

DDIM

DDPM vs. DDIM



— *Song, Meng, Ermon '20*

2. A <u>deterministic</u> sampler: **denoising diffusion implicit models**

DDIM

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

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

Interpretations from lens of SDE/ODE

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

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 as forward SDE

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

Interpretations from lens of SDE/ODE

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 as forward SDE

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

- $\exists$ reverse-time ODE w/ same *marginal* dist. as forward SDE

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

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

— *almost no restriction on target data distributions*

— *discretization schemes matter*

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

— *almost no restriction on target data distributions*

— *discretization schemes matter*

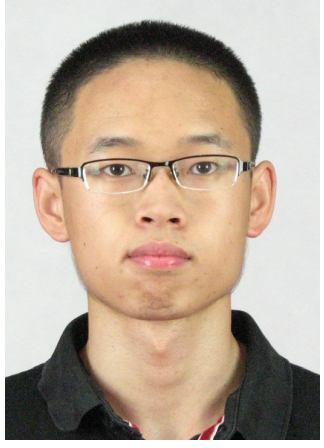
Questions:

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

This talk:

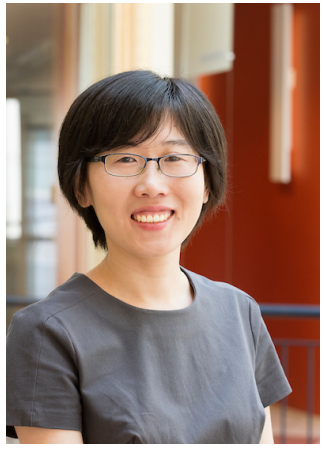
1. non-asymptotic convergence theory in discrete time
2. adaptation to (unknown) low-dimensional structure

Part 1: sharp convergence theory for DDIM



Gen Li
CUHK

Yuting Wei
UPenn



Yuejie Chi
CMU

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

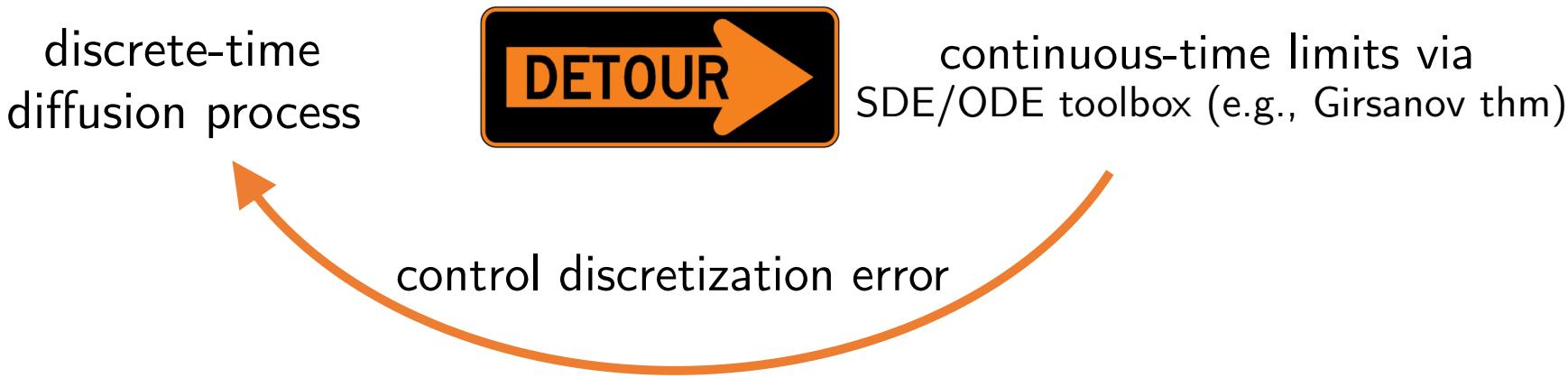
discrete-time diffusion process

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

control discretization error

Prior analyses for DDIM & DDPM

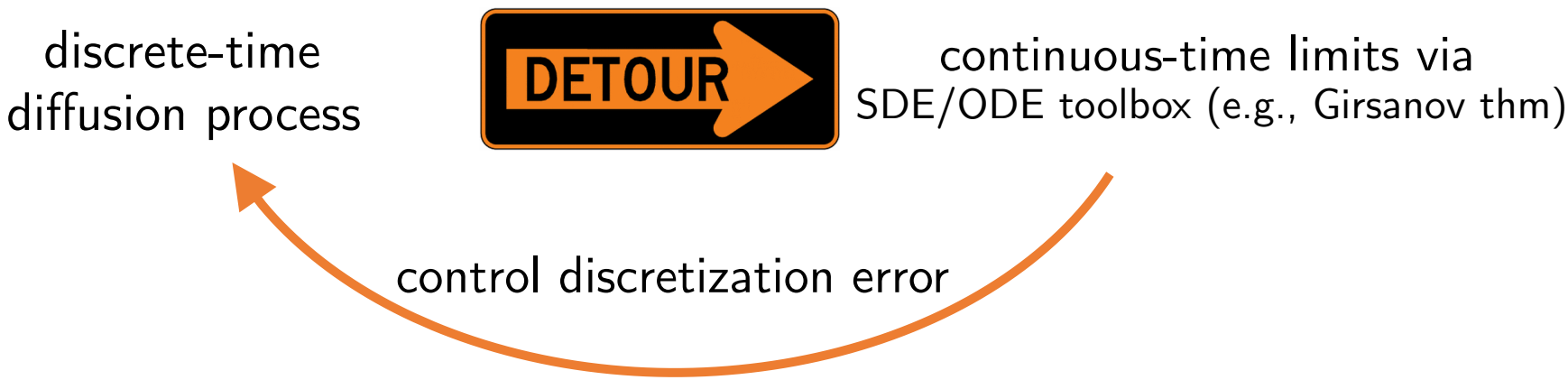
— Li, Lu, Tan '22
— Chen, Lee, Lu '22
— Chen, Chewi, Li, Li, Salim, Zhang '22
— Chen, Daras, Dimakis '23
— Chen, Chewi, Lee, Li, Lu, Salim '23
— Benton, De Bortoli, Doucet, Deligiannidis '23



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*



- Built upon toolboxes from SDE/ODE
- Prior analyses **highly inadequate** for deterministic samplers
  e.g., DDIM

Can we develop a versatile non-asymptotic framework that
analyzes diffusion models in discrete time directly?

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

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

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

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

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:* $s_t^\star(X) := \nabla \log p_{X_t}(X)$,

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

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

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

  - 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:*

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

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 $c_0, c_1 > 0$,

$$1 - \alpha_1 = \frac{1}{T^{c_0}}$$

$$1 - \alpha_t = \underbrace{\frac{c_1 \log T}{T} \min\left\{(1-\alpha_1)\left(1 + \frac{c_1 \log T}{T}\right)^t, 1\right\}}_{\text{2 phases: exp growth} \to \text{flat}}$$

Main result: DDIM / probability flow ODE

Theorem 1 (Li, Wei, Chi, Chen '24)

The DDIM sampler can achieve (up to log factor)

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

Main result: DDIM / probability flow ODE

Theorem 1 (Li, Wei, Chi, Chen '24)

The DDIM sampler can achieve (up to log factor)

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

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

Main result: DDIM / probability flow ODE

Theorem 1 (Li, Wei, Chi, Chen '24)

The DDIM sampler can achieve (up to log factor)

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

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

Main result: DDIM / probability flow ODE

Theorem 1 (Li, Wei, Chi, Chen '24)

The DDIM sampler can achieve (up to log factor)

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

- $\underbrace{\textbf{iteration complexity:}}_{\text{to yield TV dist} \le \varepsilon}$ d/ε for small enough ε
- **stability:** $\mathsf{TV}(p_{X_1}, p_{Y_1}) \propto$ error measures $\varepsilon_{\mathsf{score}}$ and $\varepsilon_{\mathsf{Jacobi}}$
- **general data distribution:** no need of smoothness, log-concavity

Comparison w/ prior DDIM theory

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

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

Comparison w/ prior DDIM theory

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

- *Chen, Daras, Dimakis '23:* no concrete poly dependency
  (ours: d/ε)
  exponential in smoothness parameter
  (ours: independent of smoothness pars)
  needs exact score functions
  (ours: allow score errors)
- *Chen, Chewi, Lee, Li, Lu, Salim '23:* requires additional stochastic correction steps & smoothness
  (different from DDIM)

Comparison w/ prior DDIM theory

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

- *Chen, Daras, Dimakis '23:* no concrete poly dependency (ours: d/ε), exponential in smoothness parameter (ours: independent of smoothness pars), needs exact score functions (ours: allow score errors)
- *Chen, Chewi, Lee, Li, Lu, Salim '23:* requires additional stochastic correction steps (different from DDIM) & smoothness
- *Huang, Huang, Lin '24:* suboptimal d-dependency (i.e., d^2/ε) (ours: d/ε)

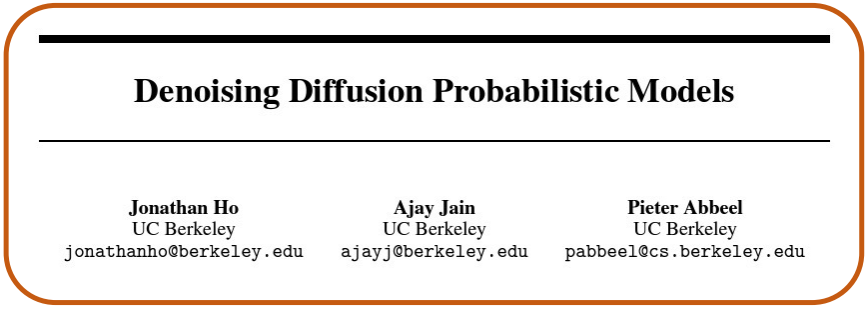
Part 2: adaptation to (unknown) low dimensionality

Jiadong Liang
UPenn

Zhihan Huang
UPenn

Yuting Wei
UPenn

Theory for mainstream diffusion models



Theorem 2 (Li, Wei, Chi, Chen '24, Li, Yan '24)

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

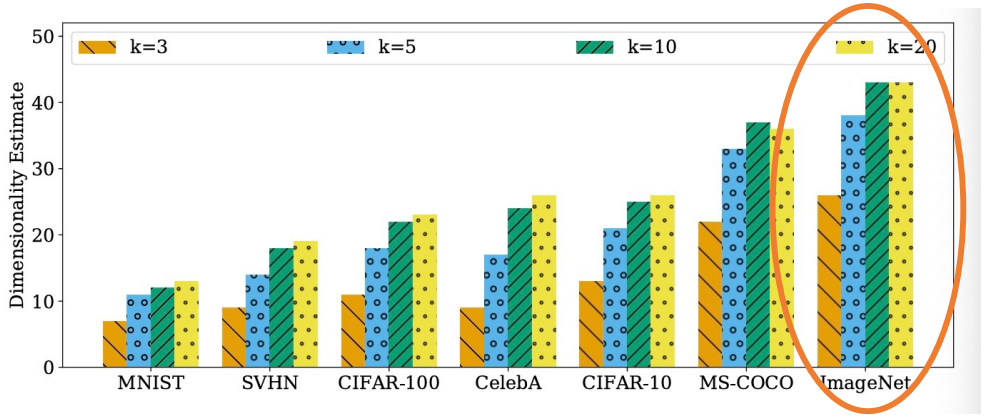
$$\widetilde{O}(d/\varepsilon) \text{ iterations}$$

- d: *ambient dimension*

d/ε iterations are too slow ...

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

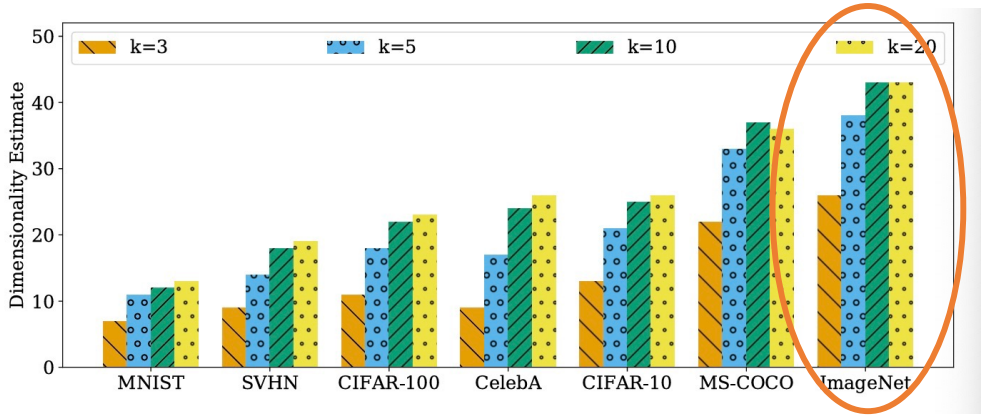
d/ε iterations are too slow ...



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

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

d/ε iterations are too slow ...



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

$k = 43$ 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

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

Intrinsic dimension

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

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

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

Main result: convergence in total variation

Theorem 3 (Liang, Huang, Chen '24)

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

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

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

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

Main result: convergence in total variation

Theorem 3 (Liang, Huang, Chen '24)

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

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

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

Main result: convergence in total variation

Theorem 3 (Liang, Huang, Chen '24)

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

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

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!*

Main result: convergence in KL divergence

Theorem 4 (Huang, Wei, Chen '24)

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

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

— *concurrent work Potaptchik et al.'24*

Main result: convergence in KL divergence

Theorem 4 (Huang, Wei, Chen '24)

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

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

— *concurrent work Potaptchik et al.'24*

- optimal scaling in k

Main result: convergence in KL divergence

Theorem 4 (Huang, Wei, Chen '24)

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

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

— *concurrent work Potaptchik et al.'24*

- optimal scaling in k
- Pinsker inequality $(\mathsf{TV} \lesssim \sqrt{\mathsf{KL}})$ is loose when deriving TV bound

$$\underbrace{k/\varepsilon^2}_{\text{Pinsker} + \text{Thm 4}} \quad \text{iterations} \quad \text{vs.} \quad \underbrace{k/\varepsilon}_{\text{Thm 3}} \quad \text{iterations}$$

Sampler	Smoothness	Convergence rate (in total variation)	Adaptation to low dimension
Chen et al.'22	L-Lipschitz	$L\sqrt{d/T}$	✗
Chen et al.'23	no requirement	$\sqrt{d^2/T}$	✗
Benton et al.'23	no requirement	$\sqrt{d/T}$	✗
Li & Yan'24a	no requirement	$k^2/\sqrt{T}$	✓
Li & Yan'24b	no requirement	d/T	✗
Our work	no requirement	k/T	✓

Table 1: comparison with prior DDPM theory

Sampler	Smoothness	Convergence rate (in total variation)	Adaptation to low dimension
Chen et al.'23	L-Smooth	$\mathrm{poly}(Ld)/\sqrt{T}$	✗
Li et al.'23	no requirement	$d^2/T + d^6/T^2$	✗
Huang et al.'24	L-smooth	$(Ld)^2/T$	✗
Li et al.'24	no requirement	$d/T + (d^2/T)^{\log T}$	✗
Our work	no requirement	k/T	✓

Table 2: comparison with prior DDIM theory

Crucial choices of coefficients

Both DDPM and DDIM update rules take the form

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

Crucial choices of coefficients

Both DDPM and DDIM update rules take the form

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

Theorem 5 (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

Both DDPM and DDIM update rules take the form

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

Theorem 5 (Liang, Huang, Chen '24)

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

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

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

- both DDIM and DDPM satisfy this!

Connection to SDE/ODE

Reverse-time SDE w/ the same marginal distribution:

$$\mathrm{d}Y_t = \big(Y_t + 2s_{T-t}(Y_t)\big)\beta(T-t)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

Connection to SDE/ODE

Reverse-time SDE w/ the same marginal distribution:

$$\mathrm{d}Y_t = \big(Y_t + 2s_{T-t}(Y_t)\big)\beta(T-t)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

(Reparametrization) $\Downarrow$ def. $\mu_t(x) := \mathbb{E}[X_0 | X_t = x]$

$$\mathrm{d}Y_t = \big(A_1(t)Y_t + A_2(t)\mu_{T-t}(Y_t)\big)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

Connection to SDE/ODE

Reverse-time SDE w/ the same marginal distribution:

$$\mathrm{d}Y_t = \big(Y_t + 2s_{T-t}(Y_t)\big)\beta(T-t)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

(Reparametrization) $\Downarrow$ def. $\mu_t(x) := \mathbb{E}[X_0 | X_t = x]$

$$\mathrm{d}Y_t = \big(A_1(t)Y_t + A_2(t)\mu_{T-t}(Y_t)\big)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

(Discretization) $\Downarrow$ exponential integrator scheme

$$\mathrm{d}Y_t = \big(A_1(t)Y_t + A_2(t)\mu_{T-t_n}(Y_{t_n})\big)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

Connection to SDE/ODE

Reverse-time SDE w/ the same marginal distribution:

$$\mathrm{d}Y_t = \big(Y_t + 2s_{T-t}(Y_t)\big)\beta(T-t)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

(Reparametrization) $\Downarrow$ def. $\mu_t(x) := \mathbb{E}[X_0 | X_t = x]$

$$\mathrm{d}Y_t = \big(A_1(t)Y_t + A_2(t)\mu_{T-t}(Y_t)\big)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

(Discretization) $\Downarrow$ exponential integrator scheme

$$\mathrm{d}Y_t = \big(A_1(t)Y_t + A_2(t)\mu_{T-t_n}(Y_{t_n})\big)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

(Discretization points) $\Updownarrow$ Choose $\{t_n\} = 0, \cdots, T-1$

Solve analytically? DDPM sampler!

Connection to SDE/ODE

Reverse-time SDE w/ the same marginal distribution:

$$\mathrm{d}Y_t = \big(Y_t + 2s_{T-t}(Y_t)\big)\beta(T-t)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

(Reparametrization) $\Downarrow$ def. $\mu_t(x) := \mathbb{E}[X_0 | X_t = x]$

$$\mathrm{d}Y_t = \big(A_1(t)Y_t + A_2(t)\mu_{T-t}(Y_t)\big)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

(Discretization) $\Downarrow$ exponential integrator scheme

$$\mathrm{d}Y_t = \big(A_1(t)Y_t + A_2(t)\mu_{T-t_n}(Y_{t_n})\big)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}B_t$$

(Discretization points) $\Updownarrow$ Choose $\{t_n\} = 0, \cdots, T-1$

Solve analytically? DDPM sampler!

Why DDPM works?: $\mu_t(x)$ is a "projection" onto low-dimensional structure

$\Rightarrow$ enhances smoothness & adapts to low dimensionality!

— *same trick works for DDIM*

Concluding remarks

- Sharp convergence theory for DDIM (or probability flow ODE)
- Diffusion models are adaptive to unknown low dimensionality!
  - choices of coefficients matter!

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

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

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