SCAN ME

How well does diffusion model generate?
- a training and sampling combined quantification

arXiv:2406.12839

Yuqing Wang [1,2]
Ye He [1]
Molei Tao [1]

1 Georgia Tech, USA
2 Simons Institute, UC Berkeley

Sep 10, 2024 Emerging Generalization Settings Workshop @ Simons

90% quantification of diffusion model **generation**

90%		quantification of diffusion model **generation**

≤10%		generalization

why am I here?

90%   quantification of diffusion model **generation**

≤10%   generalization

why am I here?

- one interesting generalization setting: understood

90% quantification of diffusion model **generation**

≤10% generalization

why am I here?

- one interesting generalization setting: understood
- another: not

90% quantification of diffusion model **generation**

≤10% generalization

why am I here?

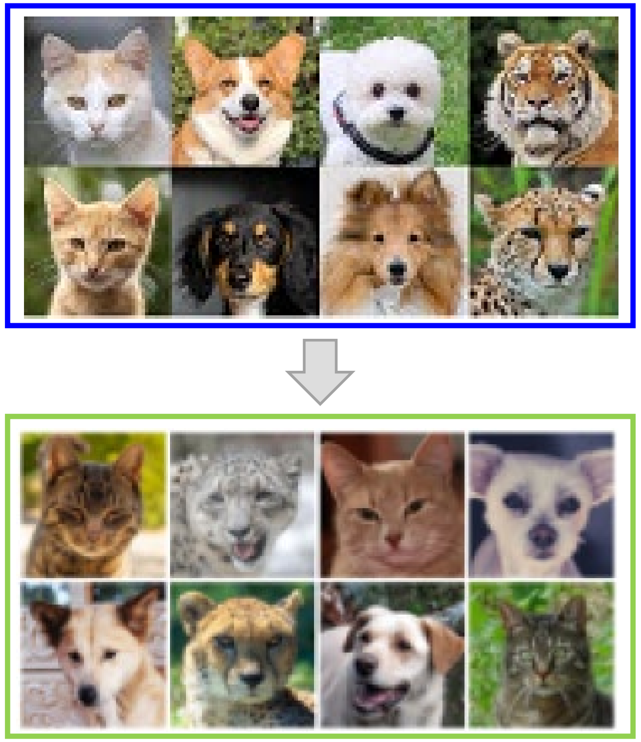
- one interesting generalization setting: understood
- another: not
- the quantifications are interesting too (hopefully)

Generative Modeling

Given samples of an unknown probability distribution (possibly in very high dim.), generate **more samples** of the same distribution.

Generative Modeling

Given samples of an unknown probability distribution (possibly in very high dim.), generate **more samples** of the same distribution.



denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

denoising diffusion model

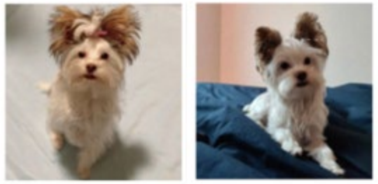
(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

Stable Diffusion, DALL·E, Midjourney; Sora;

(Chat)GPT, Gemini, Llama, Claude, …

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

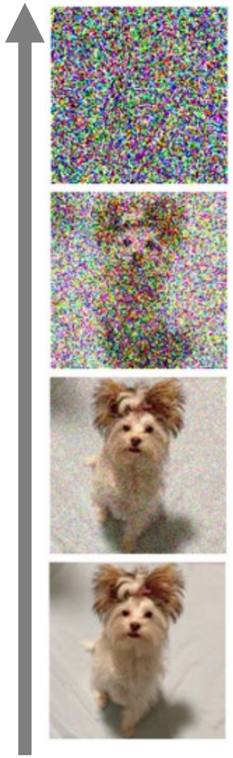


Yang+ 22

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

forward noising process: learn "score" (~ evolution of data density)

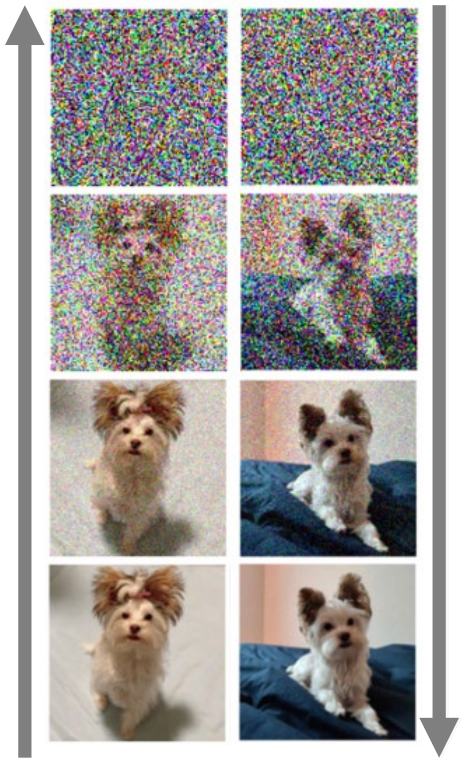


Yang+ 22

denoising diffusion model

(*Sohl-Dickstein*+ 15, *Ho*+ 20, *Song*+ 21 ...)

forward noising process: learn “score” (~ evolution of data density)



backward denoising process: use “score” to generate data from noise

Yang+ 22

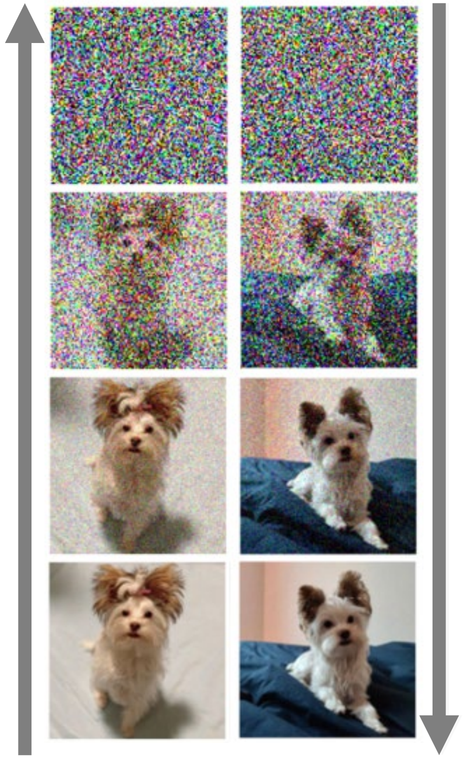
denoising diffusion model

(*Sohl-Dickstein*+ 15, *Ho*+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

forward noising process: learn "score" (~ evolution of data density)



backward denoising process: use "score" to generate data from noise

Yang+ 22

denoising diffusion model

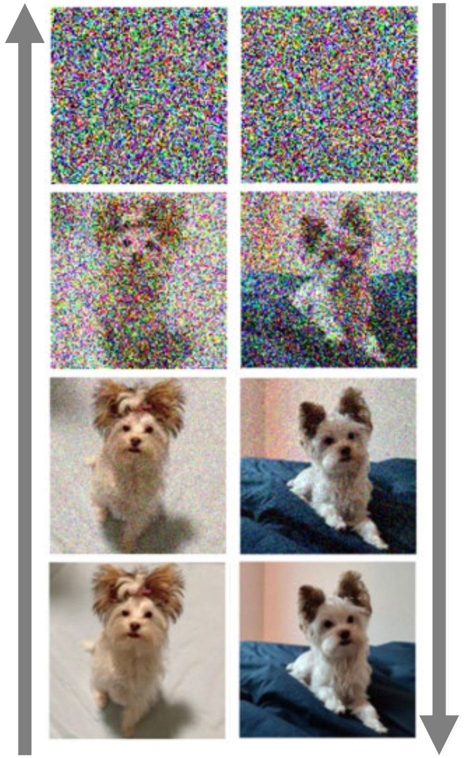
(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X dt + \sqrt{2} dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y dt + 2s(Y, T-t) dt + \sqrt{2} dB_t}$$

forward noising process: learn "score" (~ evolution of data density)



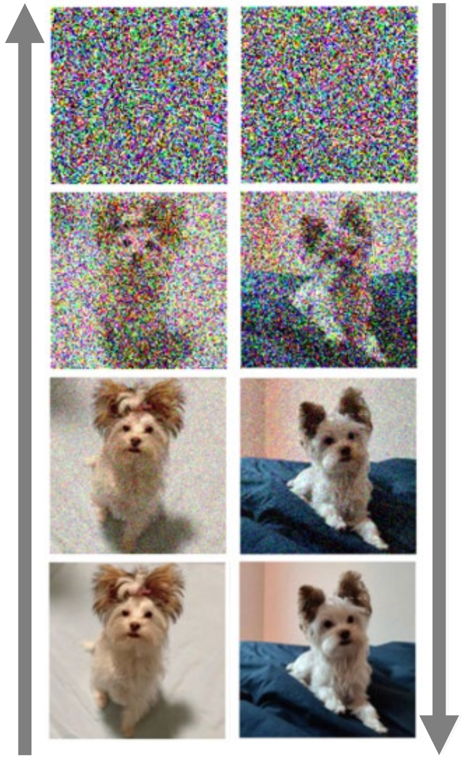
backward denoising process: use "score" to generate data from noise

Yang+ 22

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

forward noising process: learn "score" (~ evolution of data density)



backward denoising process: use "score" to generate data from noise

Yang+ 22

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

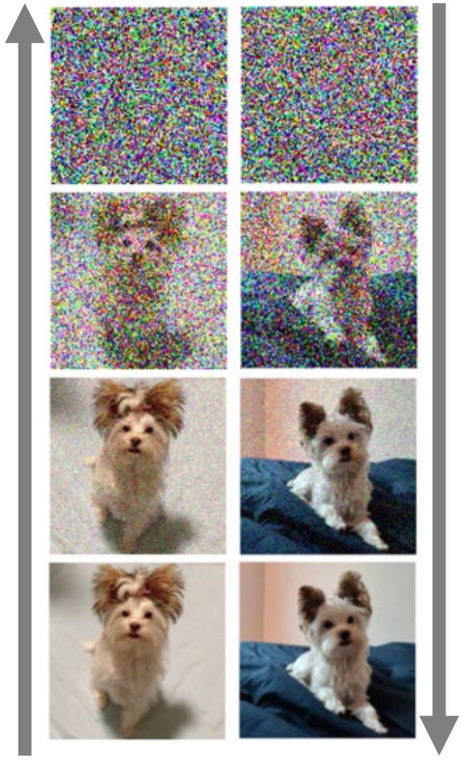
$$X(t) \sim p(x,t)$$

0.2 Introduction: Denoising Diffusion Generative Model

denoising diffusion model

(*Sohl-Dickstein*+ 15, *Ho*+ 20, *Song*+ 21 ...)

forward noising process: learn “score” (~ evolution of data density)



backward denoising process: use “score” to generate data from noise

Yang+ 22

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \overset{d}{=} Y(T-t), \forall t \qquad X(t) \sim p(x,t)$$

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t}$$

score ??

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

denoising diffusion model

(*Sohl-Dickstein*+ 15, *Ho*+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t}$$

score ??

learned during forward process

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score ??

learned during forward process

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

score matching

$$\min_\theta \int_0^T w(t)\mathbb{E}_{X_t}\|\nabla_x \log p(X_t,t) - s_\theta(X_t,t)\|^2 dt$$

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score ??

learned during forward process

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

score matching

$$\min_\theta \int_0^T w(t) \underbrace{\mathbb{E}_{X_t} \|\nabla_x \log p(X_t, t) - s_\theta(X_t, t)\|^2}_{\approx \frac{1}{n}\sum_{i=1}^n \|\nabla_x \log p(X_t^i, t) - s_\theta(X_t^i, t)\|^2} dt$$

$$X_0^i \sim \mathcal{D}$$

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score ??

learned during forward process

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

score matching unimplementable

$$\min_\theta \int_0^T w(t) \underbrace{\mathbb{E}_{X_t} \|\boxed{\nabla_x \log p}(X_t, t) - s_\theta(X_t, t)\|^2}_{\approx \frac{1}{n}\sum_{i=1}^n \|\boxed{\nabla_x \log p}(X_t^i, t) - s_\theta(X_t^i, t)\|^2} dt$$

$$X_0^i \sim \mathcal{D}$$

denoising diffusion model

(*Sohl-Dickstein*+ 15, *Ho*+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score ??

learned during forward process

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

score matching unimplementable

$$\min_\theta \int_0^T w(t)\mathbb{E}_{X_t}\|\boxed{\nabla_x \log p}(X_t, t) - s_\theta(X_t, t)\|^2 dt$$

denoising score matching

$$\min_\theta \int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0, t) - s_\theta(X_t, t)\|^2 dt$$

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score ??

learned during forward process

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

score matching unimplementable

$$\min_\theta \int_0^T w(t)\mathbb{E}_{X_t}\|\boxed{\nabla_x \log p}(X_t,t) - s_\theta(X_t,t)\|^2 dt$$

denoising score matching

$$\min_\theta \int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\boxed{\nabla_x \log p_{t|0}}(X_t|X_0,t) - s_\theta(X_t,t)\|^2 dt$$

analytically available

denoising diffusion model

(*Sohl-Dickstein*+ 15, *Ho*+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score ??

learned during forward process

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

score matching unimplementable

$$\min_\theta \int_0^T w(t)\mathbb{E}_{X_t}\|\boxed{\nabla_x \log p}(X_t,t) - s_\theta(X_t,t)\|^2 dt$$

denoising score matching

$$\min_\theta \int_0^T w(t)\boxed{\mathbb{E}_{X_0}}\mathbb{E}_{X_t|X_0}\|\boxed{\nabla_x \log p_{t|0}}(X_t|X_0,t) - s_\theta(X_t,t)\|^2 dt$$

analytically available

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score ??

learned during forward process

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

score matching unimplementable

$$\min_\theta \int_0^T w(t)\mathbb{E}_{X_t}\|\boxed{\nabla_x \log p}(X_t, t) - s_\theta(X_t, t)\|^2 dt$$

denoising score matching less trivial than appeared, but provable

$$\min_\theta \int_0^T w(t)\boxed{\mathbb{E}_{X_0}}\mathbb{E}_{X_t|X_0}\|\boxed{\nabla_x \log p_{t|0}}(X_t|X_0, t) - s_\theta(X_t, t)\|^2 dt$$

analytically available

1.1 Quantification of Diffusion Model's Generation Quality: Overview

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

1. forward training/learning process:
   optimization → score s

1.1 Quantification of Diffusion Model's Generation Quality: Overview

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

1. forward training/learning process:
   optimization → score s

2. backward sampling/inference process:
   numerical simulation → sample Y

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

1.1 Quantification of Diffusion Model's Generation Quality: Overview

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

1. forward training/learning process:
   optimization → score s

2. backward sampling/inference process:
   numerical simulation → sample Y

? quality of generated samples

$$\mathrm{KL}\big(\mathrm{Law}(X_0)|\mathrm{Law}(Y_T)\big) \leq \cdots$$

denoising diffusion model

(*Sohl-Dickstein*+ 15,
Ho+ 20, *Song*+ 21 ...)

1. forward training/learning process:
   optimization → score s

2. backward sampling/inference process:
   numerical simulation → sample Y

? quality of generated samples

$$\mathrm{KL}\big(\mathrm{Law}(X_0)|\mathrm{Law}(Y_T)\big) \leq \cdots$$

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

"main stream"

if score is approximated with error ≤ ε in the sense of ______, then generated and training samples have statistical distance/divergence ≤ ___, under assumptions ___________.

Lee+ 22, *de Bortoli* 22, *Yang* & *Wibisono* 22, *S Chen*+ 23,
H Chen+ 23, *Benton*+ 23, *Conforti*+ 23, …

denoising diffusion model

(*Sohl-Dickstein*+ 15, *Ho*+ 20, *Song*+ 21 ...)

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

1. forward training/learning process: optimization → score s

only

2. backward sampling/inference process: numerical simulation → sample Y

? quality of generated samples

$$\mathrm{KL}\big(\mathrm{Law}(X_0)|\mathrm{Law}(Y_T)\big) \leq \cdots$$

"main stream"

if score is approximated with error ≤ ε in the sense of ______, then generated and training samples have statistical distance/divergence ≤ ___, under assumptions ___________.

Lee+ 22, *de Bortoli* 22, *Yang* & *Wibisono* 22, *S Chen*+ 23, *H Chen*+ 23, *Benton*+ 23, *Conforti*+ 23, …

2. backward sampling/inference process:
   numerical simulation → sample Y

only

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

only

2. backward sampling/inference process:
   numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

- score error

only

2. backward sampling/inference process:
numerical simulation → sample Y

already highly nontrivial

sources of error

- score error

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2\cancel{s}(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

s_θ score

$$s(x,t) := \nabla_x \log p(x,t)$$

2. backward sampling/inference process:
   numerical simulation → sample Y

only

already highly nontrivial

sources of error

- score error
- integration error

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

only

2. backward sampling/inference process:
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

- score error
- integration error

common: exponential integrator

only

2. backward sampling/inference process:
   numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

- score error
- integration error

common: exponential integrator

$$dY_t = Y_t dt + 2s(Y_t, T-t)dt + \sqrt{2}dB_t$$

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

- score error
- integration error

common: exponential integrator

Y_t: exact solution, N/A due to nonlinearity

$$dY_t = Y_t dt + 2s(Y_t, T-t)dt + \sqrt{2}dB_t$$

2. **only** backward sampling/inference process: numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

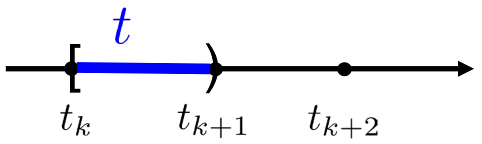
- score error
- integration error

common: exponential integrator

Y_t: exact solution, N/A due to nonlinearity

$$dY_t = Y_t dt + 2s(Y_t, T-t)dt + \sqrt{2}dB_t$$

$$d\hat{Y}_t = \hat{Y}_t dt + 2s(\hat{Y}_{\boxed{t_k}}, T-t_k)dt + \sqrt{2}dB_t$$



only

2. backward sampling/inference process:
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

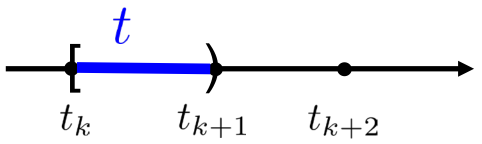
- score error
- integration error

common: exponential integrator

Y_t: exact solution, N/A due to nonlinearity

$$dY_t = Y_t dt + 2s(Y_t, T-t)dt + \sqrt{2}dB_t$$

$$d\hat{Y}_t = \hat{Y}_t dt + 2s(\hat{Y}_{\boxed{t_k}}, T-t_k)dt + \sqrt{2}dB_t$$



$\hat{Y}_t$: exactly solvable, a numerical approximation of Y_t

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

- score error
- integration error

common: exponential integrator

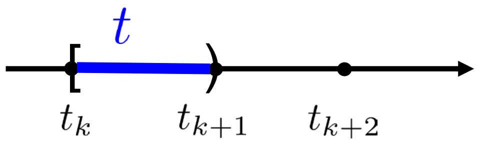
Y_t: exact solution, N/A due to nonlinearity

efficient integration:

lots of research

$$dY_t = Y_t\,dt + 2s(Y_t, T-t)\,dt + \sqrt{2}\,dB_t$$

$$d\hat{Y}_t = \hat{Y}_t\,dt + 2s(\hat{Y}_{\boxed{t_k}}, T-t_k)\,dt + \sqrt{2}\,dB_t$$



$\hat{Y}_t$: exactly solvable, a numerical approximation of Y_t

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

- score error
- integration error

common: exponential integrator

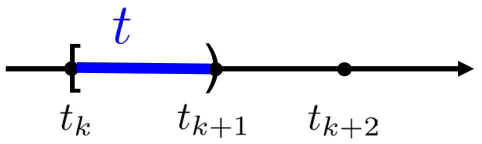
Y_t: exact solution, N/A due to nonlinearity

efficient integration:

lots of research

$$dY_t = Y_t dt + 2s(Y_t, T-t)dt + \sqrt{2}dB_t$$

$$d\hat{Y}_t = \hat{Y}_t dt + 2s(\hat{Y}_{\boxed{t_k}}, T-t_k)dt + \sqrt{2}dB_t$$



$\hat{Y}_t$: exactly solvable, a numerical approximation of Y_t

$\hat{Y}_{t_k} \mapsto \hat{Y}_{t_{k+1}}$: one function evaluation (of s)

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

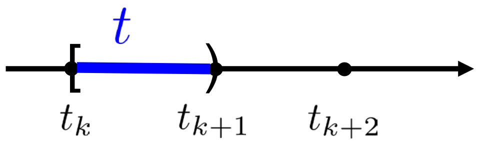
- score error
- integration error

common: exponential integrator

Y_t: exact solution, N/A due to nonlinearity

$$dY_t = Y_t\,dt + 2s(Y_t, T-t)\,dt + \sqrt{2}\,dB_t$$

$$d\hat{Y}_t = \hat{Y}_t\,dt + 2s(\hat{Y}_{\boxed{t_k}}, T-t_k)\,dt + \sqrt{2}\,dB_t$$



$\hat{Y}_t$: exactly solvable, a numerical approximation of Y_t

$\hat{Y}_{t_k} \mapsto \hat{Y}_{t_{k+1}}$: one function evaluation (of s)

efficient integration:

lots of research

e.g., DEIS, gDDIM
(*Zhang* & *Chen* 2022)
(*Zhang*, ***Tao*** & *Chen* 2022)

1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

only

2. backward sampling/inference process:
numerical simulation → sample Y

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

already highly nontrivial

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

score

sources of error

$$s(x,t) := \nabla_x \log p(x,t)$$

- score error
- integration error

common: exponential integrator

Y_t: exact solution, N/A due to nonlinearity

efficient integration:

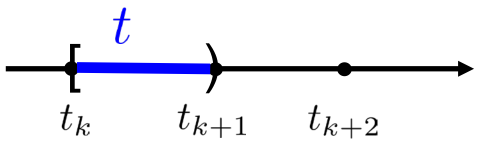
lots of research

e.g., DEIS, gDDIM
(*Zhang* & *Chen* 2022)
(*Zhang*, ***Tao*** & *Chen* 2022)

NFE~10-20

$$dY_t = Y_t\,dt + 2s(Y_t, T-t)\,dt + \sqrt{2}\,dB_t$$

$$d\hat{Y}_t = \hat{Y}_t\,dt + 2s(\hat{Y}_{\boxed{t_k}}, T-t_k)\,dt + \sqrt{2}\,dB_t$$



$\hat{Y}_t$: exactly solvable, a numerical approximation of Y_t

$\hat{Y}_{t_k} \mapsto \hat{Y}_{t_{k+1}}$: one function evaluation (of s)

2. **only** backward sampling/inference process:
numerical simulation → sample Y

already highly nontrivial

$$dX = -X dt + \sqrt{2} dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y dt + 2s(Y, T-t)dt + \sqrt{2} dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

sources of error

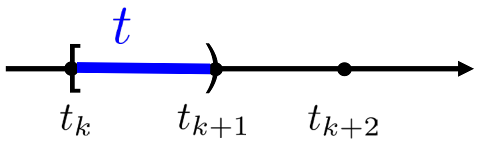
- score error
- integration error

common: exponential integrator

Y_t: exact solution, N/A due to nonlinearity

$$dY_t = Y_t dt + 2s(Y_t, T-t)dt + \sqrt{2} dB_t$$

$$d\hat{Y}_t = \hat{Y}_t dt + 2s(\hat{Y}_{\boxed{t_k}}, T-t_k)dt + \sqrt{2} dB_t$$



$\hat{Y}_t$: exactly solvable, a numerical approximation of Y_t

$\hat{Y}_{t_k} \mapsto \hat{Y}_{t_{k+1}}$: one function evaluation (of s)

efficient integration:

lots of research

e.g., DEIS, gDDIM
(*Zhang* & *Chen* 2022)
(*Zhang*, ***Tao*** & *Chen* 2022)

NFE~10-20

→ distillation, etc.

2. backward sampling/inference process: numerical simulation → sample Y

only

$$dX = -X dt + \sqrt{2} dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

already highly nontrivial

sources of error

$$dY = Y dt + 2s(Y, T-t)dt + \sqrt{2} dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

- score error
- integration error

common: exponential integrator

Y_t: exact solution, N/A due to nonlinearity

efficient integration:

lots of research

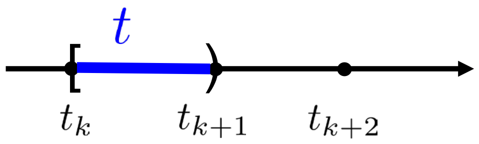
e.g., DEIS, gDDIM
(*Zhang* & *Chen* 2022)
(*Zhang*, ***Tao*** & *Chen* 2022)

NFE~10-20

→ distillation, etc.

$$dY_t = Y_t dt + 2s(Y_t, T-t)dt + \sqrt{2} dB_t$$

$$d\hat{Y}_t = \hat{Y}_t dt + 2s(\hat{Y}_{t_k}, T-t_k)dt + \sqrt{2} dB_t$$



$\hat{Y}_t$: exactly solvable, a numerical approximation of Y_t

$\hat{Y}_{t_k} \mapsto \hat{Y}_{t_{k+1}}$: one function evaluation (of s)

only

2. backward sampling/inference process:
   numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error

only 2. backward sampling/inference process: numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

ideally

$$Y(0) \overset{d}{=} X(\infty) \sim \mathcal{N}(0, I)$$

- score error
- integration error
- initialization error

only

2. backward sampling/inference process:
   numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error

ideally

$$Y(0) \stackrel{d}{=} X(\infty) \sim \mathcal{N}(0, I)$$

in reality

T large but finite

1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

2. backward sampling/inference process:
numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0,I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error

ideally

$$Y(0) \overset{d}{=} X(\infty) \sim \mathcal{N}(0,I)$$

in reality

T large but finite

then, ideally

$$Y(0) \overset{d}{=} X(T) \quad \Rightarrow \quad Y(T) \overset{d}{=} X(0)$$

only

2. backward sampling/inference process:
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error

ideally

$$Y(0) \overset{d}{=} X(\infty) \sim \mathcal{N}(0, I)$$

in reality

$$T \text{ large but finite}$$

then, ideally

$$Y(0) \overset{d}{=} X(T) \quad \Rightarrow \quad Y(T) \overset{d}{=} X(0)$$

but, in reality

$$\text{Law} X(T) \text{ unknown} \Rightarrow Y(0) \overset{d}{=} X(\infty) \sim \mathcal{N}(0, I)$$

only

2. backward sampling/inference process: numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error

ideally

$$Y(0) \overset{d}{=} X(\infty) \sim \mathcal{N}(0, I)$$

in reality

T large but finite

then, ideally

$$Y(0) \overset{d}{=} \boxed{X(T)}$$

but, in reality

$$\text{Law}X(T) \text{ unknown} \Rightarrow Y(0) \overset{d}{=} \boxed{X(\infty) \sim \mathcal{N}(0, I)}$$

discrepancy propagated thru Y dynamics → error

2. backward sampling/inference process: **only**
   numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error
- early stopping

2. backward sampling/inference process:
only
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error
- early stopping

common:
low dim. data manifold assumption

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

sources of error

- score error
- integration error
- initialization error
- early stopping

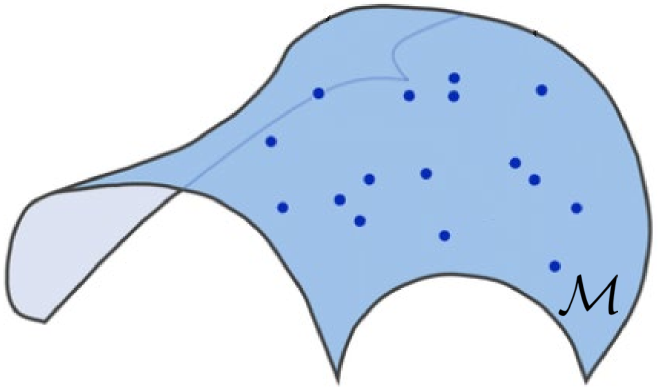
common:
low dim. data manifold assumption

$p(\cdot, 0)$ is supported on $\mathcal{M}$

$x \in \mathbb{R}^D, \quad D > \dim\mathcal{M}$

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$



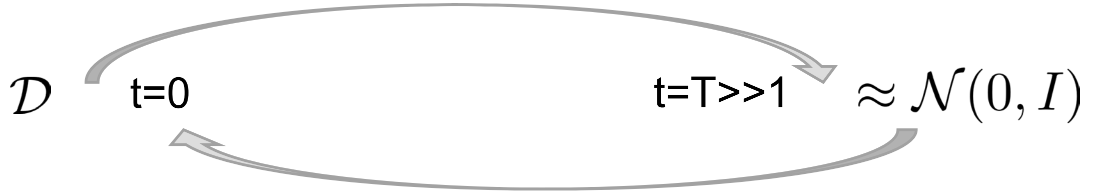
1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

2. backward sampling/inference process:
numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}\,dW_t$$



$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

score

sources of error

- score error
- integration error
- initialization error
- early stopping

common:
low dim. data manifold assumption

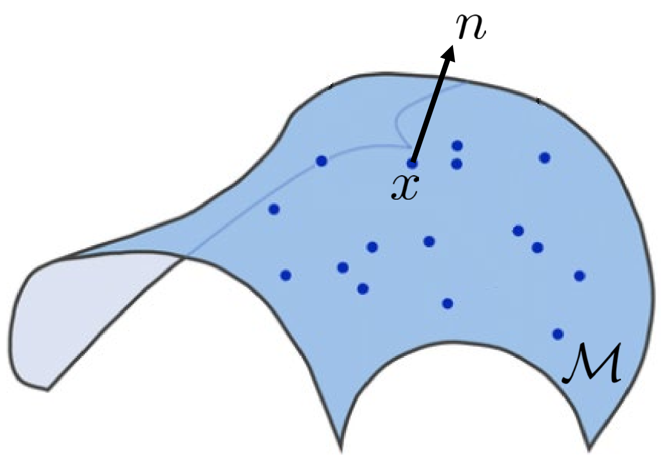
$p(\cdot, 0)$ is supported on $\mathcal{M}$

$x \in \mathbb{R}^D, \quad D > \dim \mathcal{M}$

$\Rightarrow p(\cdot, 0)$ has a jump in n direction

$$s(x, t) := \nabla_x \log p(x, t)$$

$$X(t) \sim p(x, t)$$



1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error
- early stopping

common:
low dim. data manifold assumption

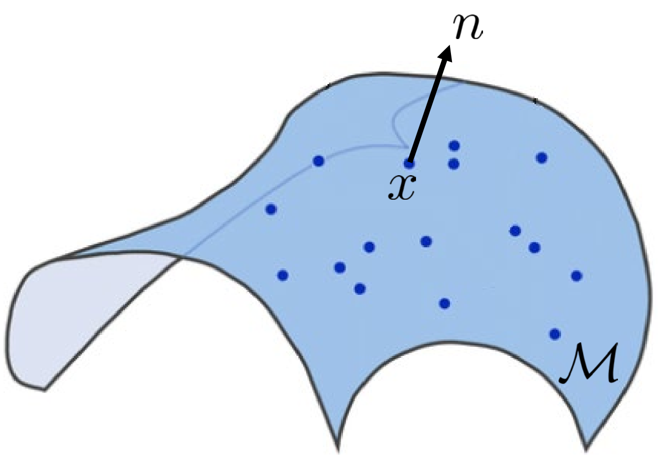
$p(\cdot, 0)$ is supported on $\mathcal{M}$

$x \in \mathbb{R}^D, \quad D > \dim\mathcal{M}$

$\Rightarrow p(\cdot, 0)$ has a jump in n direction

$\Rightarrow \langle \nabla_x \log p(x,0), n\rangle = \infty$

i.e. score at t=0 is ill-defined



2. backward sampling/inference process: **only**
numerical simulation → sample Y

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

already highly nontrivial

sources of error

- score error
- integration error
- initialization error
- early stopping

common:
low dim. data manifold assumption

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

$p(\cdot, 0)$ is supported on $\mathcal{M}$

$x \in \mathbb{R}^D, \quad D > \dim\mathcal{M}$

$\Rightarrow p(\cdot, 0)$ has a jump in n direction

$\Rightarrow \langle \nabla_x \log p(x,0), n \rangle = \infty$

⇒ Y(T) not usable

i.e. score at t=0 is ill-defined

2. backward sampling/inference process:
numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error
- early stopping

common:
low dim. data manifold assumption

$p(\cdot, 0)$ is supported on $\mathcal{M}$

$x \in \mathbb{R}^D, \quad D > \dim\mathcal{M}$

$\Rightarrow p(\cdot, 0)$ has a jump in n direction

$\Rightarrow \langle \nabla_x \log p(x,0), n \rangle = \infty$

i.e. score at t=0 is ill-defined

⇒ Y(T) not usable

use Y(T-δ) instead

only

2. backward sampling/inference process:
   numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error
- early stopping

common:
low dim. data manifold assumption

$p(\cdot, 0)$ is supported on $\mathcal{M}$

$x \in \mathbb{R}^D, \quad D > \dim\mathcal{M}$

$\Rightarrow p(\cdot,0)$ has a jump in n direction

$\Rightarrow \langle \nabla_x \log p(x,0), n\rangle = \infty$

i.e. score at t=0 is ill-defined

⇒ Y(T) not usable

use Y(T-δ) instead

accuracy quantifier

$$\mathrm{KL}\big(\mathrm{Law}(X_0)|\mathrm{Law}(Y_T)\big)$$

1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

only

2. backward sampling/inference process:
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

sources of error

- score error
- integration error
- initialization error
- early stopping

common:
low dim. data manifold assumption

$p(\cdot, 0)$ is supported on $\mathcal{M}$

$x \in \mathbb{R}^D, \quad D > \dim\mathcal{M}$

$\Rightarrow p(\cdot, 0)$ has a jump in n direction

$\Rightarrow \langle \nabla_x \log p(x, 0), n\rangle = \infty$

i.e. score at t=0 is ill-defined

Y(T) not usable

use Y(T-δ) instead

accuracy quantifier

$$\cancel{\mathrm{KL}\big(\mathrm{Law}(X_0)|\mathrm{Law}(Y_T)\big)} \longrightarrow \mathrm{KL}\big(\mathrm{Law}(X_\delta)|\mathrm{Law}(Y_{T-\delta})\big) + W_2(\mathrm{Law}(X_0), \mathrm{Law}(X_\delta))^2$$

only

2. backward sampling/inference process:
   numerical simulation → sample Y

already highly nontrivial

sources of error

assumptions on data distribution

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

sources of error

assumptions on data distribution

<u>Ex</u> (prior to diffusion model & its analysis)

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$\boxed{dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

sources of error

score

$$s(x,t) := \nabla_x \log p(x,t)$$

assumptions on data distribution

$$X(t) \sim p(x,t)$$

Ex (prior to diffusion model & its analysis)

(overdamped) Langevin dynamics:

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \xlongequal{V:=-\log p(\cdot,0)} s(Z_t, 0)dt + \sqrt{2}dB_t$$

only

2. backward sampling/inference process:
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

sources of error

score

$$s(x,t) := \nabla_x \log p(x,t)$$

assumptions on data distribution

$$X(t) \sim p(x,t)$$

<u>Ex</u> (prior to diffusion model & its analysis)

(overdamped) Langevin dynamics:

$$dZ_t = -\nabla V(Z_t)\,dt + \sqrt{2}\,dB_t \xlongequal{V := -\log p(\cdot, 0)} s(Z_t, 0)\,dt + \sqrt{2}\,dB_t \qquad Z(\infty) \sim p(\cdot, 0)$$

need long time to conv.

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

sources of error

score

$$s(x,t) := \nabla_x \log p(x,t)$$

assumptions on data distribution

Ex (prior to diffusion model & its analysis)

$$X(t) \sim p(x,t)$$

(overdamped) Langevin dynamics:

$$dZ_t = -\nabla V(Z_t)\,dt + \sqrt{2}\,dB_t \overset{V:=-\log p(\cdot,0)}{=\!=\!=\!=} s(Z_t, 0)\,dt + \sqrt{2}\,dB_t \qquad Z(\infty) \sim p(\cdot, 0)$$

How long?

need long time to conv.

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

sources of error

assumptions on data distribution

<u>Ex</u> (prior to diffusion model & its analysis)

(overdamped) Langevin dynamics:

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \overset{V:=-\log p(\cdot,0)}{=\!=\!=\!=\!=} s(Z_t, 0)dt + \sqrt{2}dB_t$$

How long? Depends on V and thus data distribution

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

$$Z(\infty) \sim p(\cdot, 0)$$

need long time to conv.

only

2. backward sampling/inference process: numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

sources of error

score

$$s(x,t) := \nabla_x \log p(x,t)$$

assumptions on data distribution

$$X(t) \sim p(x,t)$$

Ex (prior to diffusion model & its analysis)

(overdamped) Langevin dynamics:

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \xlongequal{V:=-\log p(\cdot,0)} s(Z_t, 0)dt + \sqrt{2}dB_t$$

$$Z(\infty) \sim p(\cdot, 0)$$

need long time to conv.

How long? Depends on V and thus data distribution

~1/m m-convex unimodal

2. backward sampling/inference process: numerical simulation → sample Y

only

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

already highly nontrivial

sources of error

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

assumptions on data distribution

$$X(t) \sim p(x,t)$$

<u>Ex</u> (prior to diffusion model & its analysis)

(overdamped) Langevin dynamics:

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \xlongequal{V:=-\log p(\cdot,0)} s(Z_t, 0)dt + \sqrt{2}dB_t \qquad Z(\infty) \sim p(\cdot, 0)$$

need long time to conv.

How long? Depends on V and thus data distribution

~1/m	m-convex	unimodal
can be exp. longer	nonconvex	multimodal

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

sources of error

assumptions on data distribution

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

need T time to ~~conv.~~ reach p(·,0)

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

<u>Ex</u> (prior to diffusion model & its analysis)

(overdamped) Langevin dynamics:

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \overset{V:=-\log p(\cdot,0)}{=\!=\!=\!=} s(Z_t, 0)dt + \sqrt{2}dB_t$$

$$Z(\infty) \sim p(\cdot, 0)$$

need long time to conv.

How long?	Depends on V and thus data distribution		
~1/m		m-convex	unimodal
can be exp. longer		nonconvex	multimodal

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

sources of error

assumptions on data distribution

need T time to ~~conv.~~ reach p(·,0)

agnostic to multimodality!

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

Ex (prior to diffusion model & its analysis)

(overdamped) Langevin dynamics:

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \overset{V:=-\log p(\cdot,0)}{=\!=\!=\!=} s(Z_t, 0)dt + \sqrt{2}dB_t$$

$$Z(\infty) \sim p(\cdot, 0)$$

need long time to conv.

How long? Depends on V and thus data distribution

~1/m	m-convex	unimodal
can be exp. longer	nonconvex	multimodal

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

sources of error

assumptions on data distribution

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

need T time to ~~conv.~~ reach p(·,0)

agnostic to multimodality!

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \xlongequal{V:=-\log p(\cdot,0)} s(Z_t, 0)dt + \sqrt{2}dB_t$$

$$Z(\infty) \sim p(\cdot, 0)$$

suffer from multimodality etc.

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

sources of error

assumptions on data distribution

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

need T time to ~~conv.~~ reach p(·,0)

agnostic to multimodality!

difference?

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \xlongequal{V:=-\log p(\cdot,0)} s(Z_t, 0)dt + \sqrt{2}dB_t$$

$$Z(\infty) \sim p(\cdot, 0)$$

suffer from multimodality etc.

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

sources of error

assumptions on data distribution

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0

shooting a moving target $I)$

$$dY = Y\,dt + 2s(Y, \boxed{T-t})dt + \sqrt{2}dB_t$$

score

need T time to ~~conv.~~ reach p(·,0)

$$s(x,t) := \nabla_x \log p(x,t)$$

agnostic to multimodality!

$$X(t) \sim p(x,t)$$

difference?

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \overset{V:=-\log p(\cdot,0)}{=\!=\!=\!=} s(Z_t, 0)dt + \sqrt{2}dB_t$$

$$Z(\infty) \sim p(\cdot, 0)$$

suffer from multimodality etc.

2. backward sampling/inference process:
numerical simulation → sample Y

only

already highly nontrivial

sources of error

assumptions on data distribution

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 $I)$

shooting a moving target

$$dY = Y\,dt + 2s(Y, \boxed{T-t})dt + \sqrt{2}dB_t$$

score

$$s(x,t) := \nabla_x \log p(x,t)$$

$$X(t) \sim p(x,t)$$

need T time to ~~conv.~~ reach p(·,0)

agnostic to multimodality!

difference?

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \overset{V:=-\log p(\cdot,0)}{=\!=\!=\!=\!=} s(Z_t, 0)dt + \sqrt{2}dB_t$$

$$Z(\infty) \sim p(\cdot, 0)$$

suffer from multimodality etc.

a specific annealing scheme made multimodal sampling effectively

2. only backward sampling/inference process: numerical simulation → sample Y

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 shooting a moving target $I)$

already highly nontrivial

sources of error

$$dY = Y\,dt + 2s(Y, \boxed{T-t})dt + \sqrt{2}dB_t$$

score

assumptions on data distribution

need T time to ~~conv.~~ reach p(·,0)

$$s(x,t) := \nabla_x \log p(x,t)$$

agnostic to multimodality!

$$X(t) \sim p(x,t)$$

difference?

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t \overset{V:=-\log p(\cdot,0)}{=\!=\!=\!=} s(Z_t, 0)dt + \sqrt{2}dB_t$$

$$Z(\infty) \sim p(\cdot, 0)$$

suffer from multimodality etc.

a specific annealing scheme made multimodal sampling effectively

Lee, *Risteski*, *Ge* 18
Chehab, *Hyvarinen*, *Risteski* 23

only

2. backward sampling/inference process:
   numerical simulation → sample Y

already highly nontrivial

sources of error

assumptions on data distribution

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 $I)$

shooting a moving target

$$dY = Y\,dt + 2s(Y, \boxed{T-t})\,dt + \sqrt{2}\,dB_t$$

score

need T time to ~~conv.~~ reach p(·,0)

$$s(x,t) := \nabla_x \log p(x,t)$$

agnostic to multimodality!

$$X(t) \sim p(x,t)$$

difference?

$$dZ_t = -\nabla V(Z_t)\,dt + \sqrt{2}\,dB_t \overset{V:=-\log p(\cdot,0)}{=\!=\!=\!=\!=} s(Z_t, 0)\,dt + \sqrt{2}\,dB_t$$

$$Z(\infty) \sim p(\cdot, 0)$$

suffer from multimodality etc.

a specific annealing scheme made multimodal sampling effectively

Lee, Risteski, Ge 18
Chehab, Hyvarinen, Risteski 23
*Guo, **Tao**, Chen* 24

only

2. backward sampling/inference process:
   numerical simulation → sample Y

$$dX = -X\,dt + \sqrt{2}\,dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

already highly nontrivial

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

sources of error

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

only

2. backward sampling/inference process:
   numerical simulation → sample Y

already highly nontrivial

sources of error

assumptions on data distribution

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

denoising diffusion annealing: **agnostic** to multimodality

ideally:

1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

2. backward sampling/inference process: only
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

sources of error — score, integration, initialization…

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

ideally: ✓

practically: nontrivial

2. backward sampling/inference process:
only
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

sources of error score, integration, initialization…

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

assumptions on data distribution

denoising diffusion annealing: agnostic to multimodality

ideally:

practically: nontrivial

errors propagate thru Y dynamics
in controlled fashion?

2. backward sampling/inference process: **only**
   numerical simulation → sample Y

   already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

sources of error: score, integration, initialization…

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

ideally: ✓

practically: nontrivial

errors propagate thru Y dynamics in controlled fashion?

- <u>strongly log concave</u>: contraction, relatively easy

only

2. backward sampling/inference process:
   numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

sources of error score, integration, initialization…

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

ideally: ✓

practically: nontrivial

errors propagate thru Y dynamics in controlled fashion?

- <u>strongly log concave</u>: contraction, relatively easy
- <u>isoperimetric ineq. (LSI, PI, …)</u>: ~2-3 years ago

2. backward sampling/inference process: **only**
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

sources of error score, integration, initialization…

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

ideally: ✓ practically: nontrivial

errors propagate thru Y dynamics in controlled fashion?

- strongly log concave: contraction, relatively easy
- isoperimetric ineq. (LSI, PI, …): ~2-3 years ago
- bounded 2^{nd} moment + Lipschitz score: ~1-2 years ago

1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

sources of error score, integration, initialization…

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

ideally: ✓

practically: nontrivial

errors propagate thru Y dynamics in controlled fashion?

- strongly log concave: contraction, relatively easy
- isoperimetric ineq. (LSI, PI, …): ~2-3 years ago
- bounded 2^{nd} moment + Lipschitz score: ~1-2 years ago
- bounded 2^{nd} moment: ~0-1 years ago

1.2 Quantification of Diffusion Model's Generation Quality: Existing Results - Sampling Only

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

sources of error score, integration, initialization…

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

ideally: ✓

practically: nontrivial

errors propagate thru Y dynamics in controlled fashion?

- strongly log concave: contraction, relatively easy
- isoperimetric ineq. (LSI, PI, …): ~2-3 years ago
- bounded 2^{nd} moment + Lipschitz score: ~1-2 years ago
- bounded 2^{nd} moment: ~0-1 years ago [*Benton*+ 23] [*Conforti*+ 23]

2. backward sampling/inference process: numerical simulation → sample Y

only

already highly nontrivial

$dX = -X dt + \sqrt{2} dW_t$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$dY = Y dt + 2s(Y, T-t)dt + \sqrt{2} dB_t$

sources of error score, integration, initialization…

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

ideally: ✓ practically: nontrivial

errors propagate thru Y dynamics in controlled fashion?

- <u>strongly log concave</u>: contraction, relatively easy
- <u>isoperimetric ineq. (LSI, PI, …)</u>: ~2-3 years ago
- <u>bounded 2nd moment + Lipschitz score</u>: ~1-2 years ago
- <u>bounded 2nd moment</u>: ~0-1 years ago [*Benton*+ 23] [*Conforti*+ 23]

time schedule

exponential

constant step

only

2. backward sampling/inference process:
numerical simulation → sample Y

already highly nontrivial

$$dX = -X\,dt + \sqrt{2}dW_t$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

sources of error score, integration, initialization…

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

ideally: ✓ practically: nontrivial

errors propagate thru Y dynamics
in controlled fashion?

- strongly log concave: contraction, relatively easy
- isoperimetric ineq. (LSI, PI, …): ~2-3 years ago
- bounded 2^{nd} moment + Lipschitz score: ~1-2 years ago
- bounded 2^{nd} moment: ~0-1 years ago [*Benton*+ 23] [*Conforti*+ 23]

time schedule

exponential
constant step

Variance Preserving
SDE

only

2. backward sampling/inference process:
numerical simulation → sample Y

already highly nontrivial

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

sources of error score, integration, initialization…

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

assumptions on data distribution

denoising diffusion annealing: **agnostic** to multimodality

ideally: ✓ practically: nontrivial

errors propagate thru Y dynamics in controlled fashion?

- strongly log concave: contraction, relatively easy
- isoperimetric ineq. (LSI, PI, …): ~2-3 years ago
- bounded 2^{nd} moment + Lipschitz score: ~1-2 years ago
- bounded 2^{nd} moment: ~0-1 years ago [*Benton*+ 23] [*Conforti*+ 23]

time schedule

exponential
constant step

Variance Preserving SDE

1.3 Quantification of Diffusion Model's Generation Quality: New Results - Sampling Part

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

Variance Preserving SDE

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

Variance Preserving SDE

more general

$$dX = -f(X,t)dt + g(t)dW_t$$

1.3 Quantification of Diffusion Model's Generation Quality: New Results - Sampling Part

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

Variance Preserving SDE

more general

$$dX = -f(X,t)\,dt + g(t)\,dW_t$$

$$dY = f(Y, T-t)\,dt + gg^\top s(Y, T-t)\,dt + g(T-t)\,dB_t$$

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

Variance Preserving SDE

more general

$$dX = -f(X,t)\,dt + g(t)\,dW_t$$

$$dY = f(Y, T-t)\,dt + gg^\top s(Y, T-t)\,dt + g(T-t)\,dB_t$$

popular version (e.g., EDM [*Karras*+ 22])

$$\boxed{dX = -X dt + \sqrt{2} dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y dt + 2s(Y, T - t)dt + \sqrt{2} dB_t$$

Variance Preserving SDE

more general

$$dX = -f(X, t)dt + g(t)dW_t$$

$$dY = f(Y, T - t)dt + gg^\top s(Y, T - t)dt + g(T - t)dB_t$$

popular version (e.g., EDM [*Karras*+ 22])

(generalized) Variance Exploding SDE

$$dX = g(t)dW_t$$

$$dY = g^2(T - t)s(Y, T - t)dt + g(T - t)dB_t$$

$$\boxed{dX = -X dt + \sqrt{2} dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y dt + 2s(Y, T-t)dt + \sqrt{2} dB_t$$

minimal data assumption

[*Benton*+ 23]

[*Conforti*+ 23]

\+ Variance Preserving SDE

popular version (e.g., EDM [*Karras*+ 22])

(generalized) Variance Exploding SDE

$$dX = g(t) dW_t$$

$$dY = g^2(T-t)s(Y, T-t)dt + g(T-t)dB_t$$

$$\boxed{dX = -X dt + \sqrt{2} dW_t}$$

minimal data assumption

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y dt + 2s(Y, T-t)dt + \sqrt{2} dB_t$$

[*Benton*+ 23]

[*Conforti*+ 23]

+ Variance Preserving SDE

new

popular version (e.g., EDM [*Karras*+ 22])

[*Wang*+ 24]

+ **(generalized) Variance Exploding SDE**

$$dX = g(t) dW_t$$

$$dY = g^2(T-t)s(Y, T-t)dt + g(T-t)dB_t$$

1.3 Quantification of Diffusion Model's Generation Quality: New Results - Sampling Part

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

minimal data assumption

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

[*Benton*+ 23]

[*Conforti*+ 23]

\+ Variance Preserving SDE

popular version (e.g., EDM [*Karras*+ 22])

new

[*Wang*+ 24]

\+ **(generalized) Variance Exploding SDE**

$$dX = g(t)\,dW_t$$

$$dY = g^2(T-t)s(Y, T-t)\,dt + g(T-t)\,dB_t$$

minimal data assumption

$$\boxed{dX = -X dt + \sqrt{2} dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y dt + 2s(Y, T-t)dt + \sqrt{2} dB_t$$

	time schedule		
[*Benton*+ 23] [*Conforti*+ 23]	exponential constant step	+	Variance Preserving SDE
new [*Wang*+ 24]	arbitrary	+	popular version (e.g., EDM [*Karras*+ 22]) **(generalized) Variance Exploding SDE**

$$dX = g(t) dW_t$$

$$dY = g^2(T-t)s(Y, T-t)dt + g(T-t)dB_t$$

minimal data assumption

$$\boxed{dX = -X dt + \sqrt{2} dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y dt + 2s(Y, T - t)dt + \sqrt{2} dB_t$$

	time schedule		
[*Benton*+ 23]	exponential	+	Variance Preserving SDE
[*Conforti*+ 23]	constant step		

new

popular version (e.g., EDM [*Karras*+ 22])

[*Wang*+ 24] arbitrary + **(generalized) Variance Exploding SDE**

$$dX = g(t) dW_t$$

motivation:

how to choose t_k (and g?)

$$dY = g^2(T - t)s(Y, T - t)dt + g(T - t)dB_t$$

2. backward sampling/inference process:
   numerical simulation → sample Y

new

[*Wang*+ 24]

arbitrary

+

popular version (e.g., EDM [*Karras*+ 22])

(generalized) Variance Exploding SDE

$$dX = g(t)dW_t$$

$$dY = g^2(T-t)s(Y, T-t)dt + g(T-t)dB_t$$

2. backward sampling/inference process: numerical simulation → sample Y

+

1. forward training/learning process: optimization → score s

★ bigger contribution

new

[*Wang*+ 24]

arbitrary

+

popular version (e.g., EDM [*Karras*+ 22])

(generalized) Variance Exploding SDE

$$dX = g(t)dW_t$$

$$dY = g^2(T-t)s(Y, T-t)dt + g(T-t)dB_t$$

*1.4 Quantification of Diffusion Model's Generation Quality: New Results – **Training** + Sampling*

approximating score is essential

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

1.4 *Quantification of Diffusion Model's Generation Quality: New Results – **Training** + Sampling*

approximating score is essential

“main stream” generation quality bound

if $\|s_\theta - s\| \leq \epsilon$, then $\cdots$

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

approximating score is essential

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

"main stream" generation quality bound

$$\text{if } \|\boxed{s_\theta} - s\| \leq \epsilon, \text{ then } \cdots$$

where does it come from **?**

approximating score is essential

“main stream” generation quality bound

if $\|\boxed{s_\theta} - s\| \le \epsilon$, then $\cdots$

where does it come from **?**

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

- nontrivial even when the target density is known (e.g., [*Huang*+ 24], [*He*, *Rojas*, ***Tao*** 24], [*Gupta*+ 24])

*1.4 Quantification of Diffusion Model's Generation Quality: New Results – **Training** + Sampling*

approximating score is essential

$$\boxed{dX = -X\,dt + \sqrt{2}\,dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)\,dt + \sqrt{2}\,dB_t$$

"main stream" generation quality bound

$$\text{if } \|\boxed{s_\theta} - s\| \le \epsilon, \text{ then } \cdots$$

where does it come from **?**

- nontrivial even when the target density is known (e.g., [*Huang*+ 24], [*He*, *Rojas*, ***Tao*** 24], [*Gupta*+ 24])
- it also matters to innovation / generalization (an open question, to be discussed in the end)

approximating score is essential

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

"main stream" generation quality bound

$$\text{if } \|\boxed{s_\theta} - s\| \leq \epsilon, \text{ then } \cdots$$

where does it come from **?**

- nontrivial even when the target density is known (e.g., [*Huang*+ 24], [*He*, *Rojas*, ***Tao*** 24], [*Gupta*+ 24])
- it also matters to innovation / generalization (an open question, to be discussed in the end)

ideal

$$\min_\theta \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0, t) - s_\theta(X_t, t)\|^2 dt$$

approximating score is essential

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

"main stream" generation quality bound

$$\text{if } \|\boxed{s_\theta} - s\| \leq \epsilon, \text{ then } \cdots$$

where does it come from **?**

- nontrivial even when the target density is known (e.g., [*Huang*+ 24], [*He*, *Rojas*, ***Tao*** 24], [*Gupta*+ 24])
- it also matters to innovation / generalization (an open question, to be discussed in the end)

ideal
$$\min_\theta \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0, t) - s_\theta(X_t, t)\|^2 dt$$

practice time discretization/sampling → empirical approximation → optimization

*1.4 Quantification of Diffusion Model's Generation Quality: New Results – **Training** + Sampling*

approximating score is essential

$$\boxed{dX = -X\,dt + \sqrt{2}dW_t}$$

$\mathcal{D}$ t=0 t=T>>1 $\approx \mathcal{N}(0, I)$

$$dY = Y\,dt + 2s(Y, T-t)dt + \sqrt{2}dB_t$$

"main stream" generation quality bound

$$\text{if } \|\boxed{s_\theta} - s\| \le \epsilon, \text{ then } \cdots$$

where does it come from **?**

- nontrivial even when the target density is known (e.g., [*Huang*+ 24], [*He*, *Rojas*, ***Tao*** 24], [*Gupta*+ 24])
- it also matters to innovation / generalization (an open question, to be discussed in the end)

ideal

$$\min_\theta \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0,t) - \overset{S(\theta;t,X_t)}{\cancel{s_\theta(X_t,t)}}\|^2 dt$$

practice time discretization/sampling → empirical approximation → optimization

ideal

$$\min_{\theta} \frac{1}{2} \int_0^T w(t) \mathbb{E}_{X_0} \mathbb{E}_{X_t|X_0} \|\nabla_x \log p_{t|0}(X_t|X_0, t) - S(\theta; t, X_t)\|^2 dt$$

ideal $$\min_{\theta} \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0,t) - S(\theta;t,X_t)\|^2 dt$$

$$\text{forward dynamics} \Rightarrow X_t = e^{-\mu_t}X_0 + \bar{\sigma}_t\xi$$

ideal

$$\min_{\theta} \frac{1}{2} \int_0^T w(t) \mathbb{E}_{X_0} \mathbb{E}_{X_t|X_0} \|\nabla_x \log p_{t|0}(X_t|X_0, t) - S(\theta; t, X_t)\|^2 dt$$

$$\text{forward dynamics} \Rightarrow X_t = e^{-\boxed{\mu_t}} X_0 + \bar{\sigma}_t \xi$$

0 for Variance Exploding SDE

ideal $$\min_\theta \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0,t) - S(\theta;t,X_t)\|^2 dt$$

$$\text{forward dynamics} \Rightarrow X_t = e^{-\boxed{\mu_t}}X_0 + \boxed{\bar{\sigma}_t}\xi$$

variance schedule

0 for Variance Exploding SDE

ideal

$$\min_{\theta} \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0,t) - S(\theta;t,X_t)\|^2 dt$$

$$\text{forward dynamics} \Rightarrow X_t = e^{-\boxed{\mu_t}}X_0 + \boxed{\bar{\sigma}_t}\xi$$

variance schedule

0 for Variance Exploding SDE

$$\min_{\theta} \frac{1}{2}\int_{t_0}^T w(t)\frac{1}{\bar{\sigma}_t}\mathbb{E}_{X_0}\mathbb{E}_{\xi}\|\bar{\sigma}_t S(\theta;t,X_0+\bar{\sigma}_t\xi) + \xi\|^2 dt$$

ideal

$$\min_\theta \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0,t) - S(\theta;t,X_t)\|^2 dt$$

$$\| \quad \text{forward dynamics} \Rightarrow X_t = e^{-\boxed{\mu_t}}X_0 + \boxed{\bar\sigma_t}\xi$$

variance schedule

0 for Variance Exploding SDE

$$\min_\theta \frac{1}{2}\int_{t_0}^T w(t)\frac{1}{\bar\sigma_t}\mathbb{E}_{X_0}\mathbb{E}_\xi\|\bar\sigma_t S(\theta;t,X_0+\bar\sigma_t\xi)+\xi\|^2 dt$$

time discretization

$$\min_\theta \underbrace{\frac{1}{2}\sum_{j=1}^N w(t_j)(t_j - t_{j-1})\frac{1}{\bar\sigma_{t_j}}\mathbb{E}_{X_0}\mathbb{E}_\xi\|\bar\sigma_{t_j}S(\theta;t_j,X_0+\bar\sigma_{t_j}\xi)+\xi\|^2}_{\bar{\mathcal{L}}(\theta)}$$

ideal

$$\min_\theta \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0,t) - S(\theta;t,X_t)\|^2 dt$$

$$\Big\| \quad \text{forward dynamics} \Rightarrow X_t = e^{-\boxed{\mu_t}}X_0 + \boxed{\bar{\sigma}_t}\xi$$

variance schedule

0 for Variance Exploding SDE

$$\min_\theta \frac{1}{2}\int_{t_0}^T w(t)\frac{1}{\bar{\sigma}_t}\mathbb{E}_{X_0}\mathbb{E}_{\xi}\|\bar{\sigma}_t S(\theta;t,X_0+\bar{\sigma}_t\xi)+\xi\|^2 dt$$

time schedule

time discretization

$$\min_\theta \underbrace{\frac{1}{2}\sum_{j=1}^N w(t_j)(t_j-t_{j-1})\frac{1}{\bar{\sigma}_{t_j}}\mathbb{E}_{X_0}\mathbb{E}_{\xi}\|\bar{\sigma}_{t_j}S(\theta;\boxed{t_j},X_0+\bar{\sigma}_{t_j}\xi)+\xi\|^2}_{\bar{\mathcal{L}}(\theta)}$$

ideal

$$\min_\theta \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0,t) - S(\theta;t,X_t)\|^2 dt$$

$$\|$$

$$\text{forward dynamics} \Rightarrow X_t = e^{-\boxed{\mu_t}}X_0 + \boxed{\bar\sigma_t}\xi$$

$\bar\sigma_t$: variance schedule

μ_t: 0 for Variance Exploding SDE

$$\min_\theta \frac{1}{2}\int_{t_0}^T w(t)\frac{1}{\bar\sigma_t}\mathbb{E}_{X_0}\mathbb{E}_\xi\|\bar\sigma_t S(\theta;t,X_0+\bar\sigma_t\xi)+\xi\|^2 dt$$

time discretization

t_j: time schedule

$$\min_\theta \underbrace{\frac{1}{2}\sum_{j=1}^N w(t_j)(t_j - t_{j-1})\frac{1}{\bar\sigma_{t_j}}\mathbb{E}_{X_0}\mathbb{E}_\xi\|\bar\sigma_{t_j}S(\theta;\boxed{t_j},X_0+\bar\sigma_{t_j}\xi)+\xi\|^2}_{\bar{\mathcal{L}}(\theta)}$$

empirical version

$$\min_\theta \underbrace{\frac{1}{2n}\sum_{i=1}^n\sum_{j=1}^N \beta_j\|\bar\sigma_{t_j}S(\theta;t_j,x_i+\bar\sigma_{t_j}\xi_{ij})+\xi_{ij}\|^2}_{\bar{\mathcal{L}}_{em}(\theta)}$$

ideal

$$\min_\theta \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0,t) - S(\theta;t,X_t)\|^2 dt$$

$$\Big\| \quad \text{forward dynamics} \Rightarrow X_t = e^{-\boxed{\mu_t}}X_0 + \boxed{\bar\sigma_t}\xi$$

variance schedule

0 for Variance Exploding SDE

$$\min_\theta \frac{1}{2}\int_{t_0}^T w(t)\frac{1}{\bar\sigma_t}\mathbb{E}_{X_0}\mathbb{E}_\xi\|\bar\sigma_t S(\theta;t,X_0+\bar\sigma_t\xi)+\xi\|^2 dt$$

time schedule

time discretization

$$\min_\theta \underbrace{\frac{1}{2}\sum_{j=1}^N \boxed{w(t_j)(t_j-t_{j-1})\frac{1}{\bar\sigma_{t_j}}}\mathbb{E}_{X_0}\mathbb{E}_\xi\|\bar\sigma_{t_j}S(\theta;\boxed{t_j},X_0+\bar\sigma_{t_j}\xi)+\xi\|^2}_{\bar{\mathcal{L}}(\theta)}$$

total weighting

empirical version

$$\min_\theta \underbrace{\frac{1}{2n}\sum_{i=1}^n\sum_{j=1}^N \boxed{\beta_j}\|\bar\sigma_{t_j}S(\theta;t_j,x_i+\bar\sigma_{t_j}\xi_{ij})+\xi_{ij}\|^2}_{\bar{\mathcal{L}}_{em}(\theta)}$$

ideal

$$\min_\theta \frac{1}{2}\int_0^T w(t)\mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\|\nabla_x \log p_{t|0}(X_t|X_0,t) - S(\theta;t,X_t)\|^2 dt$$

$$\Big\| \quad \text{forward dynamics} \Rightarrow X_t = e^{-\boxed{\mu_t}}X_0 + \boxed{\bar\sigma_t}\xi$$

variance schedule

0 for Variance Exploding SDE

$$\min_\theta \frac{1}{2}\int_{t_0}^T w(t)\frac{1}{\bar\sigma_t}\mathbb{E}_{X_0}\mathbb{E}_\xi\|\bar\sigma_t S(\theta;t,X_0+\bar\sigma_t\xi)+\xi\|^2 dt$$

time schedule

time discretization

$$\min_\theta \underbrace{\frac{1}{2}\sum_{j=1}^N \boxed{w(t_j)(t_j-t_{j-1})\frac{1}{\bar\sigma_{t_j}}}\mathbb{E}_{X_0}\mathbb{E}_\xi\|\bar\sigma_{t_j}S(\theta;\boxed{t_j},X_0+\bar\sigma_{t_j}\xi)+\xi\|^2}_{\bar{\mathcal{L}}(\theta)}$$

total weighting

empirical version

$$\min_\theta \underbrace{\frac{1}{2n}\sum_{i=1}^n\sum_{j=1}^N \boxed{\beta_j}\|\bar\sigma_{t_j}S(\theta;t_j,x_i+\bar\sigma_{t_j}\xi_{ij})+\xi_{ij}\|^2}_{\bar{\mathcal{L}}_{em}(\theta)}$$

training

$$\theta^{(k+1)} = \theta^{(k)} - h\nabla\bar{\mathcal{L}}_{em}(\theta^{(k)})$$

GD training

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N}\beta_j\|\bar{\sigma}_{t_j}S(\theta;t_j,x_i+\bar{\sigma}_{t_j}\xi_{ij})+\xi_{ij}\|^2$$

score parameterization $S(\theta; \cdot)$ wide (& deep) ReLU MLP

GD training $$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{N} \beta_j \|\bar{\sigma}_{t_j} S(\theta; t_j, x_i + \bar{\sigma}_{t_j} \xi_{ij}) + \xi_{ij}\|^2$$

score parameterization $S(\theta; \cdot)$ wide (& deep) ReLU MLP (challenge: U-Net? DiT?)

GD training

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{N} \beta_j \|\bar{\sigma}_{t_j} S(\theta; t_j, x_i + \bar{\sigma}_{t_j} \xi_{ij}) + \xi_{ij}\|^2$$

score parameterization $S(\theta;\cdot)$ wide (& deep) ReLU MLP (challenge: U-Net? DiT?)

GD training

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N}\underbrace{\boxed{\beta_j}}_{\text{total weighting}}\|\bar{\sigma}_{t_j} S(\theta;\underbrace{\boxed{t_j}}_{\text{time schedule}}, x_i + \underbrace{\boxed{\bar{\sigma}_{t_j}}}_{\text{variance schedule}}\xi_{ij}) + \xi_{ij}\|^2$$

score parameterization

$S(\theta;\cdot)$ wide (& deep) ReLU MLP (challenge: U-Net? DiT?)

GD training

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N}\boxed{\beta_j}\|\bar{\sigma}_{t_j}S(\theta;\boxed{t_j}, x_i + \boxed{\bar{\sigma}_{t_j}}\xi_{ij}) + \xi_{ij}\|^2$$

total weighting

time schedule

variance schedule

roadmap

non-asymptotic bound of GD optimization of $\bar{\mathcal{L}}_{em}$

score parameterization $S(\theta;\cdot)$ wide (& deep) ReLU MLP (challenge: U-Net? DiT?)

total weighting

time schedule

GD training
$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N}\boxed{\beta_j}\|\bar{\sigma}_{t_j} S(\theta;\boxed{t_j}, x_i + \boxed{\bar{\sigma}_{t_j}}\xi_{ij}) + \xi_{ij}\|^2$$

variance schedule

roadmap

C>0

non-asymptotic bound of GD optimization of $\bar{\mathcal{L}}_{em} \gtrsim C$

score parameterization $S(\theta;\cdot)$ wide (& deep) ReLU MLP (challenge: U-Net? DiT?)

GD training

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N}\boxed{\beta_j}\|\bar{\sigma}_{t_j}S(\theta;\boxed{t_j},x_i+\boxed{\bar{\sigma}_{t_j}}\xi_{ij})+\xi_{ij}\|^2$$

total weighting

time schedule

variance schedule

roadmap

C>0 → interesting generalization setting

non-asymptotic bound of GD optimization of $\bar{\mathcal{L}}_{em} \gtrsim C$

*1.4 Quantification of Diffusion Model's Generation Quality: New Results – **Training** + Sampling*

score parameterization $S(\theta;\cdot)$ wide (& deep) ReLU MLP (challenge: U-Net? DiT?)

GD training

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N}\boxed{\beta_j}\|\bar{\sigma}_{t_j} S(\theta;\boxed{t_j}, x_i + \boxed{\bar{\sigma}_{t_j}}\xi_{ij}) + \xi_{ij}\|^2$$

total weighting

time schedule

variance schedule

roadmap

C>0 → interesting generalization setting

non-asymptotic bound of GD optimization of $\bar{\mathcal{L}}_{em} \gtrsim C$

faster convergence

choice of total weighting

score parameterization $S(\theta; \cdot)$ wide (& deep) ReLU MLP (challenge: U-Net? DiT?)

GD training

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{N} \boxed{\beta_j} \|\bar{\sigma}_{t_j} S(\theta; \boxed{t_j}, x_i + \boxed{\bar{\sigma}_{t_j}} \xi_{ij}) + \xi_{ij}\|^2$$

total weighting

time schedule

variance schedule

roadmap

C>0 → interesting generalization setting

non-asymptotic bound of GD optimization of $\bar{\mathcal{L}}_{em} \gtrsim C$

faster convergence

choice of total weighting

analysis of sampling, i.e. generation/inference

score parameterization $S(\theta;\cdot)$ wide (& deep) ReLU MLP (challenge: U-Net? DiT?)

GD training

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N}\boxed{\beta_j}\|\bar{\sigma}_{t_j}S(\theta;\boxed{t_j}, x_i + \boxed{\bar{\sigma}_{t_j}}\xi_{ij}) + \xi_{ij}\|^2$$

total weighting

time schedule

variance schedule

roadmap

C>0 → interesting generalization setting

non-asymptotic bound of GD optimization of $\bar{\mathcal{L}}_{em} \gtrsim C$

faster convergence

choice of total weighting

analysis of sampling, i.e. generation/inference

full bound of generation quality

*1.4 Quantification of Diffusion Model's Generation Quality: New Results – **Training** + Sampling*

score parameterization $S(\theta;\cdot)$ wide (& deep) ReLU MLP (challenge: U-Net? DiT?)

GD training

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N}\boxed{\beta_j}\|\bar{\sigma}_{t_j} S(\theta;\boxed{t_j}, x_i + \boxed{\bar{\sigma}_{t_j}}\xi_{ij}) + \xi_{ij}\|^2$$

total weighting
time schedule
variance schedule

roadmap

non-asymptotic bound of GD optimization of $\bar{\mathcal{L}}_{em} \gtrsim C$

C>0 → interesting generalization setting

faster convergence → choice of total weighting

choice of total weighting → preferred time & variance schedules

analysis of sampling, i.e. generation/inference → **full** bound of generation quality

full bound of generation quality → accurate generation → preferred time & variance schedules

Setup

- Objective function (minimized by GD):

Setup

- Objective function (minimized by GD):

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N} \underbrace{\underbrace{\frac{w(t_j)}{\bar{\sigma}_{t_j}}(t_j - t_{j-1})}_{\beta_j} \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2}_{f(\theta;i,j)}$$

Setup

- Objective function (minimized by GD):

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N} \underbrace{\underbrace{\frac{w(t_j)}{\bar{\sigma}_{t_j}}(t_j - t_{j-1})}_{\beta_j} \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2}_{f(\theta;i,j)}$$

- Architecture: deep ReLU network

Setup

- Objective function (minimized by GD):

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{N} \underbrace{\underbrace{\frac{w(t_j)}{\bar{\sigma}_{t_j}}(t_j - t_{j-1})}_{\beta_j} \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2}_{f(\theta;i,j)}$$

- Architecture: deep ReLU network

$$S(\theta; X_{ij}) = W_{L+1}\sigma(W_L \cdots W_1\sigma(W_0[X_{ij}, t_j])),$$

Setup

- Objective function (minimized by GD):

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{N} \underbrace{\underbrace{\frac{w(t_j)}{\bar{\sigma}_{t_j}}(t_j - t_{j-1})}_{\beta_j} \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2}_{f(\theta;i,j)}$$

- Architecture: deep ReLU network

$$S(\theta; X_{ij}) = W_{L+1}\sigma(W_L \cdots W_1\sigma(W_0[X_{ij}, t_j])),$$

  - $W_{L+1} \in \mathbb{R}^{d\times m}$, $W_\ell \in \mathbb{R}^{m\times m}$, $W_0 \in \mathbb{R}^{m\times d}$

Setup

- Objective function (minimized by GD):

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{N} \underbrace{\underbrace{\frac{w(t_j)}{\bar{\sigma}_{t_j}}(t_j - t_{j-1})}_{\beta_j} \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2}_{f(\theta;i,j)}$$

- Architecture: deep ReLU network

$$S(\theta; X_{ij}) = W_{L+1}\sigma(W_L \cdots W_1\sigma(W_0[X_{ij}, t_j])),$$

  - $W_{L+1} \in \mathbb{R}^{d\times m}$, $W_\ell \in \mathbb{R}^{m\times m}$, $W_0 \in \mathbb{R}^{m\times d}$
  - $\theta := (W_0, W_1, \cdots, W_L, W_{L+1})$

Setup

- Input data (t_j, X_{ij}):

Setup

- Input data (t_j, X_{ij}):

$$X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij} \sim P_{t_j}(X|X(0) = x_i),$$

Setup

- Input data (t_j, X_{ij}):

$$X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij} \sim P_{t_j}(X|X(0) = x_i),$$

where $\bar{\sigma}^2_{t_j}$ is the variance of $X_{t_j}|X_0$, and $\xi_{ij} \sim \mathcal{N}(0, I)$

Setup

- Input data (t_j, X_{ij}):

$$X_{ij} = x_i + \bar{\sigma}_{t_j} \xi_{ij} \sim P_{t_j}(X | X(0) = x_i),$$

where $\bar{\sigma}_{t_j}^2$ is the variance of $X_{t_j} | X_0$, and $\xi_{ij} \sim \mathcal{N}(0, I)$

  - $\bar{\sigma}_t$: monotonically increasing functions of t; $\bar{\sigma}_0 = 0$

Setup

- Input data (t_j, X_{ij}):

$$X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij} \sim P_{t_j}(X|X(0) = x_i),$$

where $\bar{\sigma}^2_{t_j}$ is the variance of $X_{t_j}|X_0$, and $\xi_{ij} \sim \mathcal{N}(0, I)$
  - $\bar{\sigma}_t$: monotonically increasing functions of t; $\bar{\sigma}_0 = 0$
- Output data:

$$\frac{-\xi_{ij}}{\bar{\sigma}_{t_j}}$$

Setup

- Input data (t_j, X_{ij}):

$$X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij} \sim P_{t_j}(X|X(0) = x_i),$$

where $\bar{\sigma}_{t_j}^2$ is the variance of $X_{t_j}|X_0$, and $\xi_{ij} \sim \mathcal{N}(0, I)$
  - $\bar{\sigma}_t$: monotonically increasing functions of t; $\bar{\sigma}_0 = 0$
- Output data:

$$\frac{-\xi_{ij}}{\bar{\sigma}_{t_j}}$$

  - Very large if $\bar{\sigma}_{t_j} \approx 0$

Setup

- Number of samples:

$$X_{ij} = \underbrace{x_i}_{\substack{n \text{ samples} \\ \text{from initial} \\ \text{distribution } P_0}} + \underbrace{\bar{\sigma}_{t_j}}_{\substack{N \text{ time points} \\ \text{for forward/backward SDEs}}} \xi_{ij}$$

Theory: assumptions

- Assumptions: mild + preserve the nature of diffusion models

Theory: assumptions

- Assumptions: mild + preserve the nature of diffusion models
- For example,

Theory: assumptions

- Assumptions: mild + preserve the nature of diffusion models
- For example,

Assumption

- *Data scaling:* $\|x_i\| = \Theta(\sqrt{d})$ *for all* i.

Theory: assumptions

- Assumptions: mild + preserve the nature of diffusion models
- For example,

Assumption

- *Data scaling:* $\|x_i\| = \Theta(\sqrt{d})$ *for all* i*.*

Interpretation:

Theory: assumptions

- Assumptions: mild + preserve the nature of diffusion models
- For example,

Assumption

- *Data scaling:* $\|x_i\| = \Theta(\sqrt{d})$ *for all* i.

Interpretation:

- Recall input data:

Theory: assumptions

- Assumptions: mild + preserve the nature of diffusion models
- For example,

Assumption

- *Data scaling:* $\|x_i\| = \Theta(\sqrt{d})$ *for all* i.

Interpretation:

- Recall input data:

$$X_{ij} = x_i + \bar{\sigma}_{t_j} \xi_{ij}$$

Theory: assumptions

- Assumptions: mild + preserve the nature of diffusion models
- For example,

Assumption

- *Data scaling:* $\|x_i\| = \Theta(\sqrt{d})$ *for all* i.

Interpretation:

- Recall input data:

$$X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij}$$

- $\xi_{ij} \sim \mathcal{N}(0, I) \Rightarrow \|\xi_{ij}\| \approx \sqrt{d}$

Theory: assumptions

- Assumptions: mild + preserve the nature of diffusion models
- For example,

Assumption

- *Data scaling:* $\|x_i\| = \Theta(\sqrt{d})$ *for all* i.

Interpretation:

- Recall input data:

$$X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij}$$

- $\xi_{ij} \sim \mathcal{N}(0, I) \Rightarrow \|\xi_{ij}\| \approx \sqrt{d}$

Theory: training

Theorem

For any $\epsilon_{\text{train}} > 0$, *consider* $m \geq M(\epsilon_{\text{train}})$. *With high probability,*

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq \prod_{s=0}^{k-1} \left(1 - C_5 h\, w(t_{j^*(s)})(t_{j^*(s)} - t_{j^*(s)-1})\bar{\sigma}_{t_{j^*(s)}} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right) \bar{\mathcal{L}}_{em}(\theta^{(0)})$$

Moreover, when $K = \Theta(d^{\frac{1-a_0}{2}} n^2 N \log(\frac{d}{\epsilon_{\text{train}}}))$,

$$\bar{\mathcal{L}}_{em}(\theta^{(K)}) \leq \epsilon_{\text{train}}.$$

Theory: training

Theorem

For any $\epsilon_{\text{train}} > 0$, *consider* $m \geq M(\epsilon_{\text{train}})$. *With high probability,*

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq \prod_{s=0}^{k-1} \left(1 - C_5 h\, w(t_{j^*(s)})(t_{j^*(s)} - t_{j^*(s)-1})\bar{\sigma}_{t_{j^*(s)}} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right) \bar{\mathcal{L}}_{em}(\theta^{(0)})$$

Moreover, when $K = \Theta(d^{\frac{1-a_0}{2}} n^2 N \log(\frac{d}{\epsilon_{\text{train}}}))$,

$$\bar{\mathcal{L}}_{em}(\theta^{(K)}) \leq \epsilon_{\text{train}}.$$

- Exponential decay

Theory: training

Theorem

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq$$

$$\prod_{s=0}^{k-1}\left(1 - C_5 h\, w(t_{j^*(s)})(t_{j^*(s)} - t_{j^*(s)-1})\bar{\sigma}_{t_{j^*(s)}} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right)\bar{\mathcal{L}}_{em}(\theta^{(0)})$$

Theory: training

Theorem

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq \prod_{s=0}^{k-1} \left(1 - C_5 h \, w(t_{j^*(s)})(t_{j^*(s)} - t_{j^*(s)-1})\bar{\sigma}_{t_{j^*(s)}} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right) \bar{\mathcal{L}}_{em}(\theta^{(0)})$$

- Recall:

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n} \sum_{i=1,j=1}^{n,N} f(\theta; i, j)$$

Theory: training

Theorem

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq$$

$$\prod_{s=0}^{k-1}\left(1 - C_5 h\, w(t_{j^*(s)})(t_{j^*(s)} - t_{j^*(s)-1})\bar{\sigma}_{t_{j^*(s)}} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right)\bar{\mathcal{L}}_{em}(\theta^{(0)})$$

- Recall:

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1,j=1}^{n,N} f(\theta; i, j)$$

- $(i^*(s), j^*(s)) =$ the index of the largest loss $f(\theta^{(s)}; i, j)$

Theory: training

Theorem

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq$$
$$\prod_{s=0}^{k-1}\left(1 - C_5 h\; w(t_{j^*(s)})(t_{j^*(s)} - t_{j^*(s)-1})\bar{\sigma}_{t_{j^*(s)}} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right) \bar{\mathcal{L}}_{em}(\theta^{(0)})$$

- Recall:

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n} \sum_{i=1,j=1}^{n,N} f(\theta; i, j)$$

- $(i^*(s), j^*(s))$ = the index of the largest loss $f(\theta^{(s)}; i, j)$
- Faster convergence:

Theory: training

Theorem

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq$$
$$\prod_{s=0}^{k-1}\left(1 - C_5 h\, w(t_{j^*(s)})(t_{j^*(s)} - t_{j^*(s)-1})\bar{\sigma}_{t_{j^*(s)}} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right) \bar{\mathcal{L}}_{em}(\theta^{(0)})$$

- Recall:

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n}\sum_{i=1,j=1}^{n,N} f(\theta; i, j)$$

- $(i^*(s), j^*(s))$ = the index of the largest loss $f(\theta^{(s)}; i, j)$
- Faster convergence:

want to maximize over all the indices

Theory: training

Theorem

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq \prod_{s=0}^{k-1} \left(1 - C_5 h \, w(t_{j^*(s)})(t_{j^*(s)} - t_{j^*(s)-1})\bar{\sigma}_{t_{j^*(s)}} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right) \bar{\mathcal{L}}_{em}(\theta^{(0)})$$

- Recall:

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2n} \sum_{i=1,j=1}^{n,N} f(\theta; i, j)$$

- $(i^*(s), j^*(s))$ = the index of the largest loss $f(\theta^{(s)}; i, j)$
- Faster convergence:

want to maximize over all the indices

$\Rightarrow$ want all $f(\theta^{(s)}; i, j)$ to be the largest loss

Total weighting: theory vs practice

Corollary

When $f(\theta^{(k)}; i, j) \approx f(\theta^{(k)}; i', j')$ for all $(i, j), (i', j'), k$, GD obtains the optimal rate of convergence

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq \left(1 - C_7 h \max_{j=1,\cdots,N} w(t_j)(t_j - t_{j-1})\bar{\sigma}_{t_j} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right)^k \bar{\mathcal{L}}_{em}(\theta^{(0)}).$$

Total weighting: theory vs practice

Corollary

When $f(\theta^{(k)}; i, j) \approx f(\theta^{(k)}; i', j')$ for all $(i, j), (i', j'), k$, GD obtains the optimal rate of convergence

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq \left(1 - C_7 h \max_{j=1,\cdots,N} w(t_j)(t_j - t_{j-1})\bar{\sigma}_{t_j} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right)^k \bar{\mathcal{L}}_{em}(\theta^{(0)}).$$

- $f(\theta^{(k)}; i, j) \approx f(\theta^{(k)}; i', j')$:

Total weighting: theory vs practice

Corollary

When $f(\theta^{(k)}; i, j) \approx f(\theta^{(k)}; i', j')$ for all $(i, j), (i', j'), k$, GD obtains the optimal rate of convergence

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq \left(1 - C_7 h \max_{j=1,\cdots,N} w(t_j)(t_j - t_{j-1})\bar{\sigma}_{t_j} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right)^k \bar{\mathcal{L}}_{em}(\theta^{(0)}).$$

- $f(\theta^{(k)}; i, j) \approx f(\theta^{(k)}; i', j')$:

$$f(\theta; i, j) = \underbrace{\beta_j}_{\text{total weighting}} \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$$

Total weighting: theory vs practice

Corollary

When $f(\theta^{(k)}; i, j) \approx f(\theta^{(k)}; i', j')$ for all $(i, j), (i', j'), k$, GD obtains the optimal rate of convergence

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq \left(1 - C_7 h \max_{j=1,\cdots,N} w(t_j)(t_j - t_{j-1})\bar{\sigma}_{t_j} \frac{md^{\frac{a_0-1}{2}}}{n^3 N^2}\right)^k \bar{\mathcal{L}}_{em}(\theta^{(0)}).$$

- $f(\theta^{(k)}; i, j) \approx f(\theta^{(k)}; i', j')$:

$$f(\theta; i, j) = \underbrace{\beta_j}_{\text{total weighting}} \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$$

- Claim: this implies how to choose the total weighting β_j

Total weighting: theory vs practice

$$f(\theta; i, j) = \beta_j \, \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$$

Total weighting: theory vs practice

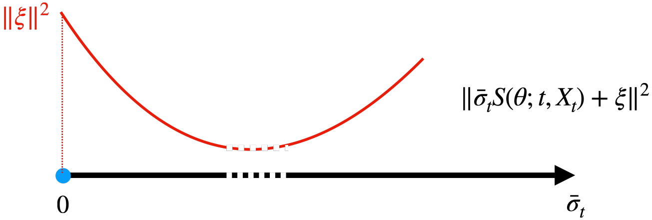
$f(\theta; i, j) = \beta_j \, \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$

- $\|\bar{\sigma}_t S(\theta; t, x_0 + \bar{\sigma}_t \xi) + \xi\|^2$:

Total weighting: theory vs practice

$$f(\theta; i, j) = \beta_j \; \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$$

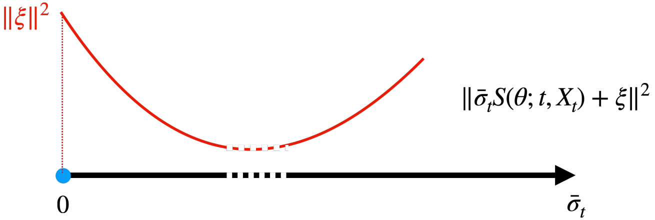
- $\|\bar{\sigma}_t S(\theta; t, x_0 + \bar{\sigma}_t \xi) + \xi\|^2$:



Total weighting: theory vs practice

$f(\theta; i, j) = \beta_j \, \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$

- $\|\bar{\sigma}_t S(\theta; t, x_0 + \bar{\sigma}_t \xi) + \xi\|^2$:

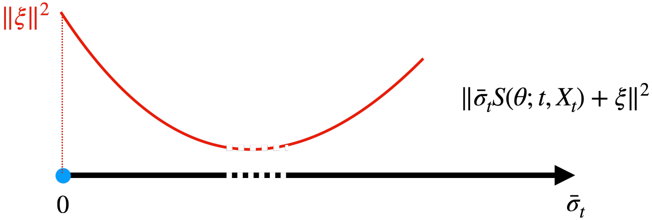


- Thus, $\beta_j \propto \frac{1}{\|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2}$

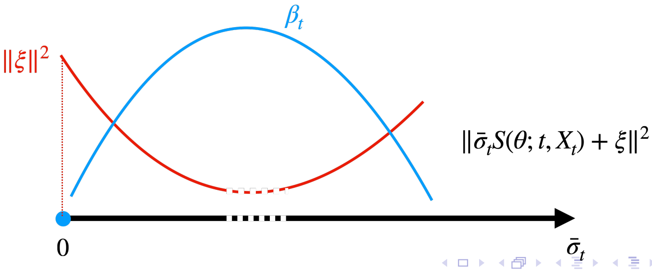
Total weighting: theory vs practice

$$f(\theta; i, j) = \beta_j \; \|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$$

- $\|\bar{\sigma}_t S(\theta; t, x_0 + \bar{\sigma}_t \xi) + \xi\|^2$:



- Thus, $\beta_j \propto \frac{1}{\|\bar{\sigma}_{t_j} S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2}$



Total weighting: theory vs practice

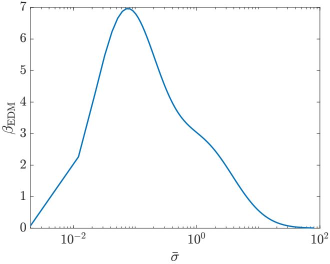
In practice:

- EDM [Karras et al., 2022]

Total weighting: theory vs practice

In practice:

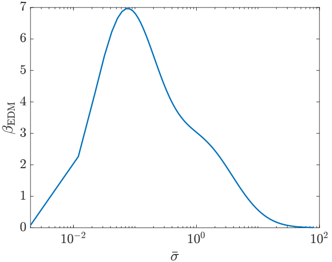
- EDM [Karras et al., 2022]



Total weighting: theory vs practice

In practice:

- EDM [Karras et al., 2022]

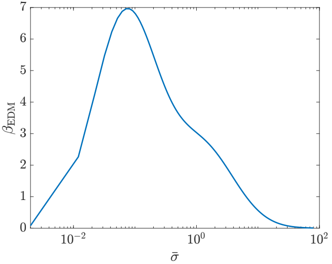


- Other total weighting functions used in practice (mostly monotone): e.g., $\beta_{\bar{\sigma}} = \frac{1}{\bar{\sigma}}$ [Song et al., 2021]

Total weighting: theory vs practice

In practice:

- EDM [Karras et al., 2022]

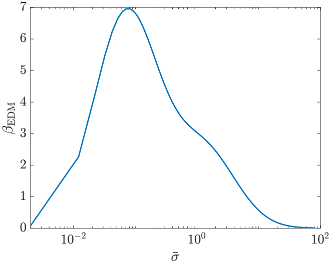


- Other total weighting functions used in practice (mostly monotone): e.g., $\beta_{\bar{\sigma}} = \frac{1}{\bar{\sigma}}$ [Song et al., 2021]
- Performance: EDM $>$ other models

Total weighting: theory vs practice

In practice:

- EDM [Karras et al., 2022]

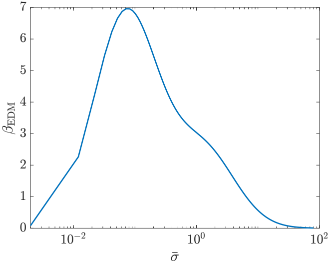


- Other total weighting functions used in practice (mostly monotone): e.g., $\beta_{\bar{\sigma}} = \frac{1}{\bar{\sigma}}$ [Song et al., 2021]
- Performance: EDM $>$ other models
  $\Rightarrow$ "bell-shape" is preferable

Total weighting: theory vs practice

In practice:

- EDM [Karras et al., 2022]



- Other total weighting functions used in practice (mostly monotone): e.g., $\beta_{\bar{\sigma}} = \frac{1}{\bar{\sigma}}$ [Song et al., 2021]
- Performance: EDM $>$ other models
  $\Rightarrow$ "bell-shape" is preferable
- Roughly, **Theory** $\approx$ **Practice**

Proof of convergence

- Recall: input data $X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij}$,
  output data $-\frac{\xi_{ij}}{\bar{\sigma}_{t_j}}$,
  where $x_i \sim P_0, \ \xi_{ij} \sim \mathcal{N}(0, I)$

Proof of convergence

- Recall: input data $X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij}$,
  output data $-\frac{\xi_{ij}}{\bar{\sigma}_{t_j}}$,
  where $x_i \sim P_0,\ \xi_{ij} \sim \mathcal{N}(0, I)$
- Framework [Allen-Zhu et al., 2019]:

Proof of convergence

- Recall: input data $X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij}$,
  output data $-\frac{\xi_{ij}}{\bar{\sigma}_{t_j}}$,
  where $x_i \sim P_0,\ \xi_{ij} \sim \mathcal{N}(0, I)$
- Framework [Allen-Zhu et al., 2019]:

$$\text{semi-smoothness} + \underbrace{\text{local strongly convex}}_{\text{key}}$$

Proof of convergence

- Recall: input data $X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij}$,
  output data $-\frac{\xi_{ij}}{\bar{\sigma}_{t_j}}$,
  where $x_i \sim P_0, \ \xi_{ij} \sim \mathcal{N}(0, I)$
- Framework [Allen-Zhu et al., 2019]:

$$\text{semi-smoothness} + \underbrace{\text{local strongly convex}}_{\text{key}}$$

  - No longer works in denoising diffusion models

Proof of convergence

- Recall: input data $X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij}$,

  output data $-\frac{\xi_{ij}}{\bar{\sigma}_{t_j}}$,

  where $x_i \sim P_0,\ \xi_{ij} \sim \mathcal{N}(0, I)$
- Framework [Allen-Zhu et al., 2019]:

$$\text{semi-smoothness} + \underbrace{\color{red}{\text{local strongly convex}}}_{\color{red}{\text{key}}}$$

  - No longer works in denoising diffusion models
  - Reason: (1) $\underbrace{\text{scaling}}_{\substack{\text{bad} \\ \xcancel{\text{small output data}}}}$;

Proof of convergence

- Recall: input data $X_{ij} = x_i + \bar{\sigma}_{t_j} \xi_{ij}$,
  output data $-\frac{\xi_{ij}}{\bar{\sigma}_{t_j}}$,
  where $x_i \sim P_0$, $\xi_{ij} \sim \mathcal{N}(0, I)$
- Framework [Allen-Zhu et al., 2019]:

$$\text{semi-smoothness} + \underbrace{\text{local strongly convex}}_{\text{key}}$$

  - No longer works in denoising diffusion models
  - Reason: (1) $\underbrace{\text{scaling}}_{\substack{\text{bad} \\ \cancel{\text{small output data}}}}$;(2) $\underbrace{\text{correlation}}_{\substack{\text{``good''} \\ \cancel{\text{data separability}}}}$

Proof of convergence

- Recall: input data $X_{ij} = x_i + \bar{\sigma}_{t_j}\xi_{ij}$,

  output data $-\frac{\xi_{ij}}{\bar{\sigma}_{t_j}}$,

  where $x_i \sim P_0,\ \xi_{ij} \sim \mathcal{N}(0, I)$
- Framework [Allen-Zhu et al., 2019]:

  $$\text{semi-smoothness} + \underbrace{\text{local strongly convex}}_{\text{key}}$$

  - No longer works in denoising diffusion models
  - Reason: (1) $\underbrace{\text{scaling}}_{\text{bad}}$;(2) $\underbrace{\text{correlation}}_{\text{“good”}}$

    ~~small output data~~ ~~data separability~~
  - Our proof: high probability bound using some high-dimensional geometry facts

Theory: more about training

Recall:

Theorem

For any $\epsilon_{\text{train}} > 0$, consider $m \geq M(\epsilon_{\text{train}})$. With high probability,

$$\bar{\mathcal{L}}_{em}(\theta^{(k)}) \leq \prod_{s=0}^{k-1} \left(1 - C_5 h\, w(t_{j^*(s)})(t_{j^*(s)} - t_{j^*(s)-1})\bar{\sigma}_{t_{j^*(s)}} \frac{m d^{\frac{a_0-1}{2}}}{n^3 N^2}\right) \bar{\mathcal{L}}_{em}(\theta^{(0)})$$

Moreover, when $K = \Theta(d^{\frac{1-a_0}{2}} n^2 N \log(\frac{d}{\epsilon_{\text{train}}}))$,

$$\bar{\mathcal{L}}_{em}(\theta^{(K)}) \leq \epsilon_{\text{train}}.$$

Theory: more about training

Properties of denoising score matching objective:

Theory: more about training

Properties of denoising score matching objective:

- $\bar{\mathcal{L}}_{em}(\theta) \to \bar{\mathcal{L}}$ as $n \to \infty$

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2}\sum_{j=1}^{N} w(t_j)(t_j - t_{j-1})\frac{1}{\bar{\sigma}_{t_j}}\frac{1}{n}\sum_{i=1}^{n}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$$

$$\bar{\mathcal{L}}(\theta) = \frac{1}{2}\sum_{j=1}^{N} w(t_j)(t_j - t_{j-1})\frac{1}{\bar{\sigma}_{t_j}}\mathbb{E}_{X_0}\mathbb{E}_{\xi}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{t_j}) + \xi\|^2$$

Theory: more about training

Properties of denoising score matching objective:

- $\bar{\mathcal{L}}_{em}(\theta) \to \bar{\mathcal{L}}$ as $n \to \infty$

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2}\sum_{j=1}^{N} w(t_j)(t_j - t_{j-1})\frac{1}{\bar{\sigma}_{t_j}}\frac{1}{n}\sum_{i=1}^{n}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$$

$$\bar{\mathcal{L}}(\theta) = \frac{1}{2}\sum_{j=1}^{N} w(t_j)(t_j - t_{j-1})\frac{1}{\bar{\sigma}_{t_j}}\mathbb{E}_{X_0}\mathbb{E}_{\xi}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{t_j}) + \xi\|^2$$

- Score matching $\to$ denoising score matching:

Theory: more about training

Properties of denoising score matching objective:

- $\bar{\mathcal{L}}_{em}(\theta) \to \bar{\mathcal{L}}$ as $n \to \infty$

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2}\sum_{j=1}^{N} w(t_j)(t_j - t_{j-1})\frac{1}{\bar{\sigma}_{t_j}}\frac{1}{n}\sum_{i=1}^{n}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$$

$$\bar{\mathcal{L}}(\theta) = \frac{1}{2}\sum_{j=1}^{N} w(t_j)(t_j - t_{j-1})\frac{1}{\bar{\sigma}_{t_j}}\mathbb{E}_{X_0}\mathbb{E}_{\xi}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{t_j}) + \xi\|^2$$

- Score matching $\to$ denoising score matching:

$$0 \le \mathbb{E}_{X_{t_j}}\|S(\theta; t_j, X_{t_j}) - \nabla_x \log p_{t_j}(X_{t_j})\|^2$$
$$= \frac{1}{\bar{\sigma}_{t_j}}\mathbb{E}_{X_0}\mathbb{E}_{\xi}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{t_j}) + \xi\|^2 + C_{t_j}$$

Theory: more about training

Properties of denoising score matching objective:

- $\bar{\mathcal{L}}_{em}(\theta) \to \bar{\mathcal{L}}$ as $n \to \infty$

$$\bar{\mathcal{L}}_{em}(\theta) = \frac{1}{2}\sum_{j=1}^{N} w(t_j)(t_j - t_{j-1})\frac{1}{\bar{\sigma}_{t_j}}\frac{1}{n}\sum_{i=1}^{n}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{ij}) + \xi_{ij}\|^2$$

$$\bar{\mathcal{L}}(\theta) = \frac{1}{2}\sum_{j=1}^{N} w(t_j)(t_j - t_{j-1})\frac{1}{\bar{\sigma}_{t_j}}\mathbb{E}_{X_0}\mathbb{E}_{\xi}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{t_j}) + \xi\|^2$$

- Score matching $\to$ denoising score matching:

$$\begin{aligned} 0 \leq & \, \mathbb{E}_{X_{t_j}}\|S(\theta; t_j, X_{t_j}) - \nabla_x \log p_{t_j}(X_{t_j})\|^2 \\ = & \frac{1}{\bar{\sigma}_{t_j}}\mathbb{E}_{X_0}\mathbb{E}_{\xi}\|\bar{\sigma}_{t_j}S(\theta; t_j, X_{t_j}) + \xi\|^2 + C_{t_j} \end{aligned}$$

  - $\Rightarrow \bar{\mathcal{L}}(\theta) + \bar{C} \geq 0$
  - $\Rightarrow \bar{\mathcal{L}}(\theta) \geq -\bar{C} > 0$, i.e., positive lower bound of $\bar{\mathcal{L}}(\theta)$

Theory: more training

- Theorem: choose arbitrarily large width m, $\epsilon_{\mathrm{train}}$ can be as small as possible

Theory: more training

- Theorem: choose arbitrarily large width m, ϵ_{train} can be as small as possible

$$0 \overset{?}{\leftarrow} \epsilon_{\text{train}} \approx \bar{\mathcal{L}}_{em}(\theta) \rightarrow \bar{\mathcal{L}}(\theta) \geq -\bar{C} > 0$$

Theory: more training

- Theorem: choose arbitrarily large width m, ϵ_{train} can be as small as possible

$$0 \overset{?}{\longleftarrow} \epsilon_{\text{train}} \approx \bar{\mathcal{L}}_{em}(\theta) \rightarrow \bar{\mathcal{L}}(\theta) \geq -\bar{C} > 0$$

- Contradiction?

Interesting generalization setting #1

No contradiction:

Interesting generalization setting #1

No contradiction:

$$\text{can have both } \bar{\mathcal{L}}_{em}(\theta) \approx 0 \text{ and } \bar{\mathcal{L}} \approx -\bar{C} > 0$$

Interesting generalization setting #1

No contradiction:

$$\text{can have both } \bar{\mathcal{L}}_{em}(\theta) \approx 0 \text{ and } \bar{\mathcal{L}} \approx -\bar{C} > 0$$

Reason:

Interesting generalization setting #1

No contradiction:

$$\text{can have both } \bar{\mathcal{L}}_{em}(\theta) \approx 0 \text{ and } \bar{\mathcal{L}} \approx -\bar{C} > 0$$

Reason:

- Overparameterized setting: fix m,

Interesting generalization setting #1

No contradiction:

$$\text{can have both } \bar{\mathcal{L}}_{em}(\theta) \approx 0 \text{ and } \bar{\mathcal{L}} \approx -\bar{C} > 0$$

Reason:

- Overparameterized setting: fix m,
  - $\bar{\mathcal{L}}_{em}(\theta) \leq \epsilon_{\mathrm{train}}$: sample size $n \ll m \;\Rightarrow\;$ n is bounded

Interesting generalization setting #1

No contradiction:

$$\text{can have both } \bar{\mathcal{L}}_{em}(\theta) \approx 0 \text{ and } \bar{\mathcal{L}} \approx -\bar{C} > 0$$

Reason:

- Overparameterized setting: fix m,
  - $\bar{\mathcal{L}}_{em}(\theta) \leq \epsilon_{\text{train}}$: sample size $n \ll m \;\Rightarrow\;$ n is bounded
  - $\bar{\mathcal{L}}_{em} \to \bar{\mathcal{L}}$ as $n \to \infty$

Interesting generalization setting #1

No contradiction:

$$\text{can have both } \bar{\mathcal{L}}_{em}(\theta) \approx 0 \text{ and } \bar{\mathcal{L}} \approx -\bar{C} > 0$$

Reason:

- Overparameterized setting: fix m,
  - $\bar{\mathcal{L}}_{em}(\theta) \leq \epsilon_{\mathrm{train}}$: sample size $n \ll m \;\Rightarrow\;$ n is bounded
  - $\bar{\mathcal{L}}_{em} \to \bar{\mathcal{L}}$ as $n \to \infty$

Consequence:

Interesting generalization setting #1

No contradiction:

$$\text{can have both } \bar{\mathcal{L}}_{em}(\theta) \approx 0 \text{ and } \bar{\mathcal{L}} \approx -\bar{C} > 0$$

Reason:

- Overparameterized setting: fix m,
  - $\bar{\mathcal{L}}_{em}(\theta) \leq \epsilon_{\text{train}}$: sample size $n \ll m \;\; \Rightarrow \;\;$ n is bounded
  - $\bar{\mathcal{L}}_{em} \to \bar{\mathcal{L}}$ as $n \to \infty$

Consequence:

- $\epsilon_n = |\bar{\mathcal{L}}_{em}(\theta) - \bar{\mathcal{L}}(\theta)|$

Interesting generalization setting #1

No contradiction:

$$\text{can have both } \bar{\mathcal{L}}_{em}(\theta) \approx 0 \text{ and } \bar{\mathcal{L}} \approx -\bar{C} > 0$$

Reason:

- Overparameterized setting: fix m,
  - $\bar{\mathcal{L}}_{em}(\theta) \leq \epsilon_{\text{train}}$: sample size $n \ll m \;\Rightarrow\;$ n is bounded
  - $\bar{\mathcal{L}}_{em} \to \bar{\mathcal{L}}$ as $n \to \infty$

Consequence:

- $\epsilon_n = |\bar{\mathcal{L}}_{em}(\theta) - \bar{\mathcal{L}}(\theta)|$
- If ϵ_{train} is small $\Rightarrow$ ϵ_n is large

Interesting generalization setting

- Recall:

$$\mathrm{KL}(p_\delta | q_{T-\delta}) \lesssim \underbrace{E_D + E_I}_{\text{sampling}} + E_S$$

where p_δ is the true density at time δ, and $q_{T-\delta}$ is the approximated density of p_δ.

Interesting generalization setting

- Recall:

$$\mathrm{KL}(p_\delta|q_{T-\delta}) \lesssim \underbrace{E_D + E_I}_{\text{sampling}} + E_S$$

  where p_δ is the true density at time δ, and $q_{T-\delta}$ is the approximated density of p_δ.
- E_S: score error

Interesting generalization setting #1

$$E_S \leq \max_{1 \leq j \leq N} \frac{\sigma^2_{t_{N-j}}}{w(t_{N-j})} \Big(\epsilon_{\text{train}} + \underbrace{\epsilon_n}_{\bar{\mathcal{L}}_{em}(\theta) - \bar{\mathcal{L}}(\theta)} + \underbrace{\epsilon_{\text{est}}}_{\substack{\text{estimation} \\ \text{error}}} + \underbrace{\epsilon_{\text{approx}}}_{\substack{\text{approximation} \\ \text{error}}} \Big)$$

Interesting generalization setting #1

$$E_S \leq \max_{1\leq j\leq N} \frac{\sigma^2_{t_{N-j}}}{w(t_{N-j})}\Big(\epsilon_{\text{train}} + \underbrace{\epsilon_n}_{\bar{\mathcal{L}}_{em}(\theta)-\bar{\mathcal{L}}(\theta)} + \underbrace{\epsilon_{\text{est}}}_{\substack{\text{estimation}\\ \text{error}}} + \underbrace{\epsilon_{\text{approx}}}_{\substack{\text{approximation}\\ \text{error}}}\Big)$$

- $\epsilon_n + \epsilon_{\text{est}} + \epsilon_{\text{approx}}$
  [Chen et al., 2023, Oko et al., 2023, Han et al., 2024]

Interesting generalization setting #1

$$E_S \leq \max_{1\leq j\leq N} \frac{\sigma^2_{t_{N-j}}}{w(t_{N-j})}\Big(\epsilon_{\text{train}} + \underbrace{\epsilon_n}_{\bar{\mathcal{L}}_{em}(\theta)-\bar{\mathcal{L}}(\theta)} + \underbrace{\epsilon_{\text{est}}}_{\substack{\text{estimation}\\ \text{error}}} + \underbrace{\epsilon_{\text{approx}}}_{\substack{\text{approximation}\\ \text{error}}}\Big)$$

- $\epsilon_n + \epsilon_{\text{est}} + \epsilon_{\text{approx}}$
  [Chen et al., 2023, Oko et al., 2023, Han et al., 2024]
- Regression generalization → Diffusion models

Interesting generalization setting #1

$$\color{red}{E_S} \leq \max_{1\leq j\leq N} \frac{\sigma^2_{t_{N-j}}}{w(t_{N-j})}\left(\epsilon_{\text{train}} + \underbrace{\epsilon_n}_{\bar{\mathcal{L}}_{em}(\theta)-\bar{\mathcal{L}}(\theta)} + \underbrace{\epsilon_{\text{est}}}_{\substack{\text{estimation}\\ \text{error}}} + \underbrace{\epsilon_{\text{approx}}}_{\substack{\text{approximation}\\ \text{error}}}\right)$$

- $\epsilon_n + \epsilon_{\text{est}} + \epsilon_{\text{approx}}$
  [Chen et al., 2023, Oko et al., 2023, Han et al., 2024]
- Regression generalization $\to$ Diffusion models
- Open problem:
  - $\epsilon_{\text{train}} + \epsilon_n \geq -2\bar{C} > 0 \;\Rightarrow$ error bound $\not\to 0$

Interesting generalization setting #1

$$E_S \leq \max_{1\leq j\leq N} \frac{\sigma^2_{t_{N-j}}}{w(t_{N-j})}\Big(\epsilon_{\mathrm{train}} + \underbrace{\epsilon_n}_{\bar{\mathcal{L}}_{em}(\theta)-\bar{\mathcal{L}}(\theta)} + \underbrace{\epsilon_{\mathrm{est}}}_{\text{estimation error}} + \underbrace{\epsilon_{\mathrm{approx}}}_{\text{approximation error}}\Big)$$

- $\epsilon_n + \epsilon_{\mathrm{est}} + \epsilon_{\mathrm{approx}}$
  [Chen et al., 2023, Oko et al., 2023, Han et al., 2024]
- Regression generalization $\rightarrow$ Diffusion models
- Open problem:
  - $\epsilon_{\mathrm{train}} + \epsilon_n \geq -2\bar{C} > 0 \;\Rightarrow$ error bound $\not\to 0$
    change decomposition $\Rightarrow$ tighter analysis?

Interesting generalization setting #1

$$\color{red}{E_S} \leq \max_{1\leq j\leq N} \frac{\sigma^2_{t_{N-j}}}{w(t_{N-j})}\Big(\epsilon_{\text{train}} + \underbrace{\epsilon_n}_{\bar{\mathcal{L}}_{em}(\theta)-\bar{\mathcal{L}}(\theta)} + \underbrace{\epsilon_{\text{est}}}_{\substack{\text{estimation}\\ \text{error}}} + \underbrace{\epsilon_{\text{approx}}}_{\substack{\text{approximation}\\ \text{error}}}\Big)$$

- $\epsilon_n + \epsilon_{\text{est}} + \epsilon_{\text{approx}}$
  [Chen et al., 2023, Oko et al., 2023, Han et al., 2024]
- Regression generalization $\rightarrow$ Diffusion models
- Open problem:
  - $\epsilon_{\text{train}} + \epsilon_n \geq -2\bar{C} > 0 \;\Rightarrow$ error bound $\not\rightarrow 0$
    change decomposition $\Rightarrow$ tighter analysis?
  - $S(\theta; t, X_t) \overset{?}{\rightarrow} \nabla_x \log p_t(X_t)$

Example full error analysis

Theorem (EDM polynomial schedule [Karras et al., 2022])

$$KL(p_\delta | q_{T-\delta}) \lesssim \underbrace{\frac{\mathrm{m}_2^2}{T^2}}_{E_I} + \underbrace{\frac{da^2 T^{\frac{1}{a}}}{\delta^{\frac{1}{a}} N} + (\mathrm{m}_2^2 + d)\left(\frac{a^2 T^{\frac{1}{a}}}{\delta^{\frac{1}{a}} N} + \frac{a^3 T^{\frac{2}{a}}}{\delta^{\frac{2}{a}} N^2}\right)}_{E_D}$$
$$+ \underbrace{\frac{1}{N}\left(C_2 + \left(1 - C_1 h\left(\frac{md^{\frac{a_0 - 1}{2}}}{n^3 N^2}\right)\right)^K\right)}_{E_S},$$

where $\delta = t_0$, $a = 7$, $a_0 \in (1/2, 1)$.

- $C_2 = \epsilon_n + \epsilon_{\text{est}} + \epsilon_{\text{approx}}$

Theory: full error analysis

$$E_S := \sum_{j=0}^{N-1} \alpha_j(\underbrace{t_j}_{\substack{\text{time}\\ \text{schedule}}}, \underbrace{\bar{\sigma}_{t_j}}_{\substack{\text{variance}\\ \text{schedule}}}) \, \mathbb{E}_{X_{t_j}} \| S(\theta; t_j, X_{t_j}) - \nabla \log p_{t_j}(X_{t_j}) \|^2$$

sampling + optimization

Theory: full error analysis

$$E_S := \sum_{j=0}^{N-1} \alpha_j(\underbrace{t_j}_{\text{time schedule}}, \underbrace{\bar{\sigma}_{t_j}}_{\text{variance schedule}})\, \mathbb{E}_{X_{t_j}} \|S(\theta; t_j, X_{t_j}) - \nabla \log p_{t_j}(X_{t_j})\|^2$$

sampling + optimization

- $\alpha_j(t_j, \bar{\sigma}_{t_j}) \neq \beta_j$, the weighting for training objective

Theory: full error analysis

$$E_S := \sum_{j=0}^{N-1} \alpha_j(\underbrace{t_j}_{\text{time schedule}}, \underbrace{\bar{\sigma}_{t_j}}_{\text{variance schedule}}) \, \mathbb{E}_{X_{t_j}} \| S(\theta; t_j, X_{t_j}) - \nabla \log p_{t_j}(X_{t_j}) \|^2$$

sampling + optimization

- $\alpha_j(t_j, \bar{\sigma}_{t_j}) \neq \beta_j$, the weighting for training objective
- First choose total weighting β_j; then apply the schedules $t_j, \bar{\sigma}_{t_j}$

Theory: full error analysis

$$E_S := \sum_{j=0}^{N-1} \alpha_j(\underbrace{t_j}_{\text{time schedule}}, \underbrace{\bar{\sigma}_{t_j}}_{\text{variance schedule}}) \, \mathbb{E}_{X_{t_j}} \| S(\theta; t_j, X_{t_j}) - \nabla \log p_{t_j}(X_{t_j}) \|^2$$

sampling + optimization

- $\alpha_j(t_j, \bar{\sigma}_{t_j}) \neq \beta_j$, the weighting for training objective
- First choose total weighting β_j; then apply the schedules $t_j, \bar{\sigma}_{t_j}$
- Next: focus on two concrete schedules used in practice

Theory: full error analysis

$$E_S := \sum_{j=0}^{N-1} \alpha_j(\underbrace{t_j}_{\text{time schedule}}, \underbrace{\bar{\sigma}_{t_j}}_{\text{variance schedule}}) \, \mathbb{E}_{X_{t_j}} \|S(\theta; t_j, X_{t_j}) - \nabla \log p_{t_j}(X_{t_j})\|^2$$

sampling + optimization

- $\alpha_j(t_j, \bar{\sigma}_{t_j}) \neq \beta_j$, the weighting for training objective
- First choose total weighting β_j; then apply the schedules $t_j, \bar{\sigma}_{t_j}$
- Next: focus on two concrete schedules used in practice
  - Theoretical implication: how to choose between two schedules

Theory: full error analysis

Two most famous choices:

Theory: full error analysis

Two most famous choices:

- Exponential schedules [Song et al., 2021]: first work

Theory: full error analysis

Two most famous choices:

- Exponential schedules [Song et al., 2021]: first work
- Polynomial schedules [Karras et al., 2022]: improved design

Theory: full error analysis

Two most famous choices:

- Exponential schedules [Song et al., 2021]: first work
- Polynomial schedules [Karras et al., 2022]: improved design

	Variance $\bar{\sigma}_t$	Time t_k
[Karras et al., 2022]	t	$\left(\bar{\sigma}_{\max}^{1/\rho} - (\bar{\sigma}_{\max}^{1/\rho} - \bar{\sigma}_{\min}^{1/\rho})\frac{N-k}{N}\right)^{\rho}$
[Song et al., 2021]	$\sqrt{t}$	$\bar{\sigma}_{\max}^2 \left(\frac{\bar{\sigma}_{\min}^2}{\bar{\sigma}_{\max}^2}\right)^{\frac{N-k}{N}}$

Theory: full error analysis

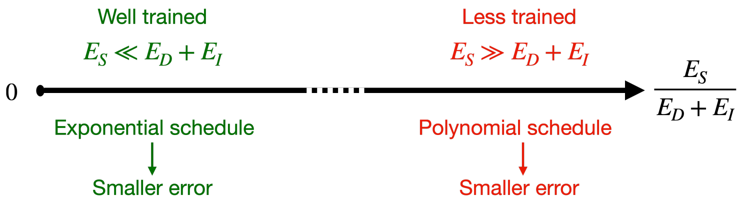
Two most famous choices:

- Exponential schedules [Song et al., 2021]: first work
- Polynomial schedules [Karras et al., 2022]: improved design

	Variance $\bar{\sigma}_t$	Time t_k
[Karras et al., 2022]	t	$\left(\bar{\sigma}_{\max}^{1/\rho} - (\bar{\sigma}_{\max}^{1/\rho} - \bar{\sigma}_{\min}^{1/\rho})\frac{N-k}{N}\right)^{\rho}$
[Song et al., 2021]	$\sqrt{t}$	$\bar{\sigma}_{\max}^2 \left(\frac{\bar{\sigma}_{\min}^2}{\bar{\sigma}_{\max}^2}\right)^{\frac{N-k}{N}}$

- Time schedule t_k: a function of k
- $\rho = 7$

Implication from theory: How to choose schedules?



- 1st quantitative result that analyzes forward training + backward sampling

$$\boxed{dX = \sqrt{2}dW_t}$$

$\mathcal{D}$ easy distribution

$$\boxed{dY = 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

- 1st quantitative result that analyzes
  forward training + backward sampling

- nontrivial analysis of training dynamics
  - overparameterized ReLU MLP

$$\boxed{dX = \sqrt{2}dW_t}$$

$\mathcal{D}$ easy distribution

$$\boxed{dY = 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

$$\boxed{dX = \sqrt{2}dW_t}$$

$\mathcal{D}$ easy distribution

$$\boxed{dY = 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

- 1st quantitative result that analyzes
  forward training + backward sampling
- nontrivial analysis of training dynamics
  - overparameterized ReLU MLP
  - weaker data assumpts (no separability, unbdd output, …)

$$\boxed{dX = \sqrt{2}dW_t}$$

$\mathcal{D}$ easy distribution

$$\boxed{dY = 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

- 1st quantitative result that analyzes
  forward training + backward sampling

- nontrivial analysis of training dynamics
  - overparameterized ReLU MLP
  - weaker data assumpts (no separability, unbdd output, …)
  - interesting generalization setting #1: SM - DSM gap

$$\boxed{dX = \sqrt{2}dW_t}$$

$\mathcal{D}$ easy distribution

$$\boxed{dY = 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

- 1st quantitative result that analyzes
  forward training + backward sampling

- nontrivial analysis of training dynamics
  - overparameterized ReLU MLP
  - weaker data assumpts (no separability, unbdd output, …)
  - interesting generalization setting #1: SM - DSM gap

- besides understanding, practical implication

$$\boxed{dX = \sqrt{2}dW_t}$$

$\mathcal{D}$ easy distribution

$$\boxed{dY = 2s(Y, T - t)dt + \sqrt{2}dB_t}$$

- 1st quantitative result that analyzes
  forward training + backward sampling

- nontrivial analysis of training dynamics
  - overparameterized ReLU MLP
  - weaker data assumpts (no separability, unbdd output, …)
  - interesting generalization setting #1: SM - DSM gap

- besides understanding, practical implication
  - total weighting

3.1 Summary

$$\boxed{dX = \sqrt{2}dW_t}$$

$\mathcal{D}$ easy distribution

$$\boxed{dY = 2s(Y, T-t)dt + \sqrt{2}dB_t}$$

- 1st quantitative result that analyzes
  forward training + backward sampling

- nontrivial analysis of training dynamics
  - overparameterized ReLU MLP
  - weaker data assumpts (no separability, unbdd output, …)
  - interesting generalization setting #1: SM - DSM gap

- besides understanding, practical implication
  - total weighting
  - variance and time schedules

Generalization Setting #2 (open)

Generalization Setting #2 (open)

Quantifications: $d(p_{\text{training data}}|p_{\text{generated data}}) \leq \cdots$

What if: generated data = uniformly drawn from training data ?

Generalization Setting #2 (open)

Quantifications: $d(p_{\text{training data}} | p_{\text{generated data}}) \leq \cdots$

What if: generated data = uniformly drawn from training data ?

accurate, but not innovative

Generalization Setting #2 (open)

Quantifications: $d(\boxed{p_{\text{training data}}} | p_{\text{generated data}}) \le \cdots$

What if: generated data = uniformly drawn from training data ?

accurate, but not innovative

Key: what exactly is this?

Generalization Setting #2 (open)

Quantifications: $d(\boxed{p_{\text{training data}}} | p_{\text{generated data}}) \leq \cdots$

What if: generated data = uniformly drawn from training data ?

accurate, but not innovative

Key: what exactly is this?

$$\min_\theta \int_0^T w(t) \underbrace{\mathbb{E}_{X_t} \|\nabla_x \log p(X_t, t) - s_\theta(X_t, t)\|^2}_{\approx \frac{1}{n}\sum_{i=1}^{n} \|\nabla_x \log p(X_t^i, t) - s_\theta(X_t^i, t)\|^2} dt$$

Generalization Setting #2 (open)

Quantifications: $d(\boxed{p_{\text{training data}}}|p_{\text{generated data}}) \leq \cdots$

What if: generated data = uniformly drawn from training data ?

accurate, but not innovative

Key: what exactly is this?

$$\min_\theta \int_0^T w(t) \underbrace{\mathbb{E}_{X_t} \|\nabla_x \log p(X_t, t) - s_\theta(X_t, t)\|^2}_{\approx \frac{1}{n}\sum_{i=1}^{n} \|\nabla_x \log p(X_t^i, t) - s_\theta(X_t^i, t)\|^2} dt$$

natural: empirical distribution, i.e. sum of Dirac's

Generalization Setting #2 (open)

Quantifications: $d(\boxed{p_{\text{training data}}} | p_{\text{generated data}}) \leq \cdots$

What if: generated data = uniformly drawn from training data ?

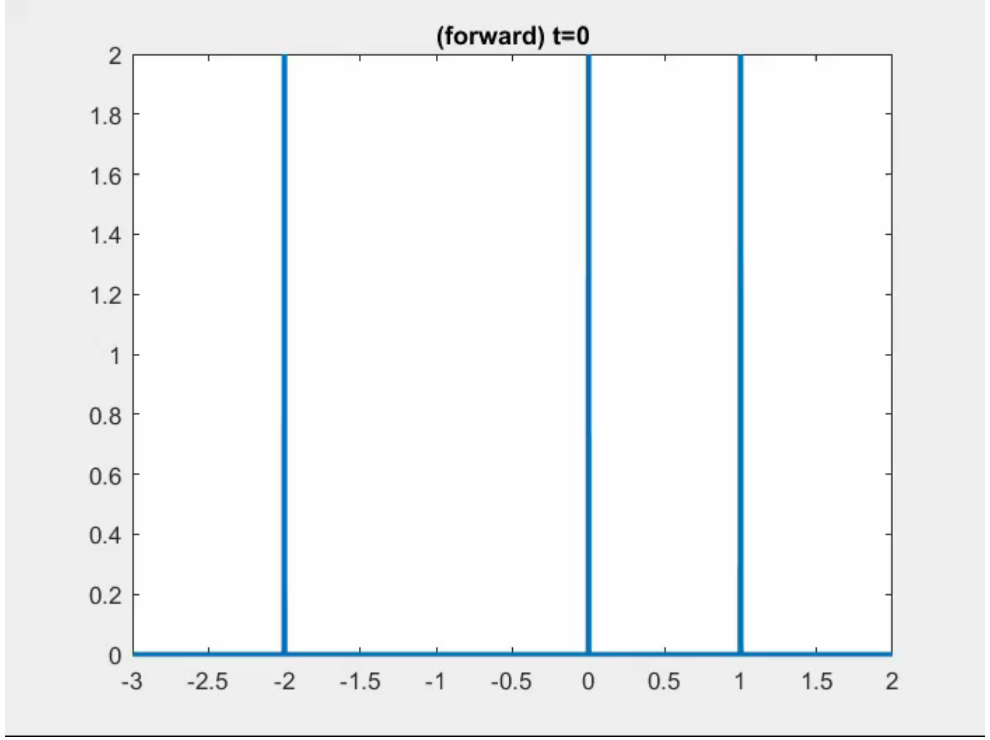
accurate, but not innovative

exact density evolution

Key: what exactly is this?

$$\min_\theta \int_0^T w(t) \underbrace{\mathbb{E}_{X_t} \|\nabla_x \log p(X_t, t) - s_\theta(X_t, t)\|^2}_{} dt$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \|\nabla_x \log p(X_t^i, t) - s_\theta(X_t^i, t)\|^2$$



natural: empirical distribution, i.e. sum of Dirac's

Generalization Setting #2 (open)

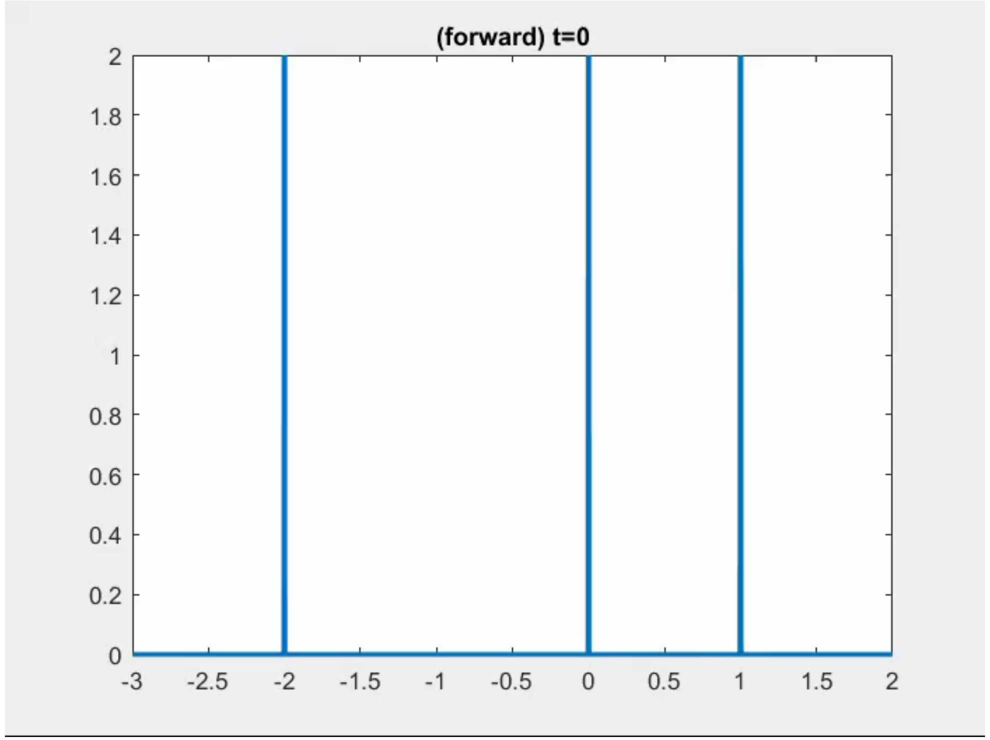
Quantifications: $d(\boxed{p_{\text{training data}}}|p_{\text{generated data}}) \leq \cdots$

perfect score

What if: generated data = uniformly drawn from training data ?

accurate, but not innovative

exact density evolution



Key: what exactly is this?

$$\min_\theta \int_0^T w(t) \underbrace{\mathbb{E}_{X_t} \|\nabla_x \log p(X_t, t) - s_\theta(X_t, t)\|^2}_{\approx \frac{1}{n}\sum_{i=1}^n \|\nabla_x \log p(X_t^i, t) - s_\theta(X_t^i, t)\|^2} dt$$

natural: empirical distribution, i.e. sum of Dirac's

Generalization Setting #2 (open)

Quantifications: $d(\boxed{p_{\text{training data}}}|p_{\text{generated data}}) \leq \cdots$

perfect score

What if: generated data = uniformly drawn from training data ?

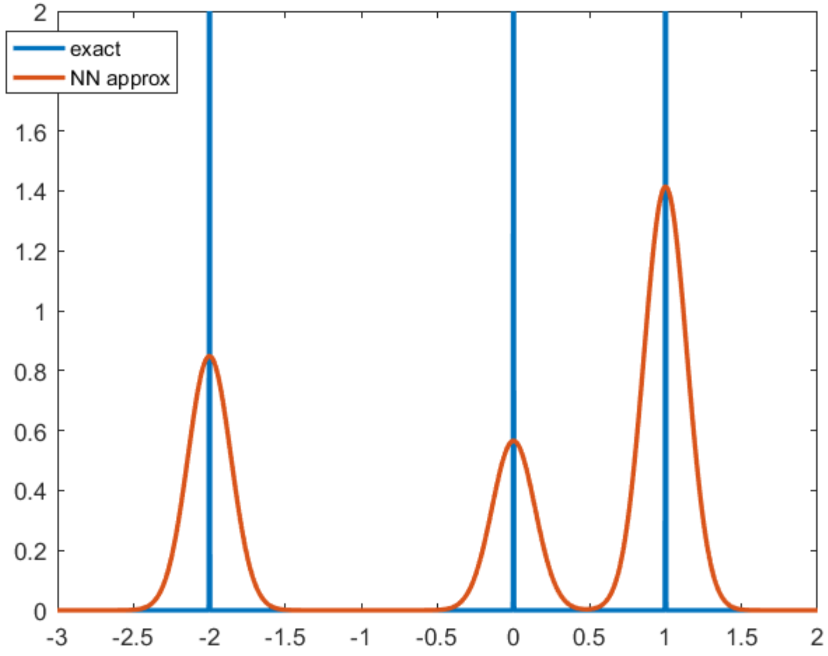
accurate, but not innovative

approximation error

Key: what exactly is this?

$$\min_\theta \int_0^T \underbrace{w(t)\mathbb{E}_{X_t}\|\nabla_x \log p(X_t,t) - s_\theta(X_t,t)\|^2}dt$$

$$\approx \frac{1}{n}\sum_{i=1}^n \|\nabla_x \log p(X_t^i,t) - s_\theta(X_t^i,t)\|^2$$



natural: empirical distribution, i.e. sum of Dirac's

Generalization Setting #2 (open)

Quantifications: $d(\boxed{p_{\text{training data}}}|p_{\text{generated data}}) \leq \cdots$

perfect score

What if: generated data = uniformly drawn from training data ?

accurate, but not innovative

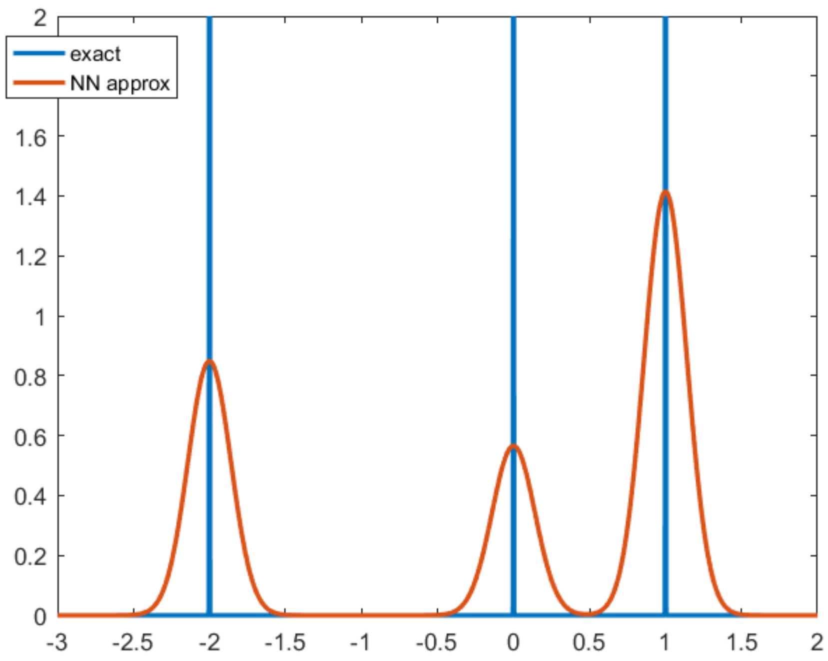
- score error
- integration error
- initialization error
- early stopping

4 types of errors

Key: what exactly is this?

$$\min_\theta \int_0^T w(t)\underbrace{\mathbb{E}_{X_t}\|\nabla_x \log p(X_t,t) - s_\theta(X_t,t)\|^2}_{} dt$$

$$\approx \frac{1}{n}\sum_{i=1}^n \|\nabla_x \log p(X_t^i,t) - s_\theta(X_t^i,t)\|^2$$



natural: empirical distribution, i.e. sum of Dirac's

Generalization Setting #2 (open)

Quantifications: $d(\boxed{p_{\text{training data}}} | p_{\text{generated data}}) \leq \cdots$

perfect score

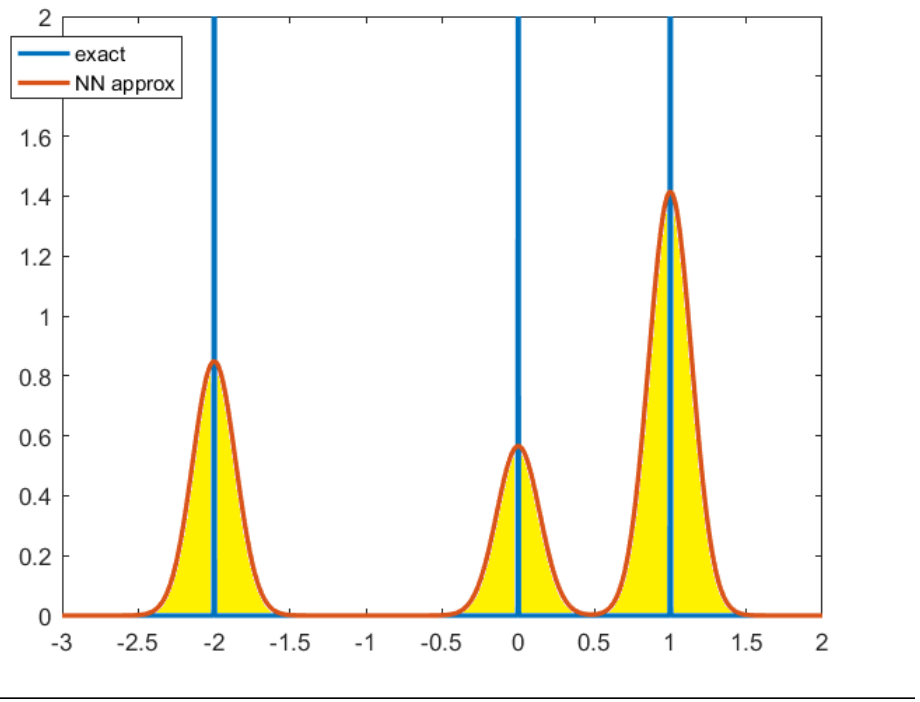
What if: generated data = uniformly drawn from training data ?

accurate, but not innovative

4 types of errors → innovation?

Key: what exactly is this?

$$\min_\theta \int_0^T w(t) \underbrace{\mathbb{E}_{X_t} \|\nabla_x \log p(X_t, t) - s_\theta(X_t, t)\|^2}_{\approx \frac{1}{n} \sum_{i=1}^n \|\nabla_x \log p(X_t^i, t) - s_\theta(X_t^i, t)\|^2} dt$$



natural: empirical distribution, i.e. sum of Dirac's

Generalization Setting #2 (open)

Quantifications: $d(\boxed{p_{\text{training data}}} | p_{\text{generated data}}) \leq \cdots$

perfect score

What if: generated data = uniformly drawn from training data ?

Kadkhodaie+ 23, *Scarvelis*+ 23, …

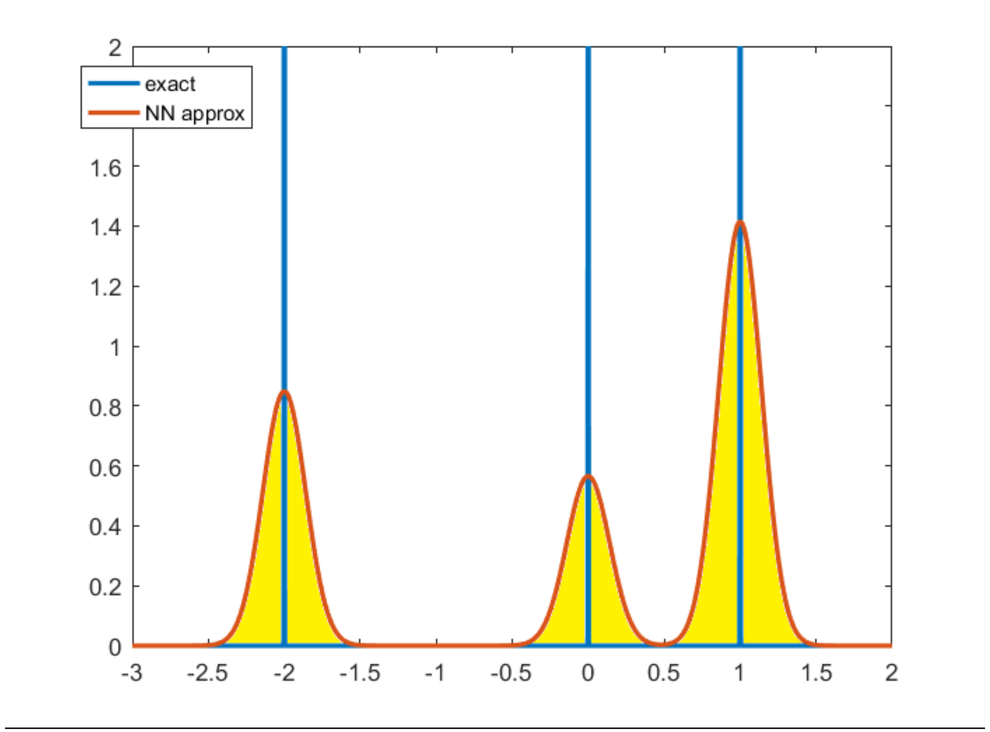
accurate, but not innovative

4 types of errors → innovation?

Key: what exactly is this?

$$\min_\theta \int_0^T w(t) \underbrace{\mathbb{E}_{X_t} \|\nabla_x \log p(X_t, t) - s_\theta(X_t, t)\|^2}_{} dt$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \|\nabla_x \log p(X_t^i, t) - s_\theta(X_t^i, t)\|^2$$



natural: empirical distribution, i.e. sum of Dirac's

Generalization Setting #2 (open)

Quantifications: $d(\boxed{p_{\text{training data}}} | p_{\text{generated data}}) \leq \cdots$

perfect score

What if: generated data = uniformly drawn from training data ?

Kadkhodaie+ 23, *Scarvelis*+ 23, …

accurate, but not innovative

4 types of errors → innovation?

hope: understanding the training dynamics

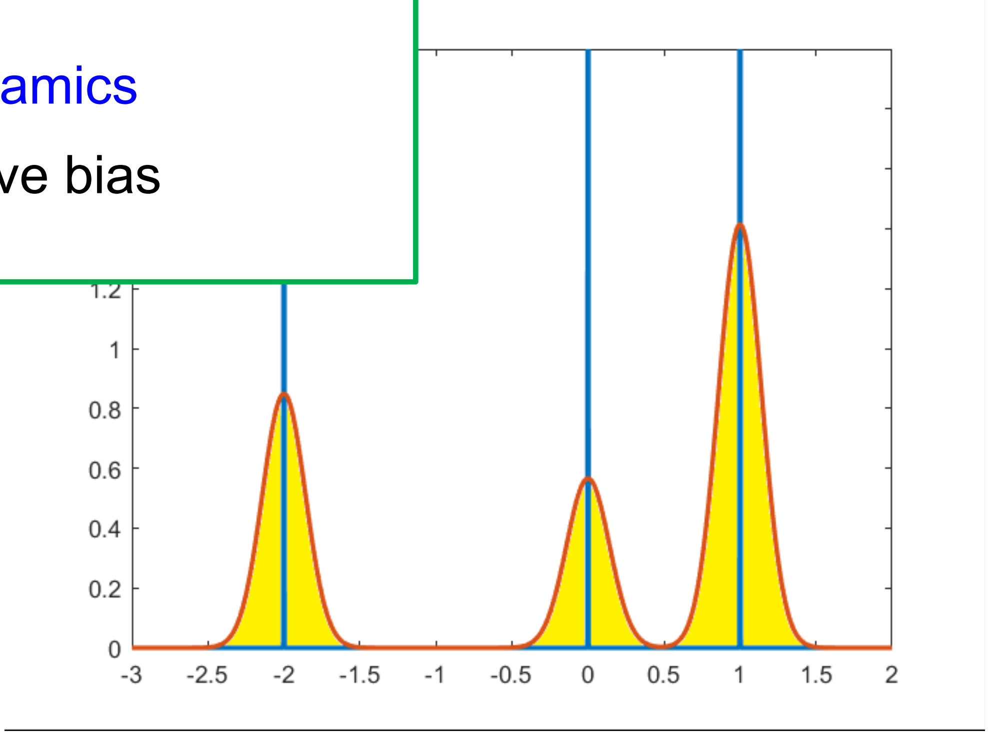
& tight analysis of sampling → inductive bias

Key:

$$\min_\theta \int_0^T \underbrace{w(t)\mathbb{E}_{X_t}\|\nabla_x \log p(X_t,t) - s_\theta(X_t,t)\|^2 dt}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} \|\nabla_x \log p(X_t^i,t) - s_\theta(X_t^i,t)\|^2$$

natural: empirical distribution, i.e. sum of Dirac's



Can discrete diffusion model add to the success of LLM?

Can discrete diffusion model add to the success of LLM?

Quantitative error analysis (possible)

Can discrete diffusion model add to the success of LLM?

Quantitative error analysis (possible) →

what is it good at?

Can discrete diffusion model add to the success of LLM?

Quantitative error analysis (possible) →

what is it good at?

difference and similarity to autoregressive model

Can discrete diffusion model add to the success of LLM?

Quantitative error analysis (possible) →

what is it good at?

difference and similarity to autoregressive model

how to best deploy it?

…

Thank you for your attention and feedback!

Support:

NSF DMS-1847802, ECCS-1936776

Cullen-Peck Scholarship

Emory-GT AI.Humanity Award

Simons Institute Research Fellowship

MoleiTaoMath

SCAN ME

itsdynamical

SCAN ME