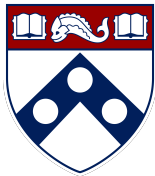


Low-dimensional adaptation of diffusion generative models



Jiadong Liang

Wharton Statistics & Data Science



Zhihan Huang
UPenn



Yuxin Chen
UPenn

“Low-dimensional adaptation of diffusion models: convergence in total variation,”

J. Liang, Z. Huang, Y. Chen, arXiv:2501.12982, COLT 2025

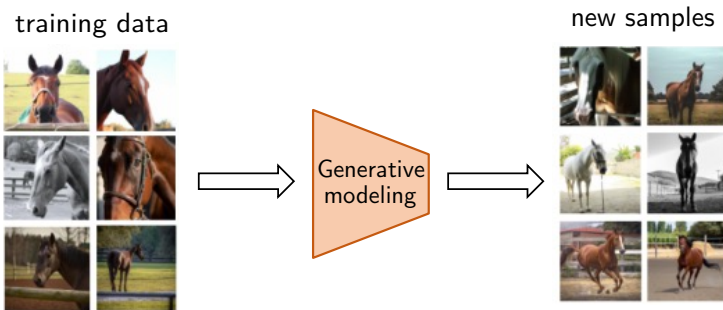
Generative modeling

training data



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

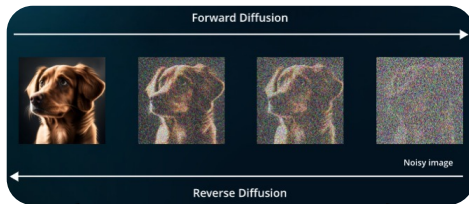
Generative modeling



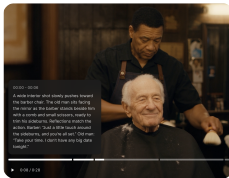
- Given training data $\underbrace{X^{\text{train},i} \sim p_{\text{data}}}_{\text{from a general distribution}} (1 \leq i \leq N)$ in $\mathbb{R}^d$
- Generate **new** samples $Y \sim p_{\text{data}}$

Inspired by nonequilibrium thermodynamics
— Sohl-Dickstein, Weiss, Maheswaranathan, Ganguli '15

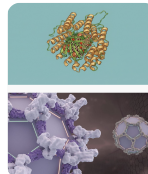
Diffusion models



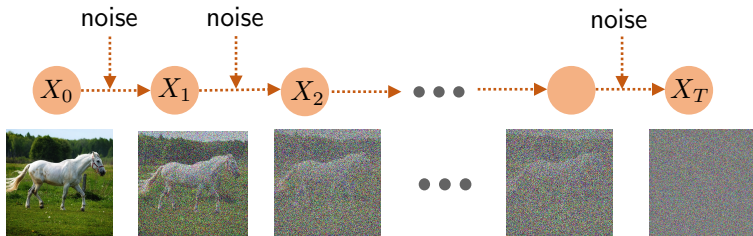
GPT-image-2



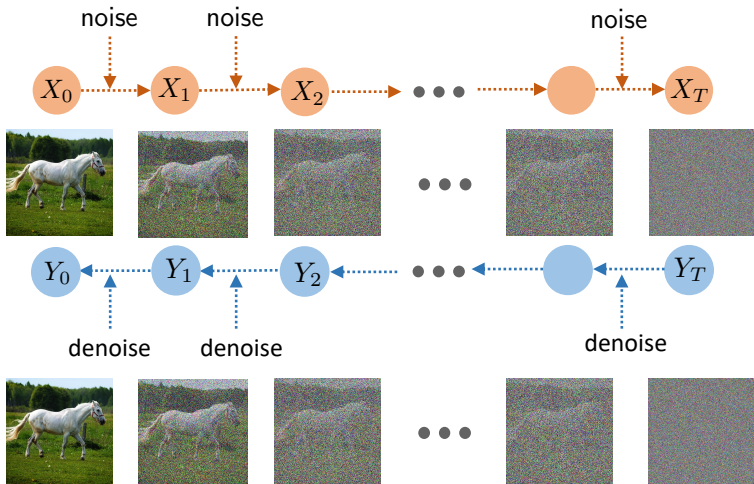
seedance-2.5



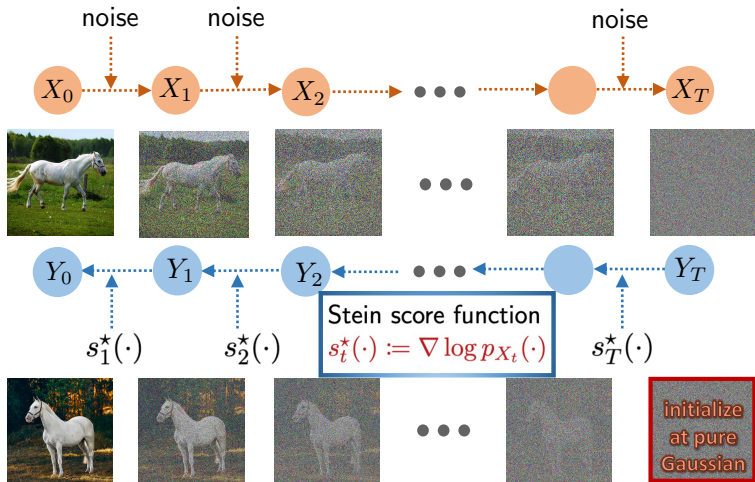
RFdiffusion



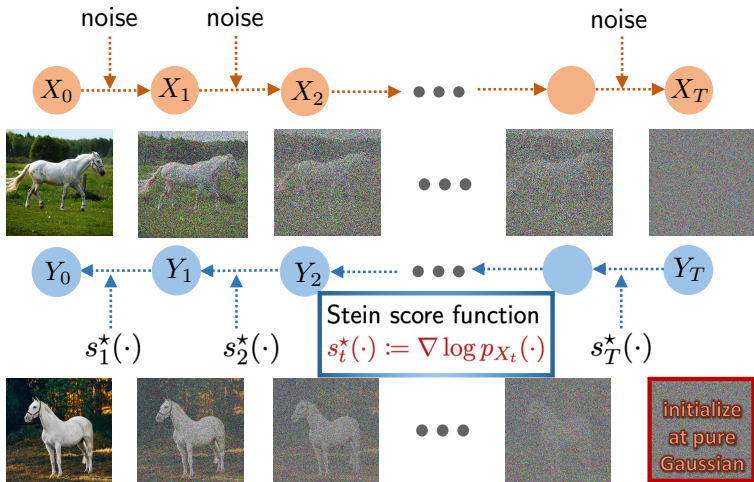
- **forward process:** diffuse data into noise



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

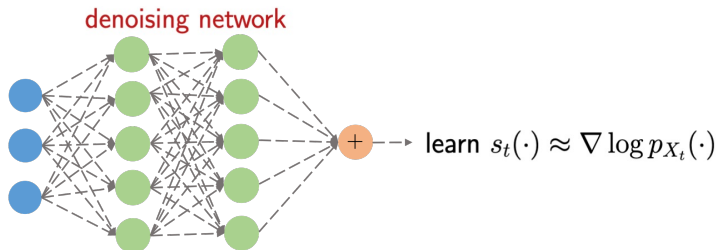


- **forward process:** diffuse data into noise
- **reverse process:** convert pure noise into data-like distributions

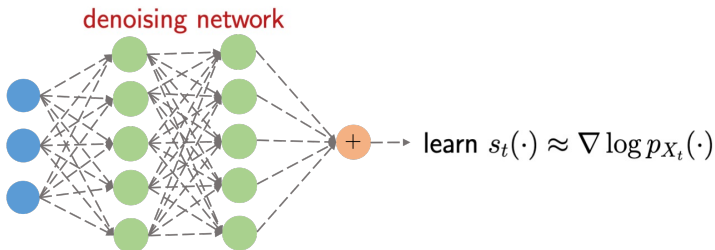


- **forward process:** diffuse data into noise
- **reverse process:** convert pure noise into data-like distributions

Goal: $Y_t \stackrel{d}{\approx} X_t, \quad t = T, \dots, 1$

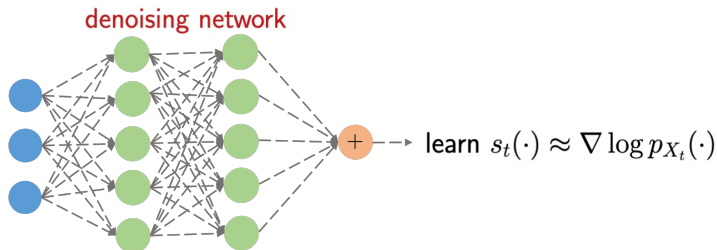


1. **score learning/matching:** learn estimates $s_t(\cdot)$ for $\nabla \log p_{X_t}(\cdot)$

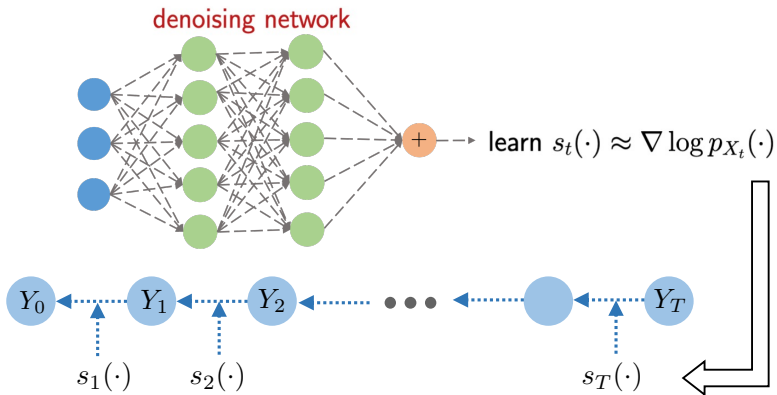


$$\epsilon_{\text{score}}^2 := \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\|s_t(X_t) - s_t^*(X_t)\|_2^2 \right]$$

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



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



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

Mainstream diffusion models

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

DENOISING DIFFUSION IMPLICIT MODELS

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

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

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

$\tilde{O}(d/\varepsilon)$ iterations.

- d : ambient dimension

— Lee, Lu, Tan '22

— Chen, Lee, Lu '22

— Benton, De Bortoli, Doucet, Deligiannidis '23

Mainstream diffusion models

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

DENOISING DIFFUSION IMPLICIT MODELS

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

Theorem 2 (Dou, Kotekal, Xu, Zhou '24, informal)

For a density on $\mathbb{R}^d$, score matching accuracy ε is achievable with

$$n = \tilde{O}\left(\varepsilon^{-\Theta(d)}\right) \quad \text{training samples.}$$

- d : ambient dimension

— Wibisono, Wu, Yang '24
— Oko, Akiyama, Suzuki '23

Ambient dimension is the common bottleneck



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

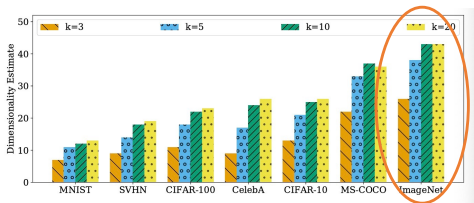
Ambient dimension is the common bottleneck



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

In practice, diffusion models learn from only 10^4 – 10^6 training samples, and DDIM/DDPM yield good samples in only 10^1 – 10^2 iterations . . .

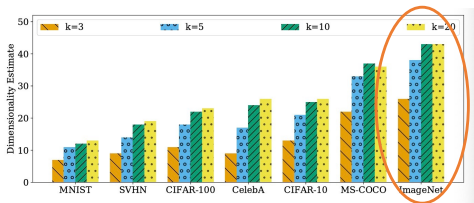
Ambient dimension is the common bottleneck



ImageNet: $d = 150,528$ pixels per image
 $k = 43$ intrinsic dimension (Pope et al. '21)

In practice, diffusion models learn from only 10^4 – 10^6 training samples, and DDIM/DDPM yield good samples in only 10^1 – 10^2 iterations ...

Ambient dimension is the common bottleneck



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

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

In practice, diffusion models learn from only 10^4 – 10^6 training samples, and DDIM/DDPM yield good samples in only 10^1 – 10^2 iterations ...

Intrinsic dimension

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

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

Intrinsic dimension

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

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

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

Our assumptions

- **Low dimensional structure:** p_{data} has intrinsic dimension $k(\ll d)$.
- **Bounded data assumptions:**

$$\mathbb{P}(\|X_0\|_2 \leq R_0) = 1.$$

Our assumptions

- **Low dimensional structure:** p_{data} has intrinsic dimension $k(\ll d)$.
- **Bounded data assumptions:**

$$\mathbb{P}(\|X_0\|_2 \leq R_0) = 1.$$

No smoothness or log-concavity assumptions on p_{data} .

Can diffusion models adapt to intrinsic low dimensionality?

Can diffusion models adapt to intrinsic low dimensionality?

- Sampling efficiency: number of reverse iterations T

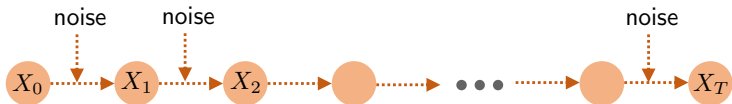
Can diffusion models adapt to intrinsic low dimensionality?

- Sampling efficiency: number of reverse iterations T
- Training sample size: number of training samples n

Can diffusion models adapt to intrinsic low dimensionality?

- Sampling efficiency: number of reverse iterations T
- Training sample size: number of training samples n

Forward process



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

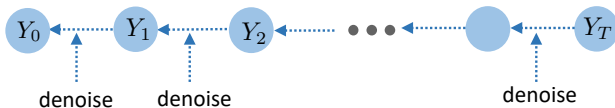
Forward process



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

- $1 - \alpha_t$: **forward step size**; its schedule can be chosen.

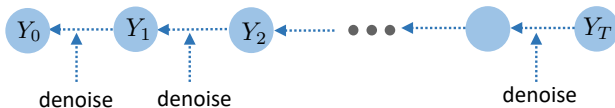
DDPM vs. DDIM



— Ho, Jain, Abbeel '20

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

DDPM vs. DDIM



— Ho, Jain, Abbeel '20

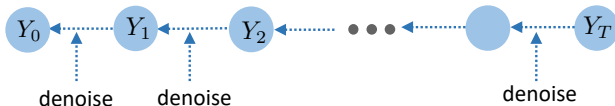
1. A stochastic sampler: **denoising diffusion probabilistic models**

DDPM

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

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

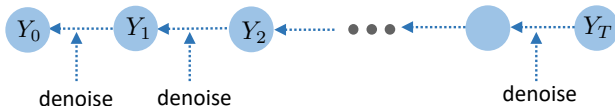
DDPM vs. DDIM



— Song, Meng, Ermon '20

2. A deterministic sampler: **denoising diffusion implicit models**
DDIM

DDPM vs. DDIM



— Song, Meng, Ermon '20

2. A deterministic sampler: **denoising diffusion implicit models**
DDIM

$$Y_T \sim \mathcal{N}(0, I_d)$$
$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(Y_t + \underbrace{\eta_t^{\text{ddim}} s_t(Y_t)}_{\text{deterministic}} \right), \quad t = T, \dots, 1$$

Interpretations from lens of SDE/ODE

forward process

$$\begin{aligned} X_t &= \sqrt{1 - \beta_t} X_{t-1} + \sqrt{\beta_t} \mathcal{N}(0, I_d) \\ \Rightarrow dX_t &= -\frac{1}{2} \beta(t) X_t dt + \sqrt{\beta(t)} dW_t \quad (\text{SDE}) \end{aligned}$$

Interpretations from lens of SDE/ODE

forward process

$$\begin{aligned} X_t &= \sqrt{1 - \beta_t} X_{t-1} + \sqrt{\beta_t} \mathcal{N}(0, I_d) \\ \Rightarrow dX_t &= -\frac{1}{2} \beta(t) X_t dt + \sqrt{\beta(t)} dW_t \quad (\text{SDE}) \end{aligned}$$

- $\exists$ reverse-time SDE w/ same *path* distribution as X_{T-t}

time discretization
 $\longrightarrow$ DDPM

- $\exists$ probability-flow ODE w/ same *marginal* dist. as X_{T-t}

time discretization
 $\longrightarrow$ DDIM

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

- *almost no restriction on target data distributions*
- *discretization schemes matter*

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

- *almost no restriction on target data distributions*
- *discretization schemes matter*

Question: which discretization adapts to low-dimensional structure?

Main result: convergence in total variation

Theorem 3 (L., Huang, Chen '24)

With exact score functions, both DDPM & DDIM (their original form) yield $\text{TV}(p_{X_1}, p_{Y_1}) \leq \varepsilon$ in

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

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

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

Main result: convergence in total variation

Theorem 3 (L., Huang, Chen '24)

With exact score functions, both DDPM & DDIM (their original form) yield $\text{TV}(p_{X_1}, p_{Y_1}) \leq \varepsilon$ in

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

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

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

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

Main result: convergence in total variation

Theorem 3 (L., Huang, Chen '24)

With exact score functions, both DDPM & DDIM (their original form) yield $\text{TV}(p_{X_1}, p_{Y_1}) \leq \varepsilon$ in

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

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

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

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

- *originally derived to optimize variational lower bounds!*

— Ho, Jain, Abbeel '20
— Song, Meng, Ermon '21

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

Table 1: comparison with prior DDPM theory

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

Table 2: comparison with prior DDIM theory

Crucial choices of coefficients

Both DDPM and DDIM update rules take the form

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

Crucial choices of coefficients

Both DDPM and DDIM update rules take the form

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

Theorem 4 (L., Huang, Chen '24)

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

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

Crucial choices of coefficients

Both DDPM and DDIM update rules take the form

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

Theorem 4 (L., Huang, Chen '24)

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

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

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

- both DDIM and DDPM satisfy this.
- this choice of step size is inevitable.

What if the scores are inexact?

Theorem 5 (L., Huang, Chen '24, informal)

Under the same low-dimensional assumptions, both DDIM and DDPM satisfy

$$\mathrm{TV}(p_{X_1}, p_{Y_1}) = \tilde{O}\left(\frac{k}{T} + \varepsilon_{\mathrm{est}}\right).$$

What if the scores are inexact?

Theorem 5 (L., Huang, Chen '24, informal)

Under the same low-dimensional assumptions, both DDIM and DDPM satisfy

$$\mathrm{TV}(p_{X_1}, p_{Y_1}) = \tilde{O}\left(\frac{k}{T} + \varepsilon_{\text{est}}\right).$$

- **DDPM:**

$$\varepsilon_{\text{est}} = \varepsilon_{\text{score}} \left(= \left\{ \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\|s_t(X_t) - s_t^*(X_t)\|_2^2 \right] \right\}^{1/2} \right).$$

What if the scores are inexact?

Theorem 5 (L., Huang, Chen '24, informal)

Under the same low-dimensional assumptions, both DDIM and DDPM satisfy

$$\mathrm{TV}(p_{X_1}, p_{Y_1}) = \tilde{O}\left(\frac{k}{T} + \varepsilon_{\text{est}}\right).$$

- **DDPM:**

$$\varepsilon_{\text{est}} = \varepsilon_{\text{score}} \left(= \left\{ \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\|s_t(X_t) - s_t^*(X_t)\|_2^2 \right] \right\}^{1/2} \right).$$

- **DDIM:** higher-order approximation errors

$$\partial^i s_t \approx \partial^i s_t^*, \quad i = 1, 2.$$

What if the scores are inexact?

Theorem 5 (L., Huang, Chen '24, informal)

Under the same low-dimensional assumptions, both DDIM and DDPM satisfy

$$\mathrm{TV}(p_{X_1}, p_{Y_1}) = \tilde{O}\left(\frac{k}{T} + \epsilon_{\text{est}}\right).$$

How large does the training set need to be?

Can diffusion models adapt to intrinsic low dimensionality?

- Sampling efficiency: number of reverse iterations T ✓
- Training sample size: number of training samples n

Adaptation during training

Theorem 6 (L., Huang, Chen, informal)

For any T , there exist estimators satisfying the score conditions required by both DDIM and DDPM with the sample size n at most

$$n = R_0^{O(k)} T^{O(k)} \varepsilon_{\text{est}}^{-2}$$

Adaptation during training

Theorem 6 (L., Huang, Chen, informal)

For any T , there exist estimators satisfying the score conditions required by both DDIM and DDPM with the sample size n at most

$$n = R_0^{O(k)} T^{O(k)} \varepsilon_{\text{est}}^{-2}$$

- **Iteration complexity:** $T \asymp k/\varepsilon$;

Adaptation during training

Theorem 6 (L., Huang, Chen, informal)

For any T , there exist estimators satisfying the score conditions required by both DDIM and DDPM with the sample size n at most

$$n = R_0^{O(k)} T^{O(k)} \varepsilon_{\text{est}}^{-2}$$

- **Iteration complexity:** $T \asymp k/\varepsilon$;
- “ $T \asymp k/\varepsilon$ ” + Thm. 6 $\implies n = R_0^{O(k)} k^{O(k)} \varepsilon^{-O(k)}$.

Adaptation during training

Theorem 6 (L., Huang, Chen, informal)

For any T , there exist estimators satisfying the score conditions required by both DDIM and DDPM with the sample size n at most

$$n = R_0^{O(k)} T^{O(k)} \varepsilon_{\text{est}}^{-2}$$

- **Iteration complexity:** $T \asymp k/\varepsilon$;
- “ $T \asymp k/\varepsilon$ ” + Thm. 6 $\implies n = R_0^{O(k)} k^{O(k)} \varepsilon^{-O(k)}$.
- **Estimator:** simple kernel-based estimators.

Paper	Sampler	Structural assumption	Training sample complexity	Low-dim. adaptation
Dou et al. '24	DDPM	β -smooth density on $\mathbb{R}^d$	$\varepsilon^{-\Theta(d)}$	$\times$
Wibisono et al. '24	DDPM	Lipschitz score on $\mathbb{R}^d$	$\varepsilon^{-(d+4)}$	$\times$
Cai & Li '25	DDIM	β -Hölder density on $\mathbb{R}^d$	$\varepsilon^{-(d+2\beta)/\beta}$	$\times$
Oko et al. '23	DDPM	β -Besov density on a k -subspace	$\varepsilon^{-(k+2\beta)/(\beta+1)}$	✓
Azangulov et al. '24	DDPM	β -smooth density on a smooth compact k -manifold	$\varepsilon^{-(2\beta+k)/(\beta+1)}$ provided $d \leq \text{poly}(1/\varepsilon)$	✓
Yakovlev–Puchkin '25	DDPM	noisy β -Hölder generator from $[0, 1]^k$	$d^{O(k)} \varepsilon^{-(2\beta+k)/\beta}$	✓
Our work	DDPM & DDIM	intrinsic covering number	$R_0^{O(k)} k^{O(k)} \varepsilon^{-O(k)}$	✓

Table 3: comparison of training sample complexity

Concluding remarks

- the original DDPM and DDIM adapt automatically to low dimensionality.
- the lower bound shows that the step sizes cannot be chosen arbitrarily.
- score learning can also adapt to low dimensionality.

Papers:

“Low-dimensional adaptation of diffusion models: convergence in total variation,”

J. Liang, Z. Huang, Y. Chen, arXiv:2501.12982, COLT 2025

Key properties

- **Tweedie's formula.** Define

$$\mu_{0|t}(x) := \mathbb{E}[X_0 \mid X_t = x], \quad \Sigma_{0|t}(x) := \text{Cov}(X_0 \mid X_t = x).$$

Then

$$s_t^*(x) = \frac{\sqrt{\bar{\alpha}_t} \mu_{0|t}(x) - x}{1 - \bar{\alpha}_t},$$
$$\nabla s_t^*(x) = \frac{\bar{\alpha}_t}{(1 - \bar{\alpha}_t)^2} \Sigma_{0|t}(x) - \frac{1}{1 - \bar{\alpha}_t} I_d.$$

- **Posterior concentration (informal).** For typical x , there is some $x_0^*(x) \in \text{supp}(p_{\text{data}})$ such that

$$\mathbb{P}\left(\|X_0 - x_0^*(x)\|_2 \gtrsim \sqrt{\frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}} k \log T \mid X_t = x\right) \leq T^{-ck}.$$

Proof sketch

- **Sampling: discretization error**

$$\mathbb{E} \left\| s_t^*(X_t) - s_{t_k}^*(X_{t_k}) \right\|_2^2 \stackrel{\text{Tweedie}}{\rightsquigarrow} \mathbb{E} \left\| \mu_{0|t}(X_t) - \mu_{0|t_k}(X_{t_k}) \right\|_2^2 = \tilde{O}(k).$$

- **Training: estimate s_t^* via $\mu_{0|t}$**

$$\mu_{0|t}(x) = \frac{\int x_0 \phi_{1-\bar{\alpha}_t}(x - \sqrt{\bar{\alpha}_t} x_0) p_{\text{data}}(dx_0)}{\int \phi_{1-\bar{\alpha}_t}(x - \sqrt{\bar{\alpha}_t} x_0) p_{\text{data}}(dx_0)},$$
$$\hat{\mu}_{0|t}(x) = \frac{\sum_{i=1}^n X_0^{(i)} \phi_{1-\bar{\alpha}_t}(x - \sqrt{\bar{\alpha}_t} X_0^{(i)})}{\sum_{i=1}^n \phi_{1-\bar{\alpha}_t}(x - \sqrt{\bar{\alpha}_t} X_0^{(i)})}.$$