Approximation and Generalization Abilities of Score-Based Neural Network Generative Models for Sub-Gaussian Distributions

Guoji Fu, Wee Sun Lee

National University of Singapore
School of Computing

Score-based Generative Models (Diffusion Models)

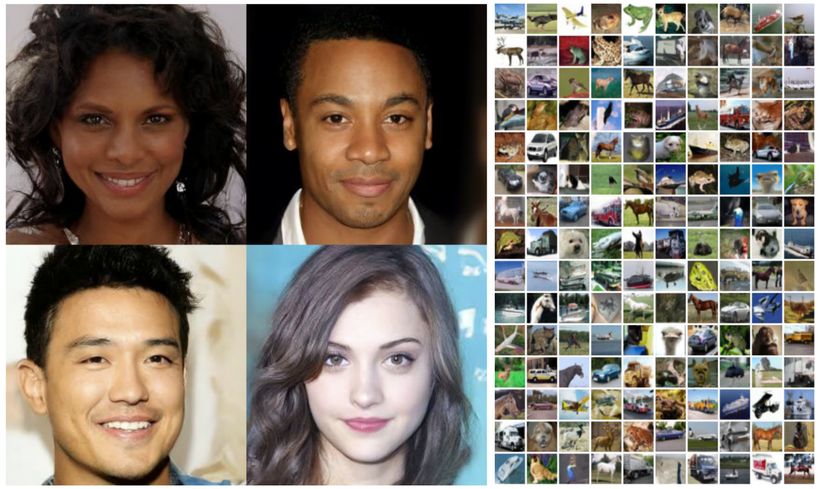


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

DDPM, 2020

Stable diffusion, 2022

SORA, 2024

Score-based Generative Models (Diffusion Models)

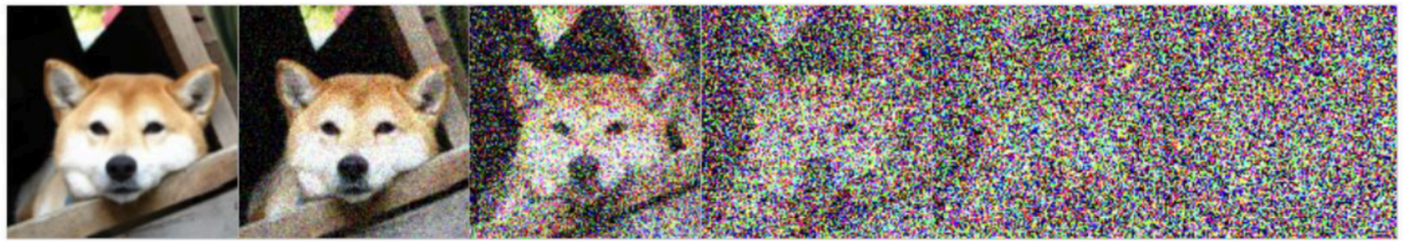
Understanding Forward & Reverse Processes through SDEs [Song et al., 2021]

Forward process: gradually adds noise to transform the signal into noise

OU Process:

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad (0 \le t \le T), \qquad X_0 \sim P_{\text{data}}$$

Forward/Diffusion Process



Reverse/Denoise Process

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log p_{T-t}(Y_t))\,\mathrm{d}t + \sqrt{2}\mathrm{d}B'_t \quad (0 \le t \le T), \qquad Y_0 \sim P_T$$

Reverse process: starting with the noise and recover the signal from noise

Connection to DDPM: Discretize OU process we obtain DDPM

Score-based Generative Models (Diffusion Models)

OU Process: **Reverse process:**

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log p_{T-t}(Y_t))\,\mathrm{d}t + \sqrt{2}\mathrm{d}B'_t \quad (0 \le t \le T), \qquad Y_0 \sim P_T$$

Generating new data:

- Learn a score estimator $\phi(\cdot, t)$ to replace $\nabla \log p_t(\cdot)$
- OU process converges exponentially to Gaussian distribution, replace P_T by $\mathcal{N}(0, I_d)$
- $\nabla \log p_t(\cdot)$ can blew up as $t \to 0$; adopt early stopping at $t_0 > 0$

$$\mathrm{d}\widehat{Y}_t^{\gamma_d} = (\widehat{Y}_t^{\gamma_d} + 2\phi(\widehat{Y}_t^{\gamma_d}, T-t))\mathrm{d}t + \sqrt{2}\mathrm{d}B'_t \quad (0 \le t \le T - t_0), \quad \widehat{Y}_0^{\gamma_d} \sim \mathcal{N}(0, I_d) =: \gamma_d$$

Denoising score matching:

$$\ell(\phi, X_0) := \int_{t=0}^{T} \mathbb{E}_{X_t|X_0}\left[\left\|\phi(X_t, t) - \nabla \log p_t(X_t|X_0)\right\|_2^2\right]\mathrm{d}t$$

Problem statement:

- Given $\{x^{(i)}\}_{i=1}^n \sim P_0^{\otimes n}$
- Our goal:

$$\mathrm{TV}(P_0, \widehat{P}_{t_0}^{\gamma_d}) \le ?$$

Analysis of Diffusion Models

Generalization analysis

Table 1: $P_0, \widehat{P}_0$: true and learned data distributions; p_0: true density function; KDE: kernel-based density estimator; $\hat{\psi}_t(\cdot)$ kernel function; s: smoothness parameter; t_0: early stopping time

Paper	Assumption	Estimator	Metric	Bound
(Zhang et al., 2024)	Sub-Gaussian P_0	KDE: $\frac{\nabla \hat{p}_t(\cdot)}{\hat{p}_t(\cdot)} \mathbb{1}_{\hat{p}_t(\cdot) > \rho_n}$	$\mathsf{TV}(P_{t_0}, \widehat{P}_{t_0})$	$\tilde{\mathcal{O}}(n^{-1/2} t_0^{-d/4})$
	Sub-Gaussian P_0 Sobolev class p_0		$\mathsf{TV}(P_0, \widehat{P}_0)$	$\tilde{\mathcal{O}}\left(n^{\frac{-s}{d+2s}}\right)$
(Wibisono et al., 2024)	Sub-Gaussian P_0 L-Lipschitz score	KDE: $\frac{\nabla \hat{p}_t(\cdot)}{\hat{p}_t(\cdot) \vee \rho_n}$	$\mathsf{TV}(P_0, \widehat{P}_0)$	$\tilde{\mathcal{O}}\left(L^{\frac{d+2}{d+4}} n^{\frac{-1}{d+4}}\right)$
(Dou et al., 2024)	$\mathsf{supp}(P_0) = [0,1]$ Hölder class p_0	KDE: $\frac{\nabla \hat{\psi}_t(\cdot)}{\hat{p}_t(\cdot) \vee \rho(\cdot, t)}$	$\mathsf{TV}(P_0, \widehat{P}_0)$	$\mathcal{O}\left(n^{\frac{-s}{1+2s}}\right)$
(Oko et al., 2023)	$\mathsf{supp}(P_0) = [0,1]^d$ lower bounded p_0 Besov class p_0	ReLU DNNs	$\mathsf{TV}(P_0, \widehat{P}_0)$	$\tilde{\mathcal{O}}(n^{\frac{-s}{d+2s}})$
Theorem 3	Sub-Gaussian P_0	ReLU DNNs	$\mathsf{TV}(P_{t_0}, \widehat{P}_{t_0})$	$\tilde{\mathcal{O}}(n^{-1/2} t_0^{-d/4})$
Corollary 1	Sub-Gaussian P_0 Sobolev class p_0		$\mathsf{TV}(P_0, \widehat{P}_0)$	$\tilde{\mathcal{O}}(n^{\frac{-s}{d+2s}})$
Corollary 2	$\mathsf{supp}(P_0) = [0,1]^d$ Besov class p_0			

Existing results

NN-based method:

- Density lower bounded;

KDE-based methods:

- Lipschitz score functions;
- Not widely used in practices

Our results

NN-based method:

- Convergence rates under only sub-Gaussian dist.;

Nearly minimax optimal rates:

- Without density lower bounded;
- Without Lipschitz score;

Problem Settings & Overview Idea

Assumption 1: (Sub-Gaussian Distribution) *The target data distribution P_0 is α-sub-Gaussian.*

for some $0 < \alpha < \infty$ if for all $\theta \in \mathbb{R}^d$:

$$\mathbb{E}_{X\sim P}\left[\theta^\top (X - \mathbb{E}_{X\sim P}[X])\right] \le \exp\left(\frac{\alpha^2\|\theta\|_2^2}{2}\right)$$

Overview Idea: Assume merely **sub-Gaussian Assumption 1** holds,

- **Step 1: Construct a regularized KDE-based score estimator** $\dfrac{\nabla \hat{p}_\sigma(\cdot)}{\hat{p}_\sigma(\cdot)\vee\rho_n} = \dfrac{\nabla\left(\frac{1}{n}\sum_{i=1}^n \varphi_\sigma(\cdot - x^{(i)})\right)}{\frac{1}{n}\sum_{i=1}^n \varphi_\sigma(\cdot - x^{(i)})\vee\rho_n}$

  we show via empirical Bayes techniques: $\mathbb{E}_{\{x^{(i)}\}_{i=1}^n}\left[\int_{\mathbb{R}^d}\left\|\dfrac{\nabla p_\sigma(x)}{p_\sigma(x)} - \dfrac{\nabla \hat{p}_\sigma(x)}{\hat{p}_\sigma(x)\vee\rho_n}\right\|_2^2 p_\sigma(x)\mathrm{d}x\right] \lesssim \tilde{\mathcal{O}}(n^{-1}\sigma^{-d-2})$

  (NOT require Lipschitz score assumption as opposed to [Wibisono, et, al. 2024])

- **Step 2: Neural network approximation for KDE estimator (score approximation)**

  $\mathcal{NN}(\text{width} \le n^{\frac{3}{d}}\log_2 n), \text{depth} \le \mathcal{O}(\log^2 n))$ such that: $\displaystyle\int_{t=t_0}^T \mathbb{E}_{x_t}\left[\left\|\nabla \log p_t(x_t) - \phi(x_t,t)\right\|_2^2\right]\mathrm{d}t \lesssim \tilde{\mathcal{O}}(t_0^{-d/2}n^{-1})$

- **Step 3: Generalization bounds by Bernstein's inequality (score estimation)**

  With high probability: $\displaystyle\int_{t=t_0}^T \mathbb{E}_{x_t}\left[\left\|\nabla \log p_t(x_t) - \widehat{\phi}(x_t,t)\right\|_2^2\right]\mathrm{d}t \lesssim \tilde{\mathcal{O}}(t_0^{-d/2}n^{-1})$

Distribution Estimation Error in TV Distance

Assumption 2: (Sobolev Class of Density)

The density p_0 belongs to the Sobolev ball with $0 < s \leq 2$.

Assumption 3: (Besov Class of Density)

The density $p_0 \in L^q([0,1]^d) \cap U(B^s_{q,q'}([0,1]^d); C)$ for some $C > 0$, where $1 \leq q \leq \infty, 0 < q' \leq \infty$ and $0 < s \leq 2$.

Theorem (Distribution Estimation error in TV distance)

Let $t_0 = n^{-\frac{2}{d+2s}}$ and $T = n^{\mathcal{O}(1)}$. Assume that Assumption 2 or Assumption 3 hold and the network size $\mathcal{NN}(\text{width} \leq n^{\frac{3}{d}} \log_2 n), \text{depth} \leq \mathcal{O}(\log^2 n))$, with probability at least $1 - 1/n$, it holds that

$$\mathbb{E}_{\{x^{(i)}\}_{i=1}^n}\left[\text{TV}(P_0, \widehat{P}_{t_0}^{\gamma_d})\right] \lesssim n^{-\frac{s}{d+2s}}$$

which is minimax optimal up to a polynomial log factor, that is, it holds that

$$n^{-\frac{s}{2s+d}} \leq \inf_{\widehat{s}\text{: estimator}} \sup_{P_0} \mathbb{E}_{\{x^{(i)}\}_{i=1}^n}\left[\text{TV}(P_0, \widehat{P}_{\widehat{s}})\right]$$

Distribution Estimation Error in TV Distance

Theorem (Distribution Estimation error in TV distance)

Let $t_0 = n^{-\frac{2}{d+2s}}$ and $T = n^{\mathcal{O}(1)}$. Assume that Assumption 2 or Assumption 3 hold and the network size $\mathcal{NN}(\text{width} \leq n^{\frac{3}{d}} \log_2 n), \text{depth} \leq \mathcal{O}(\log^2 n))$, with probability at least $1 - 1/n$, it holds that

$$\mathbb{E}_{\{x^{(i)}\}_{i=1}^n} \left[\text{TV}(P_0, \widehat{P}_{t_0}^{\gamma_d}) \right] \lesssim n^{-\frac{s}{d+2s}}$$

Observations:

(1) Dependence of network width in dimension d, i.e., $\text{width} \leq \tilde{\mathcal{O}}(n^{\frac{3}{d}})$ $d \nearrow$ $\Longrightarrow \text{width} \searrow$ (Seems counterintuitive!!)

(2) Bias-variance trade-off:

- **Increasing bias:** $d \nearrow \Longrightarrow n^{-\frac{s}{d+2s}} \nearrow$
  $\Longrightarrow$ statistical difficulty $\nearrow$ (curse of dimensionality)

- **Reduce variance by smoothing**: $d \nearrow \Longrightarrow t_0 \nearrow$ (more Gaussian smoothing) $t_0 = n^{-\frac{2}{d+2s}}$
  $\Longrightarrow$ learning P_{t_0} become easier
  $\Longrightarrow$ the required size of network $\searrow$

Conclusion & Future Work

Conclusion:

We have show for score-based neural network generative models (a.k.a, diffusion models):

- Convergence rates under only sub-Gaussian distribution assumption;
- Nearly minimax optimal rates without density lower bounded & Lipschitz score assumptions;

Future work:

1. Theory of diffusion models on low-dimensional spaces, e.g., low-dim. manifolds
   Our current rates suffer from curse of dimensionality

2. Going beyond sub-Gaussian distributions, e.g., non-smooth, heavy-tail, sub-Weibull
   Our current score estimation analysis rely on sub-Gaussian assumption.

3. Higher-order smoothness
   Our current results require the smoothness parameter $0 < s \leq 2$.