Diffusion Models are Minimax Optimal Distribution Estimators

Kazusato Oko (The University of Tokyo / AIP RIKEN)

Joint work with Shunta Akiyama (The University of Tokyo)

Taiji Suzuki (The University of Tokyo / AIP RIKEN)

THE UNIVERSITY OF TOKYO AIP

May 4, 2023

ME-FoMo spotlight talk

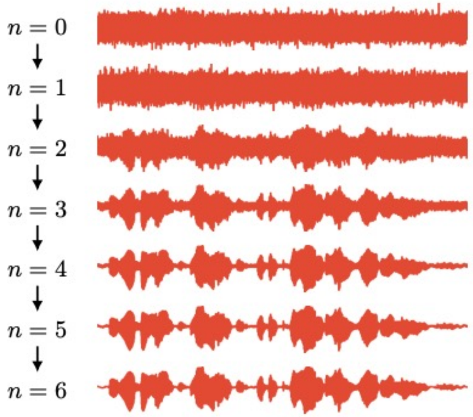
ICLR 2023 Workshop

Success of diffusion models

- Image, video, audio, etc…

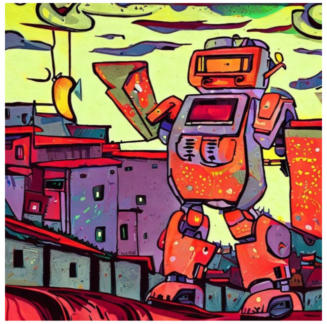
Image generated by DALL·E2

Video generated by Video Diffusion Models



Visualization of WaveGrad

- Several pictures I made with Stable Diffusion:



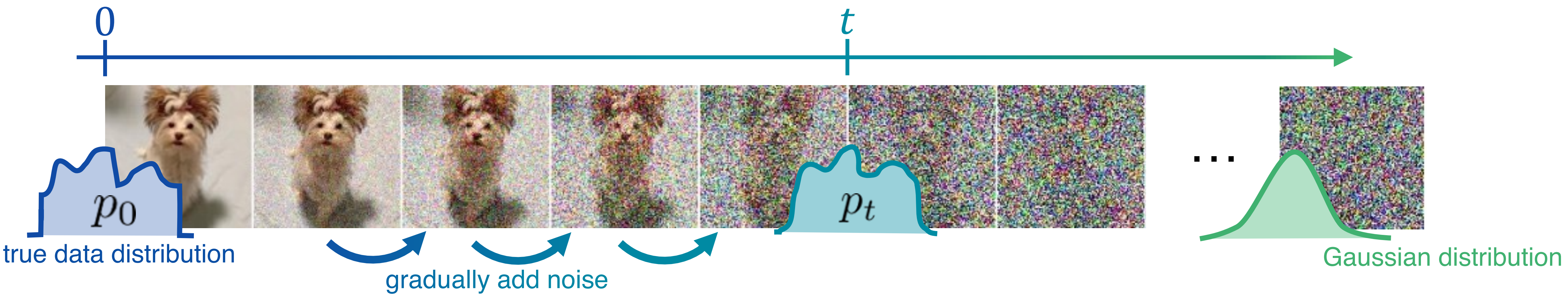
A mecha robot in a favela in an expressionist style

A high tech solarpunk utopia in the Amazon forest

Pikachu in the style of Munch’s “The Scream”

DALL·E2: A. Ramesh, et al. “Hierarchical Text-Conditional Image Generation with CLIP Latents”. *arXiv:2204.06125*, 2022; Video Diffusion Models: J. Ho, et al. “Video diffusion models”. *arXiv:2204.03458, 2022*; WaveGrad: N. Chen et al. “WaveGrad: Estimating Gradients for Waveform Generation”. *ICLR* 2021; Stable Diffusion: R. Rombach, et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. *arXiv:2112.10752*, 2022.

Formulation as SDE (Song et al., 2020): forward process



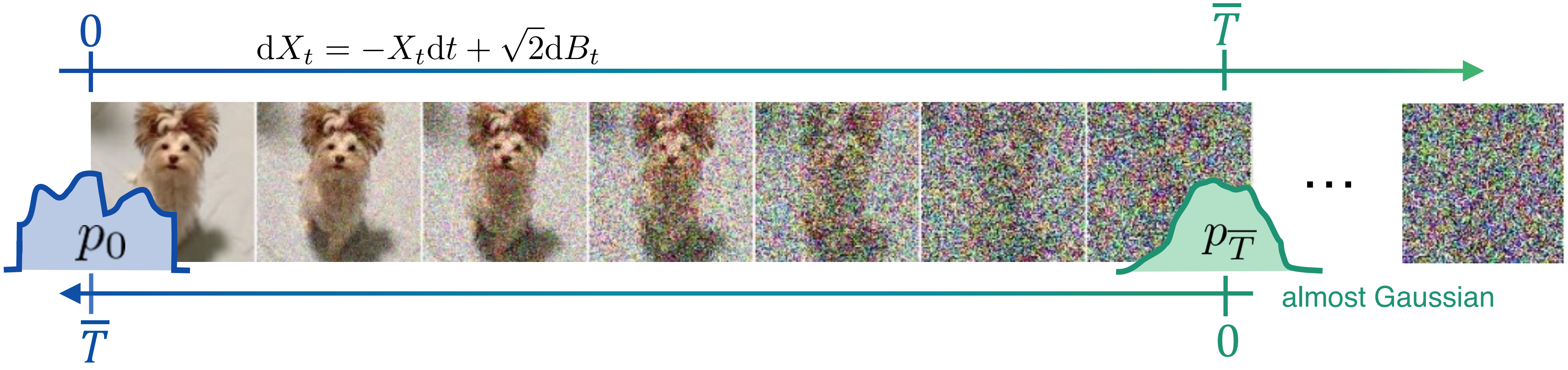
$$X_0 \sim p_0, \quad \mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \text{(OU process)}$$

Brownian motion

- The distribution of X_t at time t :

$$p_t(x) = \int p_0(y) \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - \mu_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y$$

Formulation as SDE (Song et al., 2020): backward process



$$Y_0 \sim p_{\overline{T}}, \qquad \mathrm{d}Y_t = (Y_t + 2\nabla \log p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

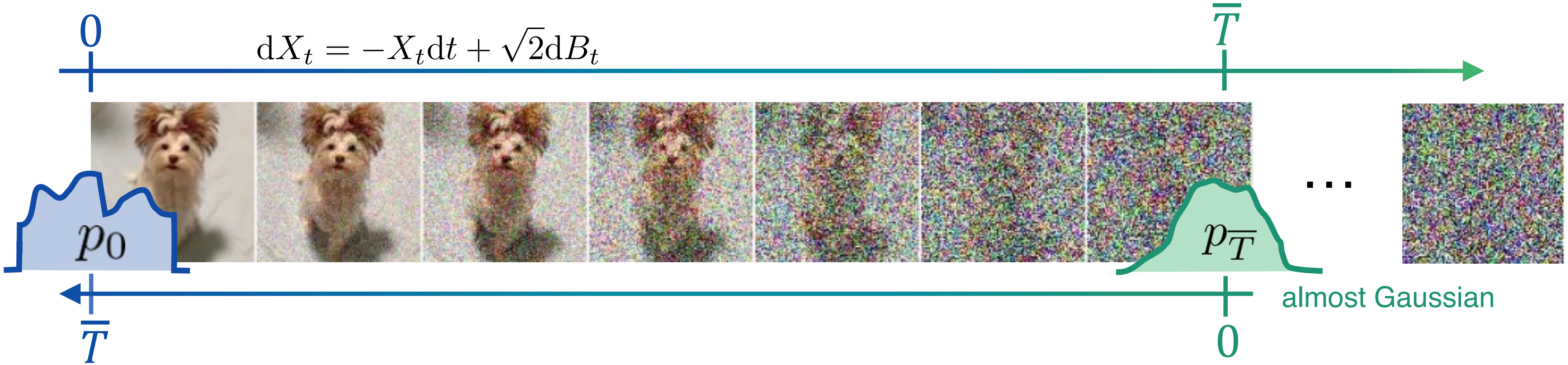
Brownian motion

$\Longrightarrow$ $Y_{\overline{T}} \sim p_0$ (recovers the true data distribution)

Y. Song et al. "Score-based generative modeling through stochastic differential equations". *ICLR* 2021

U. G. Haussmann & E. Pardoux. "Time Reversal of Diffusions". *The annals of Probability*, 14(4): 1188–1205, 1986.

Formulation as SDE (Song et al., 2020): backward process



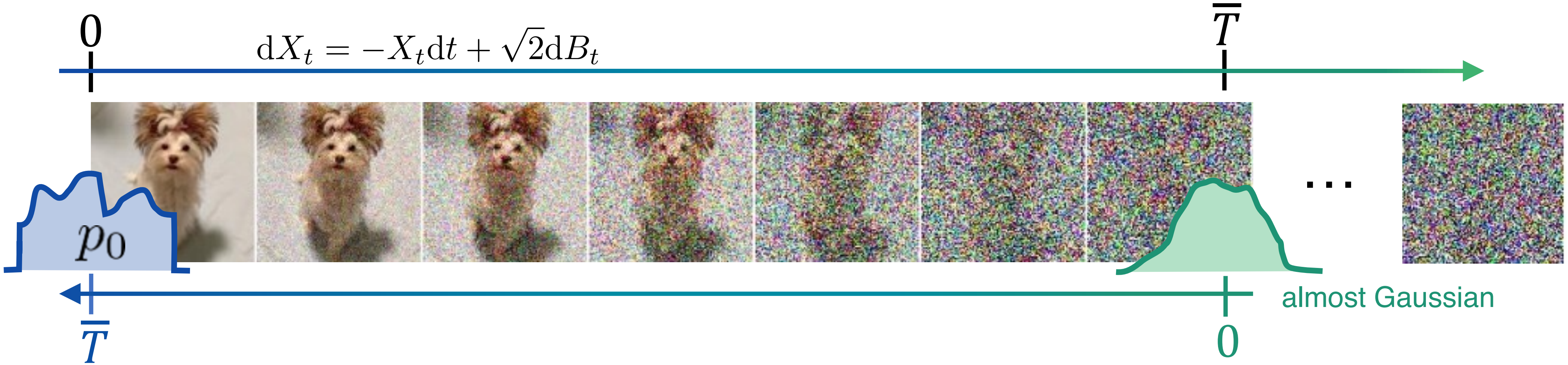
$$Y_0 \sim p_{\overline{T}}, \qquad \mathrm{d}Y_t = (Y_t + 2\boxed{\nabla \log p_{\overline{T}-t}(Y_t)})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

Brownian motion

The exact value of the score $\nabla \log p_t(x)$ cannot be obtained because it depends on p_0

Y. Song et al. "Score-based generative modeling through stochastic differential equations". *ICLR* 2021
U. G. Haussmann & E. Pardoux. "Time Reversal of Diffusions". *The annals of Probability*, 14(4): 1188–1205, 1986.

Formulation as SDE (Song et al., 2020): backward process



$$Y_0 \sim p_{\overline{T}}, \qquad \mathrm{d}Y_t = (Y_t + 2\boxed{\nabla \log p_{\overline{T}-t}(Y_t)})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

Brownian motion

The exact value of the score $\nabla \log p_t(x)$ cannot be obtained because it depends on p_0

$$\hat{Y}_0 \sim \mathcal{N}(0, I), \qquad \mathrm{d}\hat{Y}_t = (\hat{Y}_t + 2\boxed{\hat{s}(\hat{Y}_t, \overline{T} - t)})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

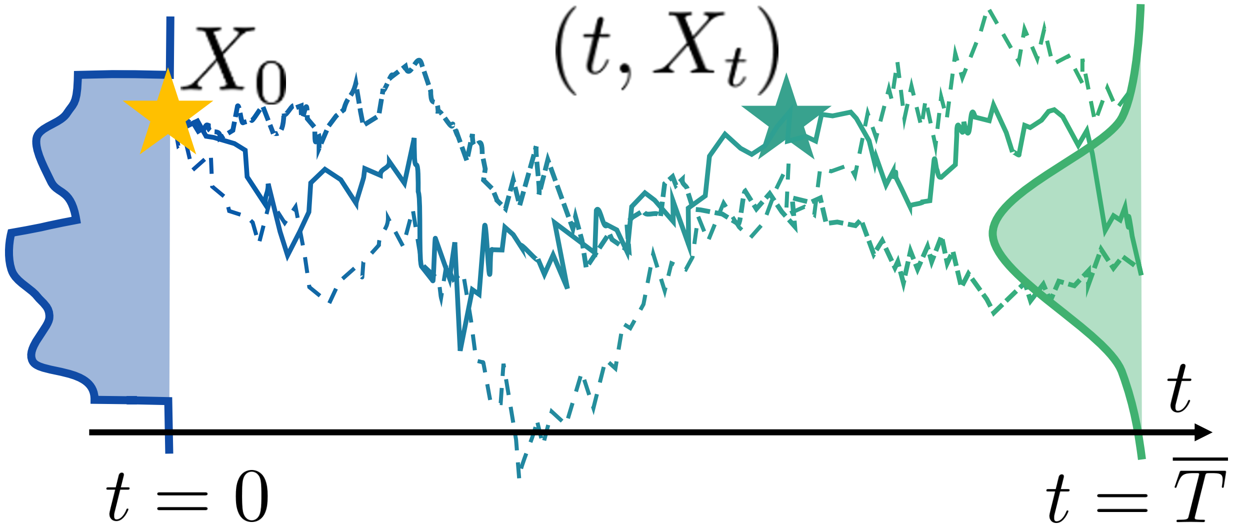
the score network, trained with finite sample

Y. Song et al. "Score-based generative modeling through stochastic differential equations". *ICLR* 2021
U. G. Haussmann & E. Pardoux. "Time Reversal of Diffusions". *The annals of Probability*, 14(4): 1188–1205, 1986.

Score matching to train the score network

- The true score minimizes the following loss:

$$\mathbb{E}_{X_0\sim p_0}\left[\underbrace{\int_{t=0}^{\overline{T}} \mathbb{E}_{X_t|X_0\sim\mathcal{N}(m_tX_0,\sigma_t^2)}\left[\|s(X_t|X_0,t)-\nabla\log p_t(X_t|X_0)\|^2\right]\mathrm{d}t}_{\text{Computed by sampling } (t,X_t)\sim \mathrm{Unif}[0,\overline{T}]\times\mathcal{N}(m_tX_0,\sigma_t^2)}\right]$$



Score matching to train the score network

- The true score minimizes the following loss:

$$\mathbb{E}_{X_0\sim p_0}\left[\int_{t=0}^{\overline{T}} \mathbb{E}_{X_t|X_0\sim\mathcal{N}(m_tX_0,\sigma_t^2)}\left[\|s(X_t|X_0,t)-\nabla\log p_t(X_t|X_0)\|^2\right]\mathrm{d}t\right]$$

- Replace the expectation w.r.t. p_0 by finite sample $x_1,\cdots,x_n \overset{\text{i.i.d}}{\sim} p_0$

$$\operatorname*{argmin}_{s\in\mathcal{S}:\ \text{DNNs}} \frac{1}{n}\sum_{i=1}^{n}\left[\int_{t=0}^{\overline{T}} \mathbb{E}_{X_t|x_i\sim\mathcal{N}(m_tx_i,\sigma_t^2)}\left[\|s(X_t|x_i,t)-\nabla\log p_t(X_t|x_i)\|^2\right]\mathrm{d}t\right]$$

empirical score matching loss

How close the generated distribution is to p_0?

P. Vincent. "A connection between score matching and denoising autoencoders." *Neural computation*, 23(7):1661– 1674, 2011.
More formally, the lower limit of the integral interval in the empirical score matching loss is needed to be replaced by a very small truncation time $\underline{T}$ for a technical reason.

Score matching to train the score network

- The true score minimizes the following loss:

$$\boxed{\mathbb{E}_{X_0\sim p_0}}\left[\int_{t=0}^{T}\mathbb{E}_{X_t|X_0\sim\mathcal{N}(m_tX_0,\sigma_t^2)}\left[\|s(X_t|X_0,t)-\nabla\log p_t(X_t|X_0)\|^2\right]\mathrm{d}t\right]$$

- Replace the expectation w.r.t. p_0 by finite sample $x_1,\cdots,x_n\overset{\text{i.i.d}}{\sim}p_0$

$$\boxed{\underset{s\in\mathcal{S}:\ \text{DNNs}}{\operatorname{argmin}}}\ \boxed{\frac{1}{n}\sum_{i=1}^{n}}\left[\int_{t=0}^{T}\mathbb{E}_{X_t|x_i\sim\mathcal{N}(m_tx_i,\sigma_t^2)}\left[\|s(X_t|x_i,t)-\nabla\log p_t(X_t|x_i)\|^2\right]\mathrm{d}t\right]$$

empirical score matching loss

How good is the diffusion modeling
as a distribution estimator based on n data?

P. Vincent. "A connection between score matching and denoising autoencoders." *Neural computation*, 23(7):1661– 1674, 2011.

More formally, the lower limit of the integral interval in the empirical score matching loss is needed to be replaced by a very small truncation time $\underline{T}$ for a technical reason.

Existing work on error analysis

- Most works analyzed how the score estimation error, **given as an assumption**, affects the generated data distribution
  - Continuous time: Song et al. (2021); De Bortoli et al. (2022a)
  - Discrete time: Lee et al. (2022a;b); Chen et al. (2023)
  - Non-quantitative bound under manifold assumption: Pidstrigach (2022)

- Sample complexity bounds
  - W1 bound of $n^{-1/d}$ with manifold assumption: De Bortoli et al. (2021)
    - **Not considering "generalization" and unimprovable**
    - Based on the convergence of the empirical measure $W_1(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, p_0) \simeq n^{-1/d}$ (Weed and Bach, 2019)

- Concurrent work (appeared after the submission of this work): Chen et al. (2023)

Song et al."Maximum likelihood training of score-based diffusion models". *NeurIPS* 2021; Chen et al: "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions". *ICLR* 2023.; Lee et al. "Convergence of score-based generative modeling for general data distributions." *NeurIPS 2022 Workshop on Score-Based Methods*, 2022a.; Lee et al. "Convergence for score-based generative modeling with polynomial complexity", *NeurIPS 2022*, 2022b.; De Bortoli et al. "Diffusion Schrödinger bridge with applications to score-based generative modeling". *NeurIPS* 2021; De Bortoli et al. "Convergence of denoising diffusion models under the manifold hypothesis". *TMLR* 2022.; Weed and Bach. "Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance". *Bernoulli*, 25(4A):2620–2648, 2019. Chen et al.: "Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data". *arXiv:2302.07194*, 2023.

Analysis of diffusion models from statistical learning theory

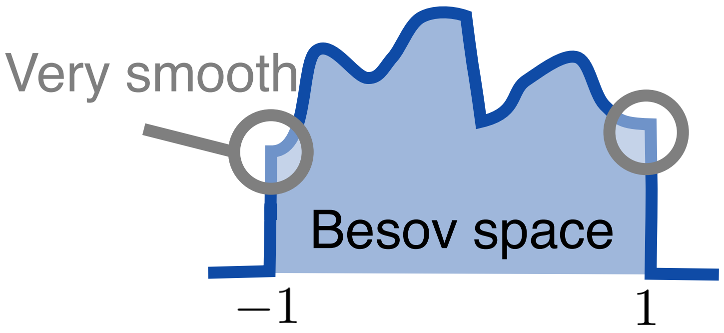
A1 p_0 is supported on $[-1,1]^d$, upper and lower bounded in the support, and

$$p_0 \in B^s_{p,q,C}$$

with $s > (1/p - 1/2)_+$ as a density function on $[-1,1]^d$.

- $B^s_{p,q,C}$: Besov space $B^s_{p,q}$ with the norm bounded by C (some constant)
  - Intuition: $\|f\|_{B^s_{p,q}(\Omega)} = \|f\|_{L^p(\Omega)} + \|D^s f\|_{L^p(\Omega)}$

A2 p_0 is sufficiently smooth on the edge of the support $[-1,1]^d \setminus [-1+n^{-\frac{1-\delta}{d}}, 1-n^{-\frac{1-\delta}{d}}]^d$.



Analysis of diffusion models from statistical learning theory

- Hypothesis network class: sparsity-constrainted deep ReLU networks

(Schmidt-Hieber, 2020; Suzuki, 2019)

$$\mathcal{S}(L \text{ (depth)}, W \text{ (width)}, S \text{ (sparsity-constraint; num. of non-zero params)}, B \text{ (magnitude)})$$

$$:= \Big\{(A^L \mathrm{ReLU}(\cdot) + b^L) \circ \cdots \circ (A^1 x + b^1) \,\Big|\, A^i \in \mathbb{R}^{w_i \times w_{i+1}}, b^i \in \mathbb{R}^{w_{i+1}}, \|w\|_\infty \le W,$$
$$\sum_{i=1}^{L} (\|A^i\|_0 + \|b^i\|_0) \le S, \max \|A^i\|_\infty \vee \|b^i\|_\infty \le B \Big\}$$

- Sparsity-constraint yields tighter generalization error bounds

J. Schmidt-Hieber. “Nonparametric regression using deep neural networks with Relu activation function.” *The Annals of Statistics*, 48(4):1875–1897, 2020.
T. Suzuki. “Adaptivity of deep Relu network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality”. *ICLR* 2019.

Main result ①: minimax optimality in TV

$\hat{s}$

Theorem 1

The generated data distribution by using the score network $\hat{s}$ that minimizes the empirical score matching loss over $\mathcal{S}(L, W, S, B)$ yields that

$$\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\mathrm{TV}(\hat{Y}_{\overline{T}}, X_0)\right] \lesssim n^{-\frac{s}{2s+d}} \log^9(n),$$

under the appropriate choice of $\overline{T}, \underline{T}, L, W, S,$ and B.

This rate is **the minimax optimal** (up to polylog), because it also holds that

$$n^{-\frac{s}{2s+d}} \lesssim \inf_{\hat{\mu}:\text{estimator}} \sup_{p_0 \in B^s_{p,q,C}} \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\mathrm{TV}(\hat{\mu}, X_0)\right].$$

More formally, $\hat{Y}_{\overline{T}}$ is needed to be replaced by $\hat{Y}_{\overline{T}-\underline{T}}$ for a technical reason in the above.

Basis decomposition tailored for score approximation

- B-spline basis decomposition of $p_0 (\in B^s_{p,q,C})$: $p_0(x) \approx \sum_{j=1}^{N} \alpha_j M^d_{a^j, b^j}(x)$ B-spline basis

(Devore & Popov, 1988)

- Approximation of $p_t(x)$:

$$p_t(x) = \int p_0(y) \underbrace{\frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - \mu_t y\|^2}{2\sigma_t^2}\right)}_{=: K_t(x|y)} \mathrm{d}y$$

approximation via B-spline basis

$$\approx \sum_{j=1}^{N} \alpha_j \underbrace{\int M^d_{a^j, b^j}(y) K_t(x|y) \mathrm{d}y}_{=: E_{a^j, b^j}(x,t)}$$

diffused B-spline basis

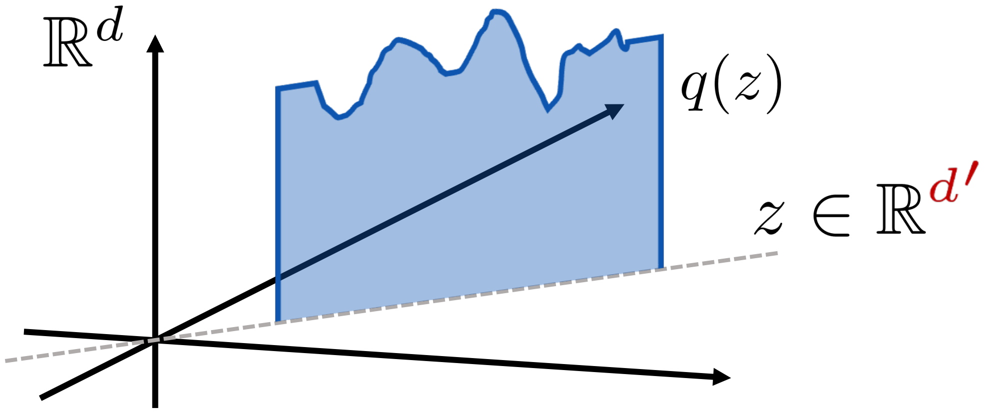
Approximated by NN very efficiently (polylog size)

❖ Approximate $\nabla p_t(x)$ in the same way and use $\nabla \log p_t(x) = \dfrac{\nabla p_t(x)}{p_t(x)}$

R. A. DeVore, & V. A. Popov. "Interpolation of Besov spaces. "*Transactions of the American Mathematical Society*, 305(1):397–414, 1988.

Main result ②: manifold hypothesis

$\hat{s}$



- p_0 lies on a d'-dimensional plane $(d' \leq d)$
- Density function on the canonical coordinate system on the plane q belongs to $B^s_{p,q,C}$

Theorem 2

Based on $\{x_i\}_{i=1}^n$, we can train the score network $\hat{s}$ that satisfies

$$\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\mathrm{W}_1(\hat{Y}_{\overline{T}}, X_0)\right] \lesssim n^{-\frac{s+1-\delta}{2s+d'}}.$$

($\delta(>0)$: arbitrarily fixed constant)

Diffusion models can avoid the curse of dimensionality!

More formally, $\hat{Y}_{\overline{T}}$ is needed to be replaced by $\hat{Y}_{\overline{T}-\underline{T}}$ for a technical reason in the above.