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

ICML International Conference On Machine Learning

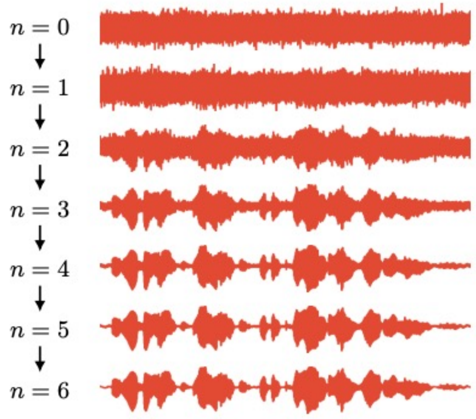
Fortieth International Conference on Machine Learning
Jul 23-29, 2023

Motivation

- Practical success of diffusion models in a wide range of data generating tasks

Image generated by DALL·E2

Video generated by Video Diffusion Models



Visualization of WaveGrad (**audio**)

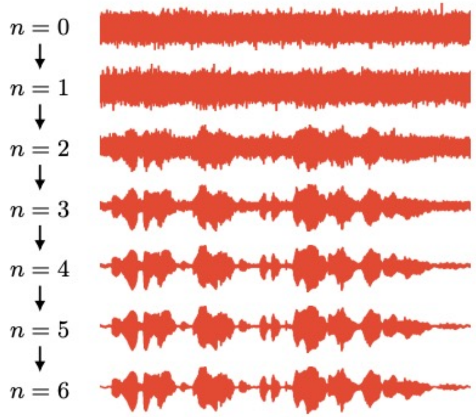
DALL·E2: A. Ramesh, et al. "Hierarchical Text-Conditional Image Generation with CLIP Latents". *arXiv:2204.06125*, 2022; Video Diffusion Models: J. Ho, et al. "Video diffusion models". *NeurIPS* 2022; WaveGrad: N. Chen et al. "WaveGrad: Estimating Gradients for Waveform Generation". *ICLR* 2021

Motivation

- Practical success of diffusion models in a wide range of data generating tasks

Image generated by DALL·E2

Video generated by Video Diffusion Models



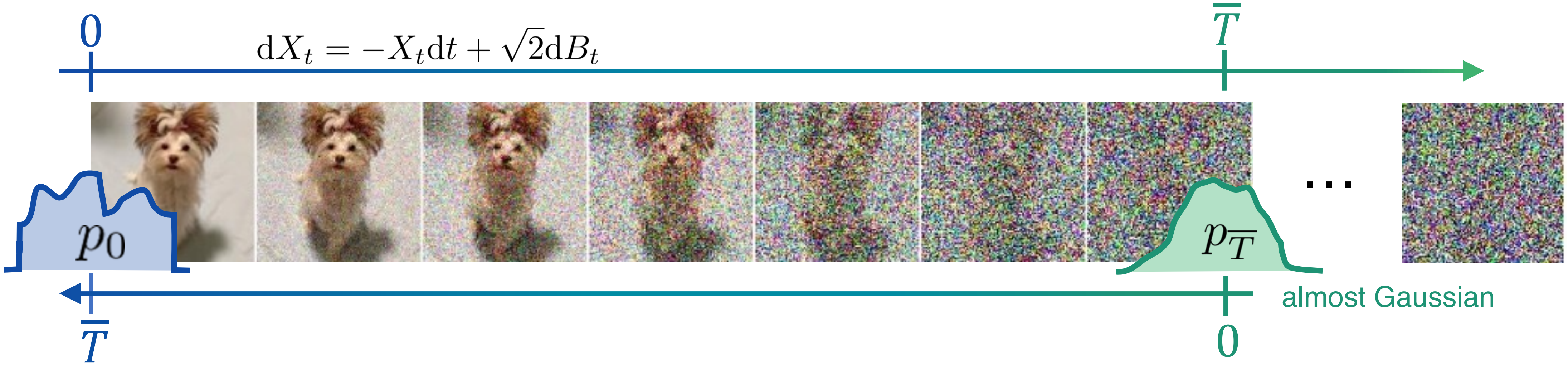
Visualization of WaveGrad (**audio**)

- Theoretical understandings of diffusion models are limited

We analyze diffusion models as a distribution learner via statistical learning theory

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”. *NeurIPS* 2022; WaveGrad: N. Chen et al. “WaveGrad: Estimating Gradients for Waveform Generation”. *ICLR* 2021

Formulation as SDE (Song et al., 2020)



$$Y_0 \sim \mathcal{N}(0,I) \approx p_{\overline{T}}, \qquad \mathrm{d}Y_t = (Y_t + 2\nabla \log p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

Brownian motion

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

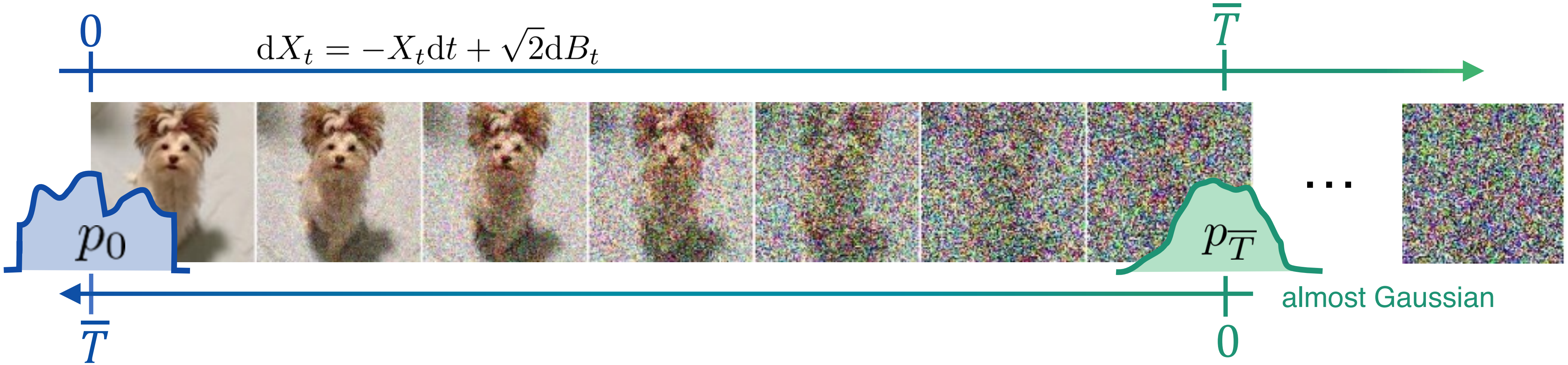
Note:

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

$(\mu_t = e^{-t},\ \sigma_t^2 = 1 - e^{-2t})$

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)



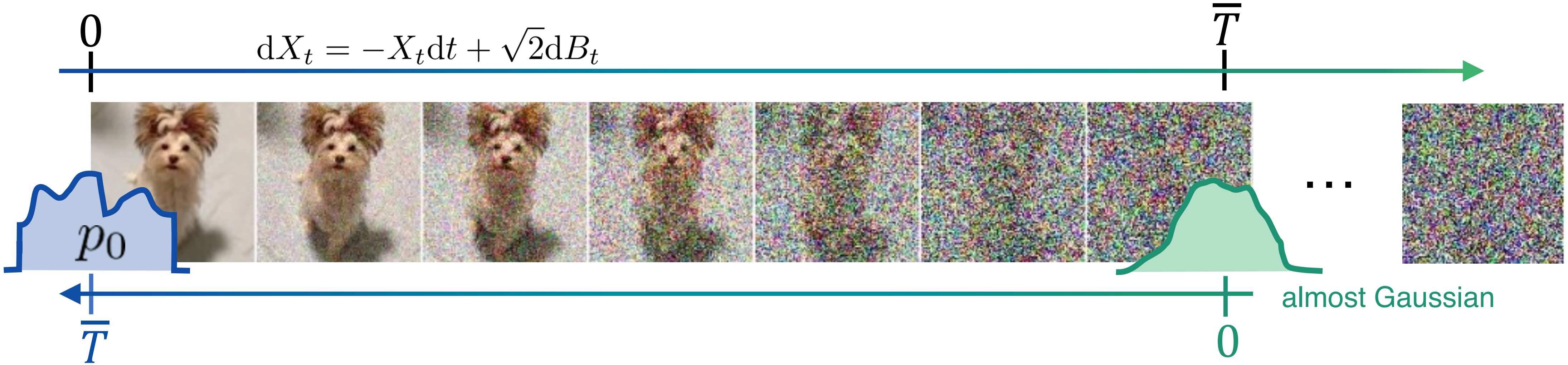
$Y_0 \sim \mathcal{N}(0,I) \approx p_{\overline{T}}, \quad \mathrm{d}Y_t = (Y_t + 2\nabla \log p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$

Brownian motion

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)



$Y_0 \sim \mathcal{N}(0,I) \approx p_{\overline{T}},$ $\quad \mathrm{d}Y_t = (Y_t + 2\nabla \log p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$

Brownian motion

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

$$\mathrm{d}\hat{Y}_t = (\hat{Y}_t + 2\hat{s}(\hat{Y}_t, \overline{T} - t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

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.

Existing work on error analysis

- If $\int_t \mathbb{E}_{X_t \sim p_t}[\|s(X_t,t) - \nabla \log p_t(X_t)\|^2]\mathrm{d}t \leq \varepsilon$, we have $\mathrm{TV}(\widehat{Y}_0, X_0) \leq \mathrm{poly}(\varepsilon, \eta, d)$ (propagation of the score matching error and discretization error)
  - Continuous time ($\eta = 0$): Song et al. (2021); De Bortoli et al. (2021)
  - Discrete time ($\eta > 0$):); De Bortoli et al. (2022); Lee et al. (2022a;b); Chen et al. (2023)
  - Non-quantitative bound under manifold assumption: Pidstrigach (2022)

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.

Existing work on error analysis

- If $\int_t \mathbb{E}_{X_t \sim p_t}[\|s(X_t,t) - \nabla \log p_t(X_t)\|^2]\mathrm{d}t \leq \varepsilon$, we have $\mathrm{TV}(\widehat{Y}_0, X_0) \leq \mathrm{poly}(\varepsilon, \eta, d)$ (propagation of the score matching error and discretization error)
  - Continuous time ($\eta = 0$): Song et al. (2021); De Bortoli et al. (2021)
  - Discrete time ($\eta > 0$):); De Bortoli et al. (2022); Lee et al. (2022a;b); Chen et al. (2023)
  - Non-quantitative bound under manifold assumption: Pidstrigach (2022)

We do not know how small ε can be with n training sample

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.

Existing work on error analysis

- If $\int_t \mathbb{E}_{X_t \sim p_t}[\|s(X_t,t) - \nabla \log p_t(X_t)\|^2]\mathrm{d}t \leq \varepsilon$, we have $\mathrm{TV}(\widehat{Y}_0, X_0) \leq \mathrm{poly}(\varepsilon, \eta, d)$ (propagation of the score matching error and discretization error)
  - Continuous time ($\eta = 0$): Song et al. (2021); De Bortoli et al. (2021)
  - Discrete time ($\eta > 0$):); De Bortoli et al. (2022); Lee et al. (2022a;b); Chen et al. (2023)
  - Non-quantitative bound under manifold assumption: Pidstrigach (2022)

We do not know how small ε can be with n training sample

- Estimation rate analysis
  - W1 bound of $n^{-1/d}$: De Bortoli et al. (2021)

  - 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.

Existing work on error analysis

- If $\int_t \mathbb{E}_{X_t \sim p_t}[\|s(X_t,t) - \nabla \log p_t(X_t)\|^2]\mathrm{d}t \leq \varepsilon$, we have $\mathrm{TV}(\widehat{Y}_0, X_0) \leq \mathrm{poly}(\varepsilon, \eta, d)$ (propagation of the score matching error and discretization error)
  - Continuous time ($\eta = 0$): Song et al. (2021); De Bortoli et al. (2021)
  - Discrete time ($\eta > 0$):); De Bortoli et al. (2022); Lee et al. (2022a;b); Chen et al. (2023)
  - Non-quantitative bound under manifold assumption: Pidstrigach (2022)

We do not know how small ε can be with n training sample

- Estimation rate analysis
  - W1 bound of $n^{-1/d}$: De Bortoli et al. (2021)
    - **can structural assumptions on the data improve this bound?: this work**
  - 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.

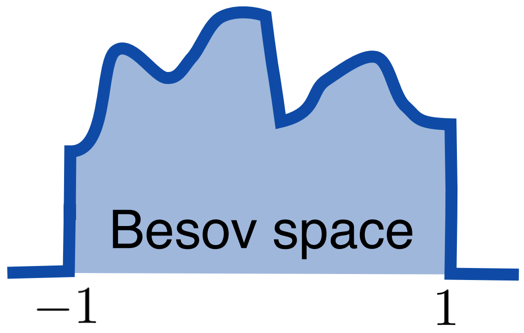
Problem settings

- Assume the true data belongs to some function space

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 $[-1,1]^d$.



Problem settings

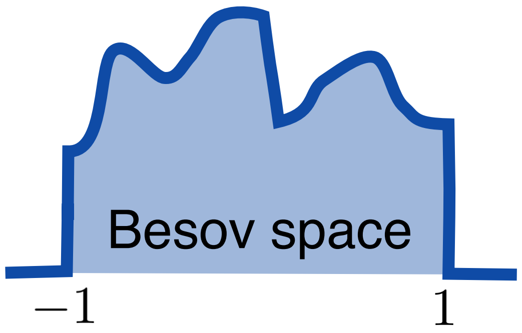
- Assume the true data belongs to some function space

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: $\|f\|_{B^s_{p,q}(\Omega)} = \|f\|_{L^p(\Omega)} + \|D^s f\|_{L^p(\Omega)}$



Problem settings

- Assume the true data belongs to some function space

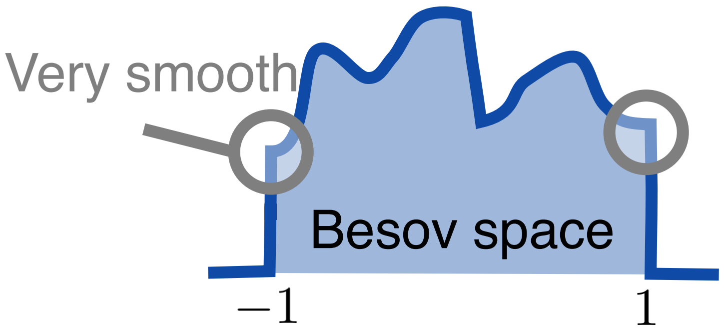
A1 p_0 is supported on $[-1,1]^d$, upper and lower bounded in the support, and

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

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 $[-1,1]^d \setminus [-1+n^{-\frac{1-\delta}{d}}, 1-n^{-\frac{1-\delta}{d}}]^d$.



Problem settings

- Select the network from a certain class so that it minimizes the empirical loss

$$\underset{s \in \mathcal{S}:\ \text{DNNs}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left[\int_t \mathbb{E}_{X_t \sim p_t(X_t|X_0=x_i)}[\|s(X_t,t) - \nabla \log p_t(X_t|X_0=x_i)\|^2]\mathrm{d}t\right]$$

$x_1, \cdots, x_n \overset{\text{i.i.d}}{\sim} p_0$ empirical score matching loss

  - Because $p_t(X_t|X_0=x_i) = \mathcal{N}(e^{-t}x_i, 1-e^{-2t})$, the minimizer can be computed only with n finite sample
  - This is equivalent to usual squared loss minimization + weight func.

$$\min_{s \in \text{DNN}} \frac{1}{n}\sum_{i=1}^{n} \lambda(t_j)\|s(x_{t_i,i}, t_i) - \nabla \log p_{t_i}(x_{t_i,i}|x_i)\|$$

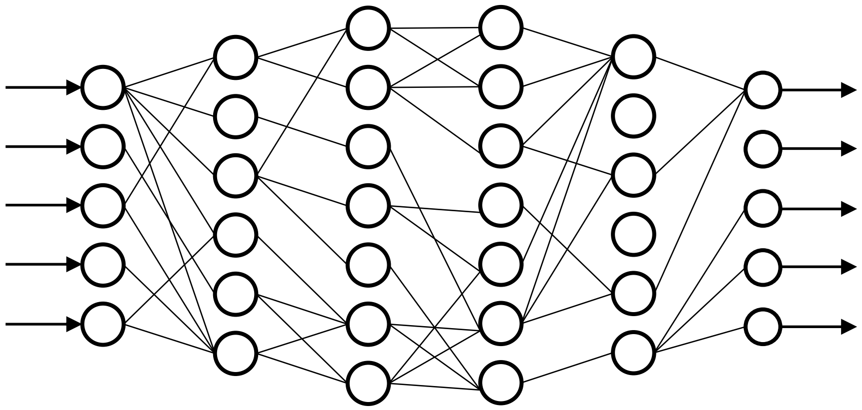
Problem settings

- Hypothesis network class: sparsity-constrainted deep ReLU networks

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

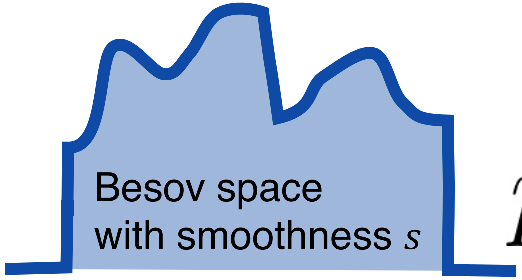
$$:= \left\{ (A^L \mathrm{ReLU}(\cdot) + b^L) \circ \cdots \circ (A^1 x + b^1) \,\middle|\, A^i \in \mathbb{R}^{w_i \times w_{i+1}}, b^i \in \mathbb{R}^{w_{i+1}}, \|w\|_\infty \le W, \sum_{i=1}^{L} (\|A^i\|_0 + \|b^i\|_0) \le S, \max \|A^i\|_\infty \vee \|b^i\|_\infty \le B \right\}$$

(Schmidt-Hieber, 2020; Suzuki, 2019)



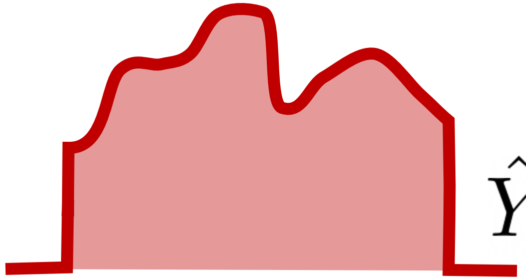
J. Schmidt-Hieber. "Nonparametric regression using deep neural networks with Relu activation function." *The Annals of Statistics*, 48(4):1875–1897, 2020.
T. Suzuki. "Adaptivity of deep Relu network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality". *ICLR*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

Problem settings



p_0

Evaluate the discrepancy



$\hat{Y}_{\overline{T}}$

Sample n data

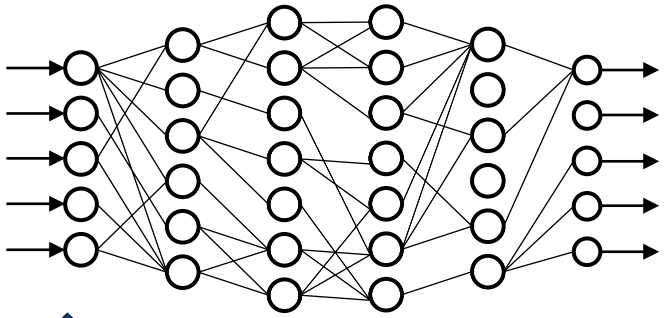
$x_1, \cdots, x_n \overset{\text{i.i.d}}{\sim} p_0$

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

Generate new data

Minimize the empirical score matching loss
over a certain class of DNNs

$$\underset{s \in \mathcal{S}:\ \text{DNNs}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left[\int_t \mathbb{E}_{X_t \sim p_t(X_t|X_0=x_i)}[\|s(X_t,t) - \nabla \log p_t(X_t|X_0=x_i)\|^2]\mathrm{d}t\right]$$



Main result ①: minimax optimality in TV

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

$$\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\mathrm{TV}(\hat{Y}_{\overline{T}}, X_0)\right] \lesssim n^{-\frac{s}{2s+d}} \log^8 n$$

under an appropriate choice of $\overline{T}, L, W, S$ 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.

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

$$\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\mathrm{TV}(\hat{Y}_{\overline{T}}, X_0)\right] \lesssim n^{-\frac{s}{2s+d}} \log^8 n$$

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

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

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

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

Basis decomposition tailored for score approximation

- B-spline basis decomposition of $p_0 (\in B^s_{p,q,C})$: $p_0(x) \approx \sum_{j=1}^{N} \alpha_j M^d_{a^j, b^j}(x)$
  (Devore & Popov, 1988)

  B-spline basis

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

Basis decomposition tailored for score approximation

- B-spline basis decomposition of $p_0(\in B^s_{p,q,C})$: $p_0(x) \approx \sum_{j=1}^{N} \alpha_j M^d_{a^j,b^j}(x)$
  (Devore & Popov, 1988)
  B-spline basis

- 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

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

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

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

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

$$=: E_{a^j,b^j}(x,t)$$ **diffused B-spline basis**

Approximated by NNs very efficiently (polylog size)

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

Basis decomposition tailored for score approximation

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

  (Devore & Popov, 1988) B-spline basis

- 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

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

$$=: E_{a^j, b^j}(x, t)$$ **diffused B-spline basis**

Approximated by NNs 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

- The exponent of $n^{-\frac{s}{2s+d}}$ depends on the dimension d "curse of dimensionality"

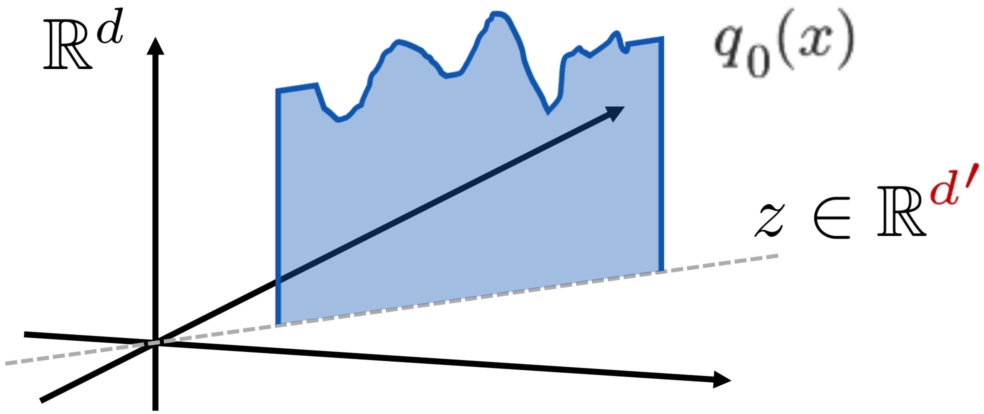
J. B. Tenenbaum, V. D. Silva, & J. C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction", *Science*, 290(5500):2319–2323, 2000.

Main result ②: manifold hypothesis

$\hat{s}$

- The exponent of $n^{-\frac{s}{2s+d}}$ depends on the dimension d "curse of dimensionality"
- Real-world data has intrinsic low-dimensionality (e.g., Tenenbaum et al., 2000)

Assume that p_0 lies on a d'-dimensional plane $(d' \leq d)$



- Density function q_0 on the canonical coordinate system on the plane belongs to $B^s_{p,q,C}$

J. B. Tenenbaum, V. D. Silva, & J. C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction", *Science*, 290(5500):2319–2323, 2000.

Main result ②: manifold hypothesis

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)

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

Main result ②: manifold hypothesis

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**
- Key idea: decomposition of the score

$$\nabla \log p_t(x) = \underline{\nabla \log q_t(A^\top x)} - \underline{\frac{1}{\sigma_t^2}(I-A)(I-A^\top)x}$$

Diffusion on the manifold $A^\top$: projection

Main result ②: manifold hypothesis

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**
- Key idea: decomposition of the score

$$\nabla \log p_t(x) = \underline{\nabla \log q_t(A^\top x)} - \underline{\frac{1}{\sigma_t^2}(I-A)(I-A^\top)x}$$

Diffusion on the manifold $\quad A^\top$: projection

- Even when $d' = d$, the rate in W1 is faster than that in TV($n^{-\frac{s}{2s+d}}$)

➡ additional techniques are required

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

Summary

- Revealed the power of diffusion modeling as a distribution estimator
  - the true distribution belongs to $B^s_{p,q,C}$ (s: smoothness)
  - and the score network minimize the empirical loss over a certain class of DNNs

Summary

- Revealed the power of diffusion modeling as a distribution estimator
  - the true distribution belongs to $B^s_{p,q,C}$ (s: smoothness)
  - and the score network minimize the empirical loss over a certain class of DNNs

- Proved that diffusion models can achieve the minimax optimal estimation rates
  - TV distance: $n^{-\frac{s}{2s+d}}$
    - Diffused B-spline basis decomposition
  - W1 distance: $n^{-\frac{s+1}{2s+d'}}$
    - Analysis under the manifold hypothesis
    - Avoid the curse of dimensionality