Statistical analysis of generative diffusion models

Nikita Puchkin

HSE University

Diffusion-based generative models

General idea [Sohl-Dickstein et al., 2015]:

- use a diffusion process to convert any complex distribution into a simple one
- learn a finite-time reversal of this diffusion process

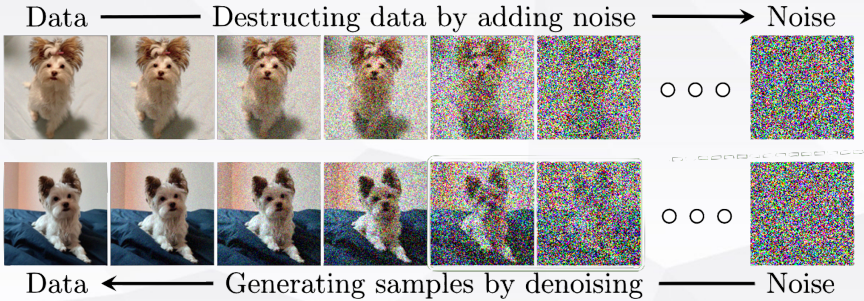

Figure: generative modelling via diffusions. Image courtesy: L. Yang et al. “Diffusion Models: A Comprehensive Survey of Methods and Applications”, ACM Computing Surveys, 2024

Motivation: application to score-based generative models

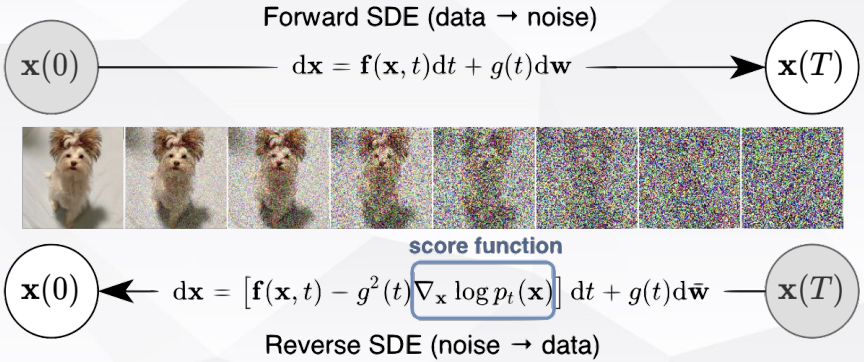

Figure: generative modelling via diffusions. Image courtesy: Y. Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”, ICLR, 2021

Example: DDPM SDE [Song et al., 2020]

$$\text{forward:} \quad \mathrm{d}X_t = -\frac{\beta_t X_t \,\mathrm{d}t}{2} + \sqrt{\beta_t}\,\mathrm{d}W_t, \quad 0 \leqslant t \leqslant T$$

$$\text{reverse:} \quad \mathrm{d}Z_t = \left(\frac{\beta_t Z_t}{2} + \beta_t \nabla \log \mathrm{p}_{T-t}(Z_t)\right)\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}B_t, \quad 0 \leqslant t \leqslant T$$

W_t and B_t are independent Wiener processes

Remark: in what follows, we consider $\beta_t = 2$

Convergence of diffusion models

Fix $0 = t_0 < t_1 < \ldots < t_N \leqslant T$ and consider

$$\mathrm{d}\widehat{Z}_t = \left(\widehat{Z}_t + 2\,\widehat{s}\,(\widehat{Z}_{t_k}, T - t_k)\right)\,\mathrm{d}t + \mathrm{d}B_t, \quad t_k \leqslant t \leqslant t_{k+1}, 1 \leqslant k \leqslant N$$

Introduce $\gamma_k = t_{k+1} - t_k$ and denote

$$\varepsilon^2_{\text{score}} = \sum_{k=0}^{N-1} \gamma_k\, \mathbb{E}_{X_{t_k} \sim \mathrm{p}_{t_k}} \left\| \nabla \log \mathrm{p}_{t_k}(X_{t_k}) - \widehat{s}\,(X_{t_k}, t_k) \right\|^2$$

Theorem: [Benton et al., 2024]

Assume that $\mathrm{Cov}(X_0) = I_D$ and that $\gamma_k \leqslant \kappa\big(1 \wedge (T - t_{k+1})\big)$ for all $k \in \{0, \ldots, N-1\}$. Then

$$\mathrm{KL}\left(\mathrm{p}_{T-t_N} \,\middle\|\, \mathrm{p}_{\widehat{Z}_{t_N}}\right) \lesssim \varepsilon^2_{\text{score}} + \kappa^2 DN + \kappa DT + De^{-2T}$$

Theorem: [Benton et al., 2024]

Assume that $\mathrm{Cov}(X_0) = I_D$ and that $\gamma_k \leqslant \kappa\big(1 \wedge (T - t_{k+1})\big)$ for all $k \in \{0, \dots, N-1\}$. Then

$$\mathrm{KL}\left(\mathrm{p}_{T-t_N} \,\middle\|\, \mathrm{p}_{\hat{Z}_{t_N}}\right) \lesssim \varepsilon^2_{\text{score}} + \kappa^2 DN + \kappa DT + De^{-2T}$$

Question: how to estimate $s^*(x,t) = \nabla \log \mathrm{p}_t(x)$?

In practice, one fixes a parametric class $\mathcal{S} = \{s_\theta : \theta \in \Theta\}$ (e. g., neural networks) and estimates $s^*(x,t)$ using i.i.d. samples drawn from p_t

Implicit score matching

Use integration by parts:

$$
\begin{aligned}
&\mathbb{E}_{X_t \sim \mathrm{p}_t} \left\| s_\theta(X_t, t) - \nabla \log \mathrm{p}_t(X_t, t) \right\|^2 \\
&= \mathbb{E}_{X \sim \mathrm{p}_t} \left\{ \left\| s_\theta(X_t, t) \right\|^2 - 2 \int s_\theta(x, t)^\top \nabla \mathrm{p}_t(x) \, \mathrm{d}x + \left\| \nabla \log \mathrm{p}_t(X_t) \right\|^2 \right\} \\
&= \mathbb{E}_{X_t \sim \mathrm{p}_t} \left\{ \left\| s_\theta(X_t, t) \right\|^2 + 2 \int \mathrm{div}\big(s_\theta(x, t)\big) \, \mathrm{p}_t(x) \, \mathrm{d}x + \left\| \nabla \log \mathrm{p}_t(X_t) \right\|^2 \right\} \\
&= \mathbb{E}_{X_t \sim \mathrm{p}_t} \left\{ \left\| s_\theta(X_t) \right\|^2 + 2 \mathrm{div}\big(s_\theta(X_t)\big) + \left\| \nabla \log \mathrm{p}_t(X_t) \right\|^2 \right\}
\end{aligned}
$$

Implicit score matching [Hyvärinen and Dayan, 2005]: $Y_1, \ldots, Y_n \sim \mathrm{p}_t^*$

$$
\frac{1}{n} \sum_{i=1}^{n} \ell_{ISM}\big(s_\theta(Y_i, t)\big) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} \left\| s_\theta(Y_i, t) \right\|^2 + \mathrm{div}\big(s_\theta(Y_i, t)\big) \right) \quad \to \quad \min_{\theta \in \Theta}
$$

Implicit score matching

- Implicit score matching works not so well in real life scenarios
- Implicit score matching does not take into account the joint distribution of X_0 and X_t
- Denoising score matching [Vincent, 2011] is more preferrable

Denoising score matching

OU Forward Process

$$\begin{cases} \mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, & 0 \leqslant t \leqslant T \\ X_0 \sim \mathrm{p}_0, & X_0 \in \mathbb{R}^D \end{cases}$$

OU Backward Process

$$\begin{cases} \mathrm{d}Z_t = (Z_t + 2\nabla \log \mathrm{p}_{T-t}(Z_t))\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \\ Z_0 \sim \mathrm{p}_T \end{cases}$$

Denoising Score Matching Objective: for a fixed *stopping time* $t_0 > 0$ we aim to minimize

$$\min_{s \in \mathcal{S}} \int_{t_0}^{T} \mathbb{E}_{X_t} \|s(X_t, t) - s^*(X_t, t)\|^2\,\mathrm{d}t, \qquad s^*(y, t) = \nabla \log \mathrm{p}_t(y, t).$$

Theorem: [Vincent, 2011]

Let $s^*(x,t) = \nabla \log p_t(x)$. Then for any $t \in [t_0, T]$ and $s(x,t)$ it holds that

$$\mathbb{E}_{X_t} \|s(X_t,t) - s^*(X_t,t)\|^2 = \mathbb{E}_{X_0} \mathbb{E}_{X_t} \left[\|s(X_t,t) - \nabla_{X_t} \log \mathrm{p}_t(X_t \mid X_0)\|^2 \mid X_0 \right] + C(t),$$

The proof is based on the observation

$$\begin{aligned}
\mathbb{E}_{X_0,X_t}\, s(X_t,t)^\top \nabla_{X_t} \log \mathrm{p}_t(X_t \mid X_0) &= \mathbb{E}_{X_0,X_t}\, s(X_t,t)^\top \nabla_{X_t} \log \mathrm{p}_t(X_t, X_0) \\
&= \int\int s(y,t)^\top \nabla_y \log \mathrm{p}_t(y,x)\, \mathrm{p}(y,x)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int\int s(y,t)^\top \nabla_y \mathrm{p}_t(y,x)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int s(y,t)^\top \nabla_y \mathrm{p}_t(y)\, \mathrm{d}y
\end{aligned}$$

Theorem: [Vincent, 2011]

Let $s^*(x,t) = \nabla \log p_t(x)$. Then for any $t \in [t_0, T]$ and $s(x,t)$ it holds that

$$\mathbb{E}_{X_t} \|s(X_t,t) - s^*(X_t,t)\|^2 = \mathbb{E}_{X_0} \mathbb{E}_{X_t} \left[\|s(X_t,t) - \nabla_{X_t} \log \mathrm{p}_t(X_t \mid X_0)\|^2 \mid X_0 \right] + C(t),$$

Denoising score matching loss:

$$\ell(s, X_0) = \int_{t_0}^{T} \mathbb{E}\left[\|s(X_t,t) - \nabla_{X_t} \log \mathrm{p}_t(X_t \mid X_0)\|^2 \mid X_0 \right] \mathrm{d}t, \quad \nabla \log \mathrm{p}_t(X_t|X_0) = -\frac{X_t - m_t X_0}{\sigma_t^2},$$

where $m_t = e^{-t}$ and $\sigma_t^2 = 1 - e^{-2t}$. Given $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathrm{p}_0$, DSM estimate is

$$\widehat{s} \in \operatorname*{argmin}_{s \in \mathcal{S}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(s, Y_i) \right\}$$

Brief literature review

- [Oko et al., 2023] – analysis of DSM estimate under the assumption that p_0 is bounded away from 0 and $+\infty$, the authors obtain a slow nonparametric rate of convergence depending on the ambient dimension
- Similar rates of convergence were obtained by [Wibisono et al., 2024] and [Zhang et al., 2024] who used kernel-based estimates
- [Chen et al., 2023] – first attempt to escape the curse of dimensionality, strong modelling assumptions ($\mathrm{supp}(p_0)$ is assumed to be a low-dimensional linear subspace)

Brief literature review

- [Tang and Yang, 2024] – analysis of DSM estimate under the manifold assumption ($\mathrm{supp}(\mathrm{p}_0)$ is a smooth low-dimensional manifold), the authors claim the rates of convergence $\mathcal{O}(n^{-2\beta/(2\beta+d)})$ but the hidden constant is proportional to e^D

- [Azangulov et al., 2024] – similar rates of convergence $\mathcal{O}(n^{-2\beta/(2\beta+d)})$ under the manifold assumtption. The authors eliminate dependence on D using smart preprocessing. However, their estimate faces problems when the data distribution slightly steps away from the manifold.

- [Yakovlev and Puchkin, 2025] – data distribution concentrates around a low-dimensional manifold, milder assumptions on p_0, similar rates of convergence $\mathcal{O}(n^{-2\beta/(2\beta+d)})$, the hidden constant grows polynomially with D

Brief literature review

- **Warning!** Theorem C.4 in [Oko et al., 2023] has a critical flow (see our paper [Yakovlev and Puchkin, 2025] for a detailed description). For this reason, the proof of the main result in [Oko et al., 2023] is incorrect
- [Tang and Yang, 2024] and [Azangulov et al., 2024] also use [Oko et al., 2023, Theorem C.4] in their proofs, so their main results are incorrect too
- One can obtain the results announced in [Oko et al., 2023, Tang and Yang, 2024, Azangulov et al., 2024] using our approach (see [Yakovlev and Puchkin, 2025, Theorem 3.6])

Assumptions on the data distribution

Assumption

Given a *generator* from the Hölder ball $g^* \in \mathcal{H}^\beta([0,1]^d, \mathbb{R}^D, H)$, $\|g^*\|_{L^\infty([0,1]^d)} \leqslant 1$, and $\sigma_{\text{data}} \in [0,1)$, the observed samples $Y_1, \dots, Y_n$ are i.i.d. copies of a random element $X_0 \in \mathbb{R}^D$ generated from the model

$$X_0 = g^*(U) + \sigma_{\text{data}}\, \xi,$$

where $U \sim \text{Un}([0,1]^d)$ and $\xi \sim \mathcal{N}(0, I_D)$ are independent

Under the assumption, the density along the forward process for any $t > 0$ is

$$\mathrm{p}_t(y) = (\sqrt{2\pi}\tilde{\sigma}_t)^{-D} \int\limits_{[0,1]^d} \exp\left\{ -\frac{\|y - m_t g^*(u)\|^2}{2\tilde{\sigma}_t^2} \right\} \mathrm{d}u, \quad \tilde{\sigma}_t^2 = m_t^2 \sigma_{\text{data}}^2 + \sigma_t^2$$

The class of score estimators

Proposition: representation of the true score function

Under Assumption 1 it holds that

$$s^*(y,t) = -\frac{y}{\tilde{\sigma}_t^2} + \frac{m_t}{\tilde{\sigma}_t^2}\left(\frac{\int\limits_{[0,1]^d} g^*(u)\exp\left(-\frac{\|y-m_t g^*(u)\|_2^2}{2\tilde{\sigma}_t^2}\right)\mathrm{d}u}{\int\limits_{[0,1]^d}\exp\left(-\frac{\|y-m_t g^*(u)\|_2^2}{2\tilde{\sigma}_t^2}\right)\mathrm{d}u}\right), \quad \tilde{\sigma}_t^2 = m_t^2\sigma_{\text{data}}^2 + \sigma_t^2$$

The class of score estimators is a ReLU-network with a tailored architecture

$$\mathcal{S}(L,W,S,B) = \left\{ s(y,t) := -\frac{y}{m_t^2\sigma^2+\sigma_t^2} + \frac{m_t \mathrm{clip}_2(f(y,t))}{m_t^2\sigma^2+\sigma_t^2} : f \in \mathrm{NN}(L,W,S,B),\ \sigma \in [0,1) \right\},$$

where L is the number of layers, $\|W\|_\infty$ is referred to as a width, S is the number of non-zero parameters, and B is weight magnitude.

Theorem: score approximation

Grant Assumption. Fix an arbitrary $\varepsilon \in (0,1)$ such that

$$D\varepsilon\sqrt{\log(1/\varepsilon)} \leqslant \tilde{\sigma}_{t_0}^2, \quad H\varepsilon\sqrt{D} \leqslant \tilde{\sigma}_{t_0}, \quad \text{and} \quad \frac{Hd^{\lfloor\beta\rfloor}\varepsilon^{\beta}\sqrt{D}}{\lfloor\beta\rfloor!} \leqslant 1 \wedge \tilde{\sigma}_{t_0}.$$

Then there exists a score function $s \in \mathcal{S}(L, W, S, B)$ with

$$L \lesssim D^2 \log^4(1/\varepsilon), \quad \|W\|_\infty \lesssim D^5 \varepsilon^{-d} \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1\right) \log^6(1/\varepsilon),$$

$$S \lesssim D^{6+2\binom{d+\lfloor\beta\rfloor+d}{d}} \varepsilon^{-d} \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1\right) (\log(1/\varepsilon))^{10+4\binom{d+\lfloor\beta\rfloor+d}{d}}, \quad \log B \lesssim D + \log^2(1/\varepsilon),$$

such that $\int_{t_0}^{T} \mathbb{E}_{X_t} \|s^*(X_t, t) - s(X_t, t)\|^2 \, \mathrm{d}t \lesssim \frac{D\varepsilon^{2\beta}}{\tilde{\sigma}_{t_0}^2}$

Step 1: local polynomial approximation. Let us introduce $N = \lceil 1/\varepsilon \rceil$, and for any $\mathbf{j} = (j_1, \dots, j_d) \in \{1, \dots, N\}^d$ we define

$$u_{\mathbf{j}} = \frac{\mathbf{j}}{N} \quad \text{and} \quad \mathcal{U}_{\mathbf{j}} = \left[\frac{j_1 - 1}{N}, \frac{j_1}{N}\right] \times \left[\frac{j_2 - 1}{N}, \frac{j_2}{N}\right] \times \cdots \times \left[\frac{j_d - 1}{N}, \frac{j_d}{N}\right].$$

$$g^{\circ}_{\mathbf{j}}(u) = \sum_{\substack{\mathbf{k} \in \mathbb{Z}^d_+ \\ |\mathbf{k}| \leqslant \lfloor \beta \rfloor}} \frac{\partial^{\mathbf{k}} g^*(u_{\mathbf{j}})}{\mathbf{k}!} (u - u_{\mathbf{j}})^{\mathbf{k}} \, \mathbb{1}[u \in \mathcal{U}_{\mathbf{j}}], \quad g^{\circ}(u) = \sum_{\mathbf{j} \in \{1, \dots, N\}^d} g^{\circ}_{\mathbf{j}}(u),$$

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \| s^{\circ}(X_t, t) - s^*(X_t, t) \|^2 \mathrm{d}t \leqslant \frac{D H^2 \varepsilon^{2\beta}}{4(\sigma^2_{\text{data}} + e^{2t_0} - 1)} \left(\frac{d^{\lfloor \beta \rfloor}}{\lfloor \beta \rfloor!} \right)^2$$

Step 2: reduction to approximation on a compact set.

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \|s^\circ(X_t,t) - s(X_t,t)\|^2 \mathrm{d}t \leqslant \int_{t_0}^{T} \frac{m_t^2}{\tilde{\sigma}_t^4} \left\{ D\varepsilon^{2\beta} + \mathbb{E}_{X_t} \left[\|f^\circ(X_t,t) - f(X_t,t)\|^2 \mathbb{1}[X_t \in \mathcal{K}_t] \right] \right\} \mathrm{d}t,$$

$$\mathcal{K}_t = \left\{ y \in \mathbb{R}^D \,:\, \min_{u \in [0,1]^d} \|y - m_t g^*(u)\| \leqslant R_t \right\}$$

$$R_t = \tilde{\sigma}_t \sqrt{D} + 16\tilde{\sigma}_t \left(\sqrt{D \log\left(\frac{\varepsilon^{-2\beta}}{D} \right)} \vee \log\left(\frac{\varepsilon^{-2\beta}}{D} \right) \right)$$

The problem of score approximation reduces to approximation of $f^\circ(y,t)$ on

$$\mathcal{C}^*_{[t_0,T]} = \left\{ (y,t) \in \mathbb{R}^D \times [t_0,T] : y \in \mathcal{K}_t \right\}.$$

Step 3: f° is a composition of simpler functions.

$$f^\circ(y,t) = \frac{\sum\limits_{\mathbf{j}\in\{1,\dots,N\}^d} \int\limits_{\mathcal{U}_\mathbf{j}} g^\circ_\mathbf{j}(u) \exp\left\{-\frac{\|y - m_t g^\circ_\mathbf{j}(u)\|^2}{2\tilde\sigma_t^2}\right\} \mathrm{d}u}{\sum\limits_{\mathbf{j}\in\{1,\dots,N\}^d} \int\limits_{\mathcal{U}_\mathbf{j}} \exp\left\{-\frac{\|y - m_t g^\circ_\mathbf{j}(u)\|^2}{2\tilde\sigma_t^2}\right\} \mathrm{d}u}.$$

Introduce an auxiliary vector-valued function

$$\frac{\|y - m_t g^\circ_\mathbf{j}(u)\|^2}{2\tilde\sigma_t^2} = V_{\mathbf{j},0}(y,t) + \mathcal{V}(t)\left\|g^\circ_\mathbf{j}(u) - g^*(u_\mathbf{j})\right\|^2 + \sum_{\substack{\mathbf{k}\in\mathbb{Z}^d_+ \\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} V_{\mathbf{j},\mathbf{k}}(y,t)\,\frac{(u-u_\mathbf{j})^\mathbf{k}}{\mathbf{k}!}.$$

$$\mathcal{V}_\mathbf{j}(y,t) = \left(\left(V_{\mathbf{j},\mathbf{k}} : \mathbf{k}\in\mathbb{Z}^d_+, 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor\right), \mathcal{V}(t)\right)^\top \in \mathbb{R}^{\binom{d+\lfloor\beta\rfloor}{d}}.$$

Step 4: approximation of $V_{\mathbf{j},0}$ and $\mathcal{V}_{\mathbf{j}}$. We represent $V_{\mathbf{j},0}(y,t)$ in the following form:

$$V_{\mathbf{j},0}(y,t) = \frac{\|y - m_t g^{\circ}_{\mathbf{j}}(u_{\mathbf{j}})\|^2}{2\tilde{\sigma}_t^2} = \frac{\|y\|^2}{2\tilde{\sigma}_t^2} - \frac{m_t y^{\top} g^{\circ}_{\mathbf{j}}(u_{\mathbf{j}})}{\tilde{\sigma}_t^2} + \frac{m_t^2 \|g^{\circ}_{\mathbf{j}}(u_{\mathbf{j}})\|^2}{2\tilde{\sigma}_t^2}.$$

Lemma: approximation of $V_{\mathbf{j},0}$

For any $\varepsilon' \in (0,1]$ there exist ReLU-networks $\tilde{V}_{\mathbf{j},0}, \tilde{V}_{\mathbf{j},\mathbf{k}}, \tilde{\mathcal{V}} \in \mathsf{NN}(\tilde{L}, \tilde{W}, \tilde{S}, \tilde{B})$ with

$$\left\|\tilde{V}_{\mathbf{j},0} - V_{\mathbf{j},0}\right\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \vee \left\|\tilde{V}_{\mathbf{j},\mathbf{k}} - V_{\mathbf{j},\mathbf{k}}\right\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \vee \left\|\tilde{\mathcal{V}} - \mathcal{V}\right\|_{L^\infty([t_0,T])} \leqslant \varepsilon',$$

$$\tilde{L} \vee \log \tilde{B} \lesssim \log^2(1/\varepsilon') + \log^2(\tilde{\sigma}_{t_0}^{-2}) + \log^2 D,$$

$$\|\tilde{W}\|_\infty \lesssim D\left(\frac{1}{t_0 + \sigma^2_{\text{data}}} \vee 1\right)\left(\log^2(1/\varepsilon') + \log^2(\tilde{\sigma}_{t_0}^{-2}) + \log^2 D\right),$$

$$\tilde{S} \lesssim D\left(\frac{1}{t_0 + \sigma^2_{\text{data}}} \vee 1\right)\left(\log^3(1/\varepsilon') + \log^3(\tilde{\sigma}_{t_0}^{-2}) + \log^3 D\right)$$

Step 5: approximation of the integrals over $\mathcal{U}_{\mathbf{j}}$.
We use the following representation:

$$\frac{\|y - m_t g_{\mathbf{j}}^{\circ}(u_{\mathbf{j}} - \varepsilon w)\|^2}{2\tilde{\sigma}_t^2} = V_{\mathbf{j},0}(y,t) + \mathscr{R}_{\mathbf{j}}(y,t)^{\top} a_{\mathbf{j}}(w) + b_{\mathbf{j}}(w),$$

$$\mathscr{R}_{\mathbf{j}}(y,t) = \left(\left(\frac{V_{\mathbf{j},\mathbf{k}}(y,t)}{2\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}} + \frac{1}{2} : \mathbf{k} \in \mathbb{Z}^d_+, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor \right), \frac{\mathcal{V}(t)}{\|\mathcal{V}\|_{L^\infty([t_0,T])}} \right)^{\top},$$

$$a_{\mathbf{j}}(w) = \left(\left(a_{\mathbf{j},\mathbf{k}}(w) : \mathbf{k} \in \mathbb{Z}^d_+, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor \right), \|\mathcal{V}\|_{L^\infty([t_0,T])} \left\| g_{\mathbf{j}}^{\circ}(u_{\mathbf{j}} - \varepsilon w) - g^*(u_{\mathbf{j}}) \right\|^2 \right),$$

$$a_{\mathbf{j},\mathbf{k}}(w) = \frac{2(-\varepsilon w)^{\mathbf{k}} \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}}{\mathbf{k}!}, \quad b_{\mathbf{j}}(w) = \sum_{\substack{\mathbf{k} \in \mathbb{Z}^d_+ \\ 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor}} \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \frac{(-\varepsilon w)^{\mathbf{k}}}{\mathbf{k}!}.$$

Lemma: approximation of the integral over $\mathcal{U}_{\mathbf{j}}$

Assume that $\frac{D\varepsilon\sqrt{\log(1/\varepsilon)}}{\tilde{\sigma}^2_{t_0}} \leqslant 1$. Fix a function $\psi_{\mathbf{j}} : [0,1]^d \to \mathbb{R}$ such that $\|\psi_{\mathbf{j}}\|_{L^\infty([0,1]^d)} \leqslant 2$. Consider the integral $\Upsilon_{\mathbf{j}}(y,t) = e^{-V_{\mathbf{j},0}(y,t)} \int_{[0,1]^d} \psi_{\mathbf{j}}(w) \exp\left\{-\mathscr{R}_{\mathbf{j}}(y,t)^\top a(w) - b(w)\right\} \mathrm{d}w$. Then, there exists $\tilde{\Upsilon}_{\mathbf{j}}(y,t) \in \mathrm{NN}(L_\Upsilon, W_\Upsilon, S_\Upsilon, B_\Upsilon)$ with $\left\|\Upsilon_{\mathbf{j}} - \tilde{\Upsilon}_{\mathbf{j}}\right\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \lesssim \varepsilon'\varepsilon$ such that

$$L_\Upsilon \lesssim \log^2(1/\varepsilon') + \log^2(\tilde{\sigma}^{-2}_{t_0}) + \log^2 D + \log\frac{1}{\varepsilon\varepsilon'} \cdot \log^2\left(\log\frac{1}{\varepsilon\varepsilon'}\right),$$

$$\|W_\Upsilon\|_\infty \lesssim D\left(\frac{1}{t_0 + \sigma^2_{\mathrm{data}}} \vee 1\right)\left(\log^3(1/\varepsilon') + \log^3(\tilde{\sigma}^{-2}_{t_0}) + \log^3 D\right) \vee \left(\log\frac{1}{\varepsilon\varepsilon'}\right)^{\binom{d+\lfloor\beta\rfloor}{d}+1},$$

$$S_\Upsilon \lesssim D\left(\frac{1}{t_0 + \sigma^2_{\mathrm{data}}} \vee 1\right)\left(\log^3(1/\varepsilon') + \log^3(\tilde{\sigma}^{-2}_{t_0}) + \log^3 D\right) + \left(\log\frac{1}{\varepsilon\varepsilon'}\right)^{2\binom{d+\lfloor\beta\rfloor}{d}+5},$$

$$\log B_\Upsilon \lesssim \log^2(1/\varepsilon') + \log^2(\tilde{\sigma}^{-2}_{t_0}) + \log^2 D.$$

Proof Overview: Step 5 (III)

The lemma suggests that there exist $\mathcal{Q}, \mathcal{P}_l \in \mathsf{NN}(\breve{L}, \breve{W}, \breve{S}, \breve{B})$, $1 \leqslant l \leqslant D$ such that

$$\sup_{(y,t)\in\mathcal{C}^*_{[t_0,T]}} \left| \mathcal{Q}(y,t) - \sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} \exp\left\{ -\frac{\|y - m_t g^{\circ}_{\mathbf{j}}(u)\|^2}{2\tilde{\sigma}_t^2} \right\} \mathrm{d}u \right| \lesssim \varepsilon\varepsilon',$$

$$\sup_{(y,t)\in\mathcal{C}^*_{[t_0,T]}} \max_{1\leqslant l\leqslant D} \left| \mathcal{P}_l(y,t) - \sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} g^{\circ}_{\mathbf{j},l}(u) \exp\left\{ -\frac{\|y - m_t g^{\circ}_{\mathbf{j}}(u)\|^2}{2\tilde{\sigma}_t^2} \right\} \mathrm{d}u \right| \lesssim \varepsilon\varepsilon'.$$

Obviously, $\mathcal{Q}$ and $\mathcal{P}_l$, $1 \leqslant l \leqslant D$, have a depth $\breve{L} = L\Upsilon$, a width $\|\breve{W}\|_\infty = N^d \|W\Upsilon\|_\infty$, at most $\breve{S} = N^d S\Upsilon$ non-zero weights, and the weight magnitude $\breve{B} = B\Upsilon$.

Step 6: division approximation. According to the definition of $\mathcal{C}^*_{[t_0,T]}$, for any $t \in [t_0, T]$ and any $y \in \mathcal{K}_t$, there exist $\mathbf{j}^* \in \{1, \dots, N\}^d$ and $u^*_{\mathbf{j}} \in \mathcal{U}^*_{\mathbf{j}}$ such that $\frac{\|y - m_t g^*(u^*_{\mathbf{j}})\|^2}{\tilde{\sigma}_t^2} \leqslant R_t$. Hence,

$$\sum_{\mathbf{j} \in \{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} \exp\left\{ -\frac{\|y - m_t g_{\mathbf{j}}(u)\|^2}{2\tilde{\sigma}_t^2} \right\} \mathrm{d}u \geqslant \varepsilon^d e^{-4 - R_t^2/\tilde{\sigma}_t^2}$$

Lemma: approximation of the division operation

Given a positive integer $K \geqslant 4$. Then, for any $\varepsilon \in (0, 1]$, there exists a ReLU-network $\mathcal{R} \in \mathrm{NN}(L, W, S, B)$ such that

$$\left| \mathcal{R}(x', y') - \frac{x}{y} \right| \leqslant 2049(4K^2 \log^2 2 + \log^2(1/\varepsilon))\varepsilon,$$

for all $y \in [2^{-K}, 1]$, $|x| \leqslant y$, $|x - x'| \vee |y - y'| \leqslant 2^{-2K}\varepsilon$. The network has $L \lesssim K^2 + \log^2(1/\varepsilon)$, $\|W\|_\infty \lesssim K^3 + K\log^2(1/\varepsilon)$, $S \lesssim K^4 + K\log^3(1/\varepsilon)$, and $B \lesssim 2^{4K}\log^2(1/\varepsilon)$

Theorem: score estimation

Grant assumptions of score approximation theorem. Let $T \geqslant 1$ and the sample size be sufficiently large. Then for an empirical risk minimizer $\widehat{s} \in \mathcal{S}(L, W, S, B)$ with

$$L \lesssim D^2 \log^4 n, \quad \|W\|_\infty \lesssim \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1\right) D^5 (n\sigma_{t_0})^{\frac{d}{4\beta+d}} \log^6 n,$$

$$S \lesssim \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1\right) D^{6+2\binom{d+\lfloor\beta\rfloor}{d}} (n\sigma_{t_0})^{\frac{d}{4\beta+d}} (\log n)^{10+4\binom{d+\lfloor\beta\rfloor}{d}}, \quad \log B \lesssim D \log^2 n$$

and any $\delta \in (0, 1)$ with probability at least $1 - \delta$ we have the following generalization bound

$$\int_{t_0}^{T} \mathbb{E}_{X_t}[\|\widehat{s}(X_t, t) - s^*(X_t, t)\|^2] \mathrm{d}t = \tilde{\mathcal{O}}\left(\frac{D^{10+2\binom{d+\lfloor\beta\rfloor}{d}} T}{(t_0 + \sigma_{\text{data}}^2) \wedge \tilde{\sigma}_{t_0}^2} (n\sigma_{t_0})^{-\frac{2\beta}{4\beta+d}} \log^3(n/\delta)\right)$$

Step 1: oracle inequality. First, we introduce the truncated version of the loss function defined for some $R \geqslant 0$

$$\ell^{tr}(s,x) = \ell(s,x)\mathbb{1}[x \in \mathcal{K}_0], \quad \mathcal{K}_0 = \{y \in \mathbb{R}^D : \inf_{u\in[0,1]^d} \|y - g^*(u)\| \leqslant \sqrt{R^2 + \sigma^2_{\text{data}} D}\}.$$

For some $a \leqslant 1$ write down the *oracle inequality*

$$\mathbb{E}_{X_0}[\ell(\widehat{s}, X_0)] \leqslant \underbrace{\mathbb{E}_{X_0}[\ell^{tr}(\widehat{s}, X_0)] - (1+a)\widehat{\mathbb{E}}_{X_0}[\ell^{tr}(\widehat{s}, X_0)]}_{(A)} + \underbrace{\mathbb{E}_{X_0}[\ell(\widehat{s}, X_0) - \ell^{tr}(\widehat{s}, X_0)]}_{(B)}$$

$$+ (1+a) \underbrace{\inf_{s\in\mathcal{S}(L,W,S,B)} \widehat{\mathbb{E}}_{X_0}[\ell(s, X_0)]}_{(C)}.$$

Proof Overview: Steps 2 and 3

Step 2: an upper bound for term (B). The evaluation of the term (B) is straightforward

$$(B) = \mathbb{E}_{X_0}[\ell(\widehat{s}, X_0) - \ell^{tr}(\widehat{s}, X_0)] \lesssim \frac{DT \log(\sigma_{t_0}^{-2})}{\sigma_{t_0}^2} \exp\left(-\frac{1}{32}\left(\frac{R^2}{D\sigma_{\text{data}}^2} \wedge \frac{R}{\sigma_{\text{data}}}\right)\right)$$

Step 3: a high-probability bound for term (A)

Lemma: [Chen et al., 2023]

Let $\mathcal{G}$ be a bounded function class, i.e. there exists $B \geqslant 0$ such that any $g \in \mathcal{G} : \mathbb{R}^D \to [0, B]$. Let also $z_1, \ldots, z_n \in \mathbb{R}^D$ be i.i.d. random variables. Then, for any $\delta \in (0,1)$, $a \leqslant 1$, and $\tau > 0$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} g(z_i) - (1+a)\mathbb{E}[g(z)] > \frac{(1+3/a)B}{3n} \log\left(\frac{\mathcal{N}(\tau, \mathcal{G}, \|.\|_\infty)}{\delta}\right) + (2+a)\tau\right) \leqslant \delta,$$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \mathbb{E}[g(z)] - \frac{1+a}{n}\sum_{i=1}^{n} g(z_i) > \frac{(1+6/a)B}{3n} \log\left(\frac{\mathcal{N}(\tau, \mathcal{G}, \|.\|_\infty)}{\delta}\right) + (2+a)\tau\right) \leqslant \delta.$$

Proof Overview: Steps 3 and 4

Using Chen's argument, we obtain that with probability at least $1 - \delta/3$

$$(A) - (2+a)\tau \lesssim \frac{(1+a^{-1})T(D+R^2)\log(\sigma_{t_0}^{-2})SL}{n\sigma_{t_0}^2} \log\left(\frac{D(1+R^2)L(W+1)T(B\vee 1)}{\delta\tau\sigma_{t_0}}\right).$$

Step 4: bounding term (C) We first provide an upper bound for the term (C)

$$(C) = \inf_{s\in\mathcal{S}} \widehat{E}_{X_0}[\ell(s, X_0)] \leqslant \underbrace{\widehat{E}_{X_0}[\ell(\bar{s}, X_0)] - (1+a)\mathbb{E}_{X_0}[\ell^{tr}(\bar{s}, X_0)]}_{(C_1)} + (1+a)\underbrace{\mathbb{E}_{X_0}[\ell^{tr}(\bar{s}, X_0)]}_{(C_2)},$$

The approximation result indicates that $(C_2) + C(T - t_0) \lesssim \frac{D\varepsilon^{2\beta}}{\tilde{\sigma}_{t_0}^2}$. Next, note that with high probability $\widehat{\mathbb{E}}_{X_0}[\ell(\bar{s}, X_0)] = \widehat{\mathbb{E}}_{X_0}[\ell^{tr}(\bar{s}, X_0)]$, which suggests that with probability at least $1 - 2\delta/3$

$$(C_1) - (2+a)\tau \lesssim \frac{T(1+a^{-1})(D+R^2)\log(1/\delta)}{n\sigma_{t_0}^2}.$$

Proof Overview: Steps 4 and 5

Combining the bounds for (C_1) and (C_2), we deduce that with probability at least $1 - 2\delta/3$

$$\begin{aligned}(C) + C(T - t_0) &\leqslant (C_1) + (C_2) + C(t - t_0) + a(C_2) \\ &\lesssim \frac{T(1 + a^{-1})(D + R^2)\log(1/\delta)}{n\sigma_{t_0}^2} + \frac{D\varepsilon^{2\beta}}{\tilde{\sigma}_{t_0}^2} + a\left(T + \frac{D\varepsilon^{2\beta}}{\tilde{\sigma}_{t_0}^2}\right) + \tau.\end{aligned}$$

Step 5: deriving the final bound Optimizing the introduced parameters, we have that with probability at least $1 - \delta$

$$\int_{t_0}^{T} \mathbb{E}_{X_t}[\|\hat{s}(X_t, t) - s^*(X_t, t)\|^2]\mathrm{d}t = \tilde{\mathcal{O}}\left(\frac{D^{10+2\binom{d+\lfloor\beta\rfloor}{d}}T}{(t_0 + \sigma_{\mathrm{data}}^2) \wedge \tilde{\sigma}_{t_0}^2}(n\sigma_{t_0})^{-\frac{2\beta}{4\beta+d}}\log^3(n/\delta)\right).$$

Generalization bound: improvement for the case $\sigma_{\text{data}} = 0$

Lemma

Let $h : \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^D$ and $h' : \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^D$ be any Borel functions such that

$$\|h\|_{L^\infty(\mathbb{R}^D \times [t_0,T])} \leqslant 2 \quad \text{and} \quad \|h'\|_{L^\infty(\mathbb{R}^D \times [t_0,T])} \leqslant 2.$$

Consider the corresponding scores

$$s(y,t) = -\frac{y}{\sigma_t^2} + \frac{m_t}{\sigma_t^2} h(y,t) \quad \text{and} \quad s'(y,t) = -\frac{y}{\sigma_t^2} + \frac{m_t}{\sigma_t^2} h'(y,t).$$

Then, for all $x \in \mathrm{Im}(g^*)$, it holds that

$$\big(\ell(s,x) - \ell(s',x)\big)^2 \leqslant \frac{36}{\sigma_{t_0}^2} \int_{t_0}^{T} \mathbb{E}_{X_t | X_0 = x} \big\| s(X_t,t) - s'(X_t,t) \big\|^2 \, \mathrm{d}t.$$

Generalization bound: improvement for the case $\sigma_{\text{data}} = 0$

Theorem

Assume all the conditions of the theorem on score approximation. Let also $T \geqslant 1$, $t_0 \leqslant 1$ and let the sample size be sufficiently large, that is, it fulfils

$$\sigma_{t_0}^2 n \geqslant \left(\frac{D\sqrt{\log(n\sigma_{t_0}^2)}}{\sigma_{t_0}^2\sqrt{2\beta+d}}\right)^{2\beta+d} \vee \left(\frac{Hd^{\lfloor\beta\rfloor}\sqrt{D}}{\lfloor\beta\rfloor!\,\sigma_{t_0}}\right)^{2+d/\beta} \vee \left(\frac{H\sqrt{D}}{\sigma_{t_0}}\right)^{2\beta+d}$$

Then, for any $\delta \in (0,1)$, with probability at least $(1-\delta)$ the excess risk of an empirical risk minimizer over the class $\mathcal{S}_0(L, W, S, B)$ satisfies the inequality

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \|\widehat{s}(X_t, t) - s^*(X_t, t)\|^2 \mathrm{d}t \lesssim \frac{D^{9+2\binom{d+\lfloor\beta\rfloor}{d}}}{\sigma_{t_0}^2} (n\sigma_{t_0}^2)^{-\frac{2\beta}{2\beta+d}} \log(2/\delta) L(t_0, n),$$

where $L(t_0, n) = (\log n)^{17+4\binom{d+\lfloor\beta\rfloor}{d}} \log T \log D \log(1/t_0)$

References I

[Azangulov et al., 2024] Azangulov, I., Deligiannidis, G., and Rousseau, J. (2024).
Convergence of diffusion models under the manifold hypothesis in high-dimensions.
arXiv preprint arXiv:2409.18804.

[Benton et al., 2024] Benton, J., Bortoli, V. D., Doucet, A., and Deligiannidis, G. (2024).
Nearly d-linear convergence bounds for diffusion models via stochastic localization.
In *The Twelfth International Conference on Learning Representations*.

[Chen et al., 2023] Chen, M., Huang, K., Zhao, T., and Wang, M. (2023).
Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data.
In *International Conference on Machine Learning*, pages 4672–4712. PMLR.

References II

[Hyvärinen and Dayan, 2005] Hyvärinen, A. and Dayan, P. (2005).
Estimation of non-normalized statistical models by score matching.
Journal of Machine Learning Research, 6(4).

[Oko et al., 2023] Oko, K., Akiyama, S., and Suzuki, T. (2023).
Diffusion models are minimax optimal distribution estimators.
In *International Conference on Machine Learning*, pages 26517–26582. PMLR.

[Sohl-Dickstein et al., 2015] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015).
Deep unsupervised learning using nonequilibrium thermodynamics.
In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR.

References III

[Song et al., 2020] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020).
Score-based generative modeling through stochastic differential equations.
arXiv preprint arXiv:2011.13456.

[Tang and Yang, 2024] Tang, R. and Yang, Y. (2024).
Adaptivity of diffusion models to manifold structures.
In *International Conference on Artificial Intelligence and Statistics*, pages 1648–1656. PMLR.

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

References IV

[Wibisono et al., 2024] Wibisono, A., Wu, Y., and Yang, K. Y. (2024).
Optimal score estimation via empirical bayes smoothing.
In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4958–4991. PMLR.

[Yakovlev and Puchkin, 2025] Yakovlev, K. and Puchkin, N. (2025).
Generalization error bound for denoising score matching under relaxed manifold assumption.
Preprint. ArXiv:2502.13662.

[Zhang et al., 2024] Zhang, K., Yin, H., Liang, F., and Liu, J. (2024).
Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions.
In *Forty-first International Conference on Machine Learning*.