Advancing Wasserstein Convergence Analysis of Score-Based Models: Insights from Discretization and Second-Order Acceleration

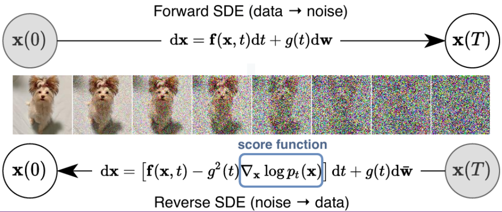
Yifeng Yu*, Lu Yu†

* Tsinghua University, † City University of Hong Kong

November 6th, 2025

Diffusion Model

- *Diffusion models* have become a pivotal framework in modern generative modeling, achieving notable success across fields such as image generation, natural language processing, and computational biology. These models add noise to data via a forward process and learn to reverse it, reconstructing data from noise.
- A widely adopted formulation of diffusion models is the score-based generative model (SGM), implemented using stochastic differential equations (SDEs) [SSDK+20].



Framework

Forward process:

$$\mathrm{d}X_t = f(X_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}B_t,$$

Backward process:

$$\mathrm{d}Y_t = [-f(Y_t, T-t) + g(T-t)^2 \nabla \log p_{T-t}(Y_t)]\,\mathrm{d}t + g(T-t)\,\mathrm{d}W_t.$$

For clarity, we adopt the simplest possible choice in this work by setting $f(X_t, t) = -X_t/2$ and $g(t) = 1$. This results in the Ornstein-Uhlenbeck process, which is described by the following SDE:

$$\mathrm{d}X_t = -\frac{1}{2}X_t\,\mathrm{d}t + \mathrm{d}B_t\,. \tag{1}$$

Then, diffusion models generate new data by reversing the SDE (1), which leads to the following backward SDE

$$\mathrm{d}X_t^{\leftarrow} = \frac{1}{2}\left(X_t^{\leftarrow} + 2\nabla \log p_{T-t}(X_t^{\leftarrow})\right)\,\mathrm{d}t + \mathrm{d}W_t\,, \tag{2}$$

Score Matching

$$\underset{\theta\in\Theta}{\text{minimize}} \quad \mathbb{E}[\|s_\theta(t, X_t) - \nabla \log p_t(X_t)\|^2],$$
$$\iff \underset{\theta\in\Theta}{\text{minimize}} \quad \mathbb{E}[\mathrm{tr}(\nabla s_\theta(x)) + \frac{1}{2}\|s_\theta(x)\|_2^2].$$

Denoising score matching. First perturbs the data point x with a pre-specified noise distribution $q_\sigma(\tilde{x}|x)$ and then employs score matching to estimate the score of the perturbed data distribution $q_\sigma(\tilde{x}) := \int q_\sigma(\tilde{x}|x)p_{\mathrm{data}}(x)\,\mathrm{d}x$.

$$\mathbb{E}_{q_\sigma(\tilde{x}|x)p_{\mathrm{data}}(x)}[\|s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)\|_2^2].$$

Sliced score matching. p_v is a simple distribution of random vectors, e.g., the multivariate standard normal.

$$\mathbb{E}_{p_v}\mathbb{E}_{p_{\mathrm{data}}}[v^\top \nabla s_\theta(x) v + \frac{1}{2}\|s_\theta(x)\|_2^2].$$

Discretization Schemes

$$X^{\leftarrow}_{t+h} = X^{\leftarrow}_{t} + \int_0^h \gamma(T-(t+v), X^{\leftarrow}_{t+v})\,\mathrm{d}v + \Delta_h W_t.$$

- **Euler-Maruyama scheme**:

$$\vartheta^{\mathrm{EM}}_{n+1} = (1+h/2)\vartheta^{\mathrm{EM}}_{n} + hs_*(T-nh, \vartheta^{\mathrm{EM}}_{n}) + \sqrt{h}\xi_n\,.$$

- **Exponential Integrator**:

$$\vartheta^{\mathrm{EI}}_{n+1} = e^{\frac{h}{2}}\vartheta^{\mathrm{EI}}_{n} + 2(e^{\frac{h}{2}}-1)s_*(T-nh, \vartheta^{\mathrm{EI}}_{n}) + \sqrt{e^h-1}\xi_n\,.$$

Randomized Midpoint Method:

$$\begin{aligned} X^{\leftarrow}_{t+h} &= X^{\leftarrow}_{t} + \int_0^h \gamma(T-t-v, X^{\leftarrow}_{t+v})\,\mathrm{d}v + \Delta_h W_t \\ &\approx X^{\leftarrow}_{t} + h\gamma(T-t-hU, X^{\leftarrow}_{t+hU}) + \Delta_h W_t\,. \end{aligned}$$

- **Vanilla Midpoint Randomization**:

Step 1 $\xi_n', \xi_n'' \sim \mathcal{N}(\mathbf{0}, I_d)$, $U_n \sim U[0,1]$. $\xi_n = \sqrt{U_n}\xi_n' + \sqrt{1-U_n}\xi_n''$.
Step 2 With the initialization $\vartheta_0^{\mathsf{REM}} \sim \hat{p}_T$, define

$$\vartheta_{n+U}^{\mathsf{REM}} = \vartheta_n^{\mathsf{REM}} + hU_n\gamma(T - nh, \vartheta_n^{\mathsf{REM}}) + \sqrt{hU_n}\xi_n',$$
$$\vartheta_{n+1}^{\mathsf{REM}} = \vartheta_n^{\mathsf{REM}} + h\gamma(T - (n+U_n)h, \vartheta_{n+U}^{\mathsf{REM}}) + \sqrt{h}\xi_n.$$

- **Exponential Integrator with Midpoint Randomization:**

Step 1 $\xi_n', \xi_n'' \sim \mathcal{N}(\mathbf{0}, I_d)$, $U_n \sim U[0,1]$. $\xi_n = \rho_n\xi_n' + \sqrt{1-\rho_n^2}\xi_n''$ with

$$\rho_n = e^{\frac{h(1+U_n)}{2}}\left(1 - e^{-hU_n}\right)\left[(e^{hU_n} - 1)(e^h - 1)\right]^{-1/2}.$$

Step 2 With the initialization $\vartheta_0^{\mathsf{REI}} \sim \hat{p}_T$, define

$$\vartheta_{n+U}^{\mathsf{REI}} = e^{hU_n/2}\vartheta_n^{\mathsf{REI}} + 2(e^{hU_n/2} - 1)s_*(T - nh, \vartheta_n^{\mathsf{REI}}) + \sqrt{e^{hU_n} - 1}\xi_n',$$
$$\vartheta_{n+1}^{\mathsf{REI}} = e^{h/2}\vartheta_n^{\mathsf{REI}} + he^{(1-U_n)h/2}s_*(T - (n+U_n)h, \vartheta_{n+U}^{\mathsf{REI}}) + \sqrt{e^h - 1}\xi_n.$$

Wasserstein Convergence Analysis

Assumption 1

The target density p_0 is m_0-strongly log-concave, and the score function $\nabla \log p_0$ is L_0-Lipschitz.

Assumption 2

There exists a constant $M_1 > 0$ such that for $n = 0, 1, \ldots, N-1$,

$$\sup_{nh \leqslant t, s \leqslant (n+1)h} \|\nabla \log p_{T-t}(x) - \nabla \log p_{T-s}(x)\| \leqslant M_1 h(1 + \|x\|), \quad \forall x\,.$$

Assumption 3

Given a small $\varepsilon_{sc} > 0$, the score estimator satisfies

$$\sup_{0 \leqslant n \leqslant N} \|\nabla \log p_{T-nh}(\vartheta_n) - s_*(T - nh, \vartheta_n)\|_{\mathbb{L}_2} \leqslant \varepsilon_{sc}\,.$$

Theorem 1 (EM, EI)

Suppose that Assumptions 1, 2 and 3 hold, it holds that

$$\mathrm{W}_2(\mathcal{L}(\vartheta_N^\alpha), p_0) \lesssim e^{-m_{\min} T}\|X_0\|_{\mathbb{L}_2} + \mathscr{C}_1^\alpha \sqrt{dh} + \mathscr{C}_2^\alpha \varepsilon_{sc},$$

where $\mathscr{C}_1^{\mathrm{EM}} = \frac{L_{\max}+1/2}{m_{\min}-1/2}$, $\mathscr{C}_2^{\mathrm{EM}} = \frac{1}{m_{\min}-1/2}$, *and* $\mathscr{C}_1^{\mathrm{EI}} = \frac{L_{\max}}{m_{\min}-1/2}$, $\mathscr{C}_2^{\mathrm{EI}} = \frac{1}{m_{\min}-1/2}$, *with* $m_{\min} = \min(1, m_0)$ *and* $L_{\max} = 1 + L_0$.

Corollary 2

Given a small $\varepsilon > 0$ *and* $\varepsilon_{sc} = \mathcal{O}(\varepsilon)$, *the Wasserstein distance satisfies* $\mathrm{W}_2(\mathcal{L}(\vartheta_N^\alpha), p_0) < \varepsilon, \alpha \in \{\mathrm{EM}, \mathrm{EI}\}$ *after* $N = \mathcal{O}\big(\frac{d}{\varepsilon^2}\log\big(\frac{\sqrt{d}}{\varepsilon}\big)\big)$ *iterations, provided that* $T = \mathcal{O}\left(\log(\frac{\sqrt{d}}{\varepsilon})\right)$ *and* $h = \mathcal{O}\left(\frac{\varepsilon^2}{d}\right)$.

Assumption 4

There exists a constant $\varepsilon_{sc} > 0$ *such that for any* $u \in [0,1]$ *and* $n = 0, \ldots, N$,

$$\|\nabla \log p_{T-(n+u)h}(\vartheta^{\alpha}_{n+u}) - s_*(T-(n+u)h, \vartheta^{\alpha}_{n+u})\|_{\mathbb{L}_2} \leqslant \varepsilon_{sc}.$$

Theorem 3 (REM, REI)

Suppose that Assumptions 1, 2 and 4 hold, then for $\alpha \in \{\mathrm{REM}, \mathrm{REI}\}$,

$$\mathrm{W}_2(\mathcal{L}(\vartheta^{\alpha}_N), p_0) \lesssim e^{-m_{\min} T}\|X_0\|_{\mathbb{L}_2} + \mathscr{C}^{\alpha}_1(d)\sqrt{h} + \mathscr{C}^{\alpha}_2 \varepsilon_{sc}\,,$$

where $\mathscr{C}^{\mathrm{REM}}_1(d) = \frac{\sqrt{d/3}L_{\max} + 1/2\sqrt{3}}{m_{\min}-1/2}$, $\mathscr{C}^{\mathrm{REM}}_2 = \frac{3}{m_{\min}-1/2}$ *and* $\mathscr{C}^{\mathrm{REI}}_1(d) = \frac{\sqrt{d/3}L_{\max}}{(m_{\min}-1/2)}$, $\mathscr{C}^{\mathrm{REI}}_2 = \frac{3}{m_{\min}-1/2}$, *with* $L_{\max}$ *and* $m_{\min}$ *as defined in Theorem 1.*

Second-order Acceleration

$$\mathrm{d}x_t = \gamma(T-t, x_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t\,,$$

We assume that $\gamma(t,x) \in C^{1,3}(\mathbb{R}_+ \times \mathbb{R}^d)$ and approximate it by a linear function in both state and time within each discretization step. Applying Itô's formula to $\gamma(T-t,x)$, we derive the following approximation for $\gamma(T-t,x_t) - \gamma(T-s,x_s)$

$$\left[\frac{\sigma^2}{2}\frac{\partial^2\gamma}{\partial x^2}(T-s,x_s) - \frac{\partial\gamma}{\partial t}(T-s,x_s)\right](t-s) + \frac{\partial\gamma}{\partial x}(T-s,x_s)(x_t-x_s)\,.$$

This allows us to express $\gamma(T-t,x_t)$ in the following form

$$\gamma(T-t,x_t) \approx \gamma(T-s,x_s) + L_s(x_t-x_s) + M_s(t-s)\,,$$

with

$$L_s = \frac{\partial\gamma}{\partial x}(T-s,x_s), \quad M_s = \frac{\sigma^2}{2}\frac{\partial^2\gamma}{\partial x^2}(T-s,x_s) - \frac{\partial\gamma}{\partial t}(T-s,x_s)\,.$$

$$x_t = \vartheta_n^{\mathsf{SO}} + \int_{nh}^{t} \Big(\frac{1}{2}\vartheta_n^{\mathsf{SO}} + \nabla \log p_{T-nh}(\vartheta_n^{\mathsf{SO}}) + L_n(x_u - \vartheta_n^{\mathsf{SO}}) + M_n(u-nh) \Big)\, \mathrm{d}u + \int_{nh}^{t} \mathrm{d}W_u$$

where

$$L_n = \frac{1}{2} I_d + \nabla^2 \log p_{T-nh}(\vartheta_n^{\mathsf{SO}}),$$

$$M_n = \frac{1}{2}\sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} \nabla \log p_{T-nh}(\vartheta_n^{\mathsf{SO}}) - \frac{\partial}{\partial t} \nabla \log p_{T-nh}(\vartheta_n^{\mathsf{SO}})\,.$$

$$\begin{aligned}\vartheta_{n+1}^{\mathsf{SO}} &= \vartheta_n^{\mathsf{SO}} + s_*^{(L)}(T-nh,\vartheta_n^{\mathsf{SO}})^{-1}\left(e^{s_*^{(L)}(T-nh,\vartheta_n^{\mathsf{SO}})h} - I_d\right)\left(\frac{1}{2}\vartheta_n^{\mathsf{SO}} + s_*(T-nh,\vartheta_n^{\mathsf{SO}})\right)\\ &+ s_*^{(L)}(T-nh,\vartheta_n^{\mathsf{SO}})^{-2}\left(e^{s_*^{(L)}(T-nh,\vartheta_n^{\mathsf{SO}})h} - s_*^{(L)}(T-nh,\vartheta_n^{\mathsf{SO}})h - I_d\right) s_*^{(M)}(T-nh,\vartheta_n^{\mathsf{SO}})\\ &+ \int_{nh}^{(n+1)h} e^{s_*^{(L)}(T-nh,\vartheta_n^{\mathsf{SO}})[(n+1)h-t]}\, \mathrm{d}W_t\,.\end{aligned}$$

Assumption 5

For some constants $\varepsilon_{sc}^{(L)}, \varepsilon_{sc}^{(M)} > 0$, *the estimate for high-order derivatives of the score function satisfies that*

$$\sup_{0 \leqslant n \leqslant N-1} \| s_*^{(L)}(T - nh, \vartheta_n^{\mathrm{SO}}) - L_n \|_{\mathbb{L}_2} \leqslant \varepsilon_{sc}^{(L)},$$

$$\sup_{0 \leqslant n \leqslant N-1} \| s_*^{(M)}(T - nh, \vartheta_n^{\mathrm{SO}}) - M_n \|_{\mathbb{L}_2} \leqslant \varepsilon_{sc}^{(M)}.$$

Assumption 6

Let $\| \cdot \|_F$ *denote the Frobenius norm. There exists a positive constant* L_F *such that*

$$\left\| \nabla^2 \log p_t(x) - \nabla^2 \log p_t(y) \right\|_F \leqslant L_F \|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 7

There exists a constant $M_2 > 0$ such that, for any $n = 0, \ldots, N-1$ and $t \in [nh, (n+1)h]$, it holds that

$$\|\nabla^2 \log p_{T-t}(x) - \nabla^2 \log p_{T-nh}(x)\| \leqslant M_2 h(1 + \|x\|), \quad \forall x \in \mathbb{R}^d .$$

Theorem 4

Suppose that Assumptions 1, 3, 5, 6 and 7 hold, then

$$\begin{aligned} \mathrm{W}_2(\mathcal{L}(\vartheta_N^{\mathsf{SO}}), p_0) \lesssim{} & e^{-m_{\min} T} \|X_0\|_{\mathbb{L}_2} + \mathscr{C}_1^{\mathsf{SO}}(d) h \\ & + \mathscr{C}_2^{\mathsf{SO}} \left(\varepsilon_{sc} + \frac{2}{3}\sqrt{hd}\,\varepsilon_{sc}^{(L)} + \frac{1}{2} h \varepsilon_{sc}^{(M)} \right) \end{aligned}$$

where $\mathscr{C}_1^{\mathsf{SO}}(d) = e^{(L_{\max} - 1/2)h} \cdot \frac{\sqrt{d}(L_{\max}^{3/2} + \sqrt{2} L_F/4)}{m_{\min} - 1/2}$ *and* $\mathscr{C}_2^{\mathsf{SO}} = \frac{e^{(L_{\max}-1/2)h}}{m_{\min} - 1/2}$ *with* $L_{\max}$ *and* $m_{\min}$ *as defined in Theorem 1.*

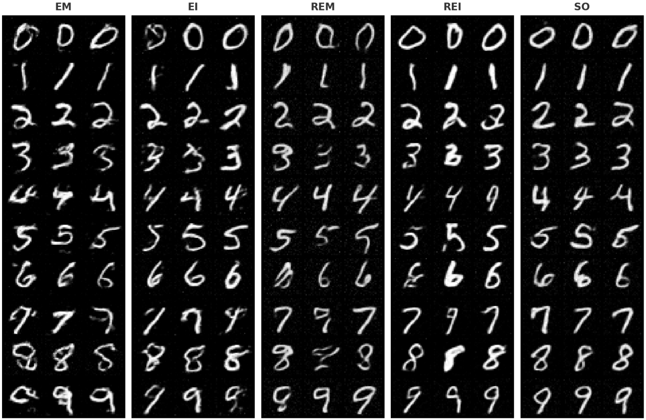
Corollary 5

For a given $\varepsilon > 0$, the Wasserstein distance satisfies $\mathsf{W}_2(\mathcal{L}(\vartheta_N^{\mathsf{SO}}), p_0) < \varepsilon$ *after* $N = \mathcal{O}\left(\frac{\sqrt{d}}{\varepsilon}\log\left(\frac{\sqrt{d}}{\varepsilon}\right)\right)$ *iterations, provided that* $T = \mathcal{O}\left(\log(\frac{\sqrt{d}}{\varepsilon})\right)$ *and* $h = \mathcal{O}(\frac{\varepsilon}{\sqrt{d}})$.

The accelerated convergence of this method is driven by two key innovations: approximating the drift term through its Itô expansion rather than endpoint evaluations, and deriving a closed-form solution to the integral equation using the Itô formula, akin to Exponential Integrator techniques.

Numerical Studies

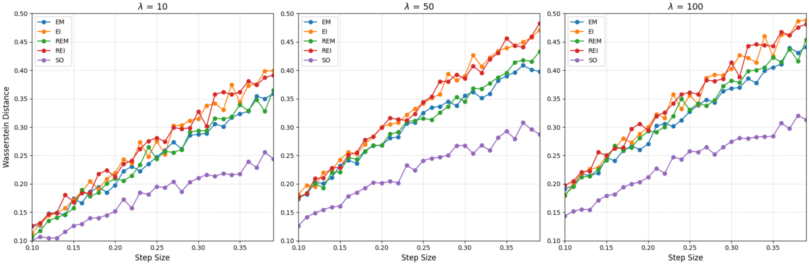
MNIST dataset: To accelerate the SO algorithm, we use Hessian-vector products (HVPs) instead of explicitly computing the Hessian.



We apply the five schemes to the posterior density of penalized logistic regression, defined by $p_0(\theta) \propto \exp(-f(\theta))$ with the potential function

$$f(\theta) = \frac{\lambda}{2}\|\theta\|^2 + \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} \log\left(1 + \exp(-y_i x_i^\top \theta)\right),$$

where $\lambda > 0$ denotes the tuning parameter.



[SSDK+20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Thanks!