Maximum Likelihood Training for Score-Based Diffusion ODEs by High-Order Denoising Score Matching

Cheng Lu, Kaiwen Zheng, Fan Bao, Chongxuan Li, Jianfei Chen, Jun Zhu

Tsinghua University

Score-based Generative Models

SDE and ODE

- ScoreSDE $p_t^{\text{SDE}}(\boldsymbol{x}_t)$: $\mathrm{d}\boldsymbol{x}_t = [\boldsymbol{f}(\boldsymbol{x}_t, t) - g(t)^2 \boldsymbol{s}_\theta(\boldsymbol{x}_t, t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\boldsymbol{w}}_t$

- ScoreODE $p_t^{\text{ODE}}(\boldsymbol{x}_t)$: $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{h}_p(\boldsymbol{x}_t, t) := \boldsymbol{f}(\boldsymbol{x}_t, t) - \frac{1}{2}g(t)^2 \boldsymbol{s}_\theta(\boldsymbol{x}_t, t)$

- Trained by minimizing weighted combination of score matching objectives:

$$\mathcal{J}_{\text{SM}}(\theta; \lambda(\cdot)) := \frac{1}{2}\int_0^T \lambda(t) \mathbb{E}_{q_t(\boldsymbol{x}_t)}\left[\|\boldsymbol{s}_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}_t)\|_2^2\right]\mathrm{d}t$$

Score matching is to minimizing (upper bound) KL-divergence of SDEs

Maximum likelihood training of ScoreSDEs

- Score Matching is to maximum likelihood training of **ScoreSDE** (Song, et al, 2021).

$$D_{\mathrm{KL}}(q_0 \parallel p_0^{\mathrm{SDE}}) \leq D_{\mathrm{KL}}(q_T \parallel p_T^{\mathrm{SDE}}) + \mathcal{J}_{\mathrm{SM}}(\theta; g(\cdot)^2)$$

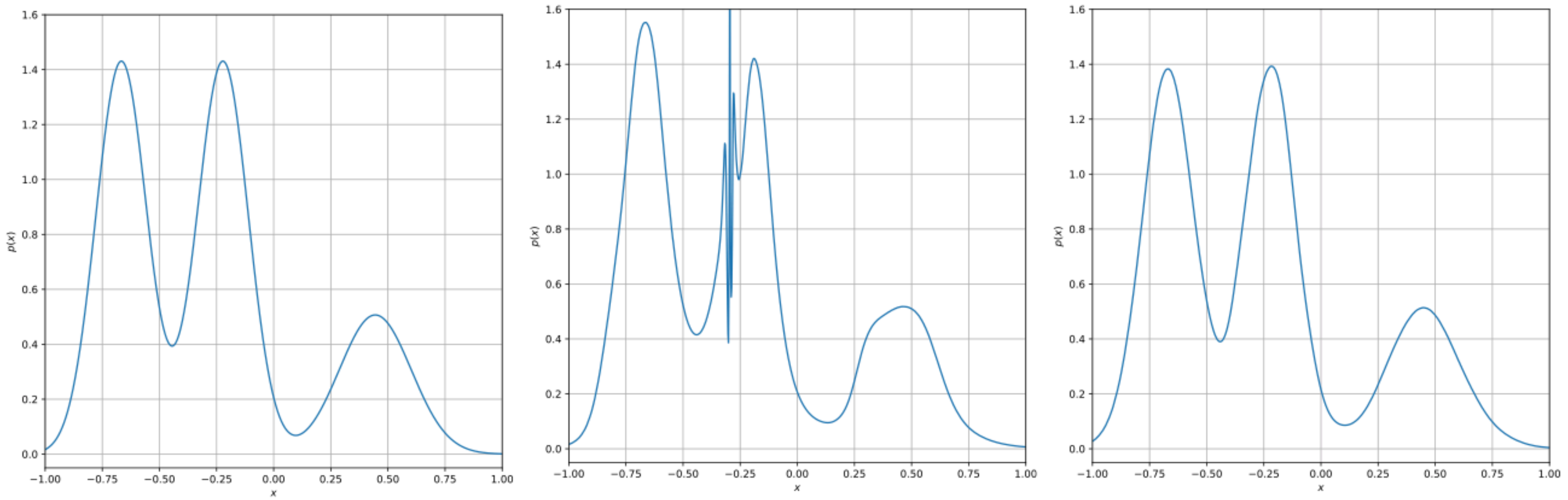
Very small,
$\approx 10^{-5}$

Weighted Score Matching

Problem: Score Matching for ScoreODEs is Unclear

First-order score matching is not enough for ScoreODEs

- An 1-D mixture-of-Gaussian distribution. ScoreODE is “Variance Exploding” type.



(a) Data

(b) ScoreODE (first-order SM)

(c) ScoreODE (third-order SM)

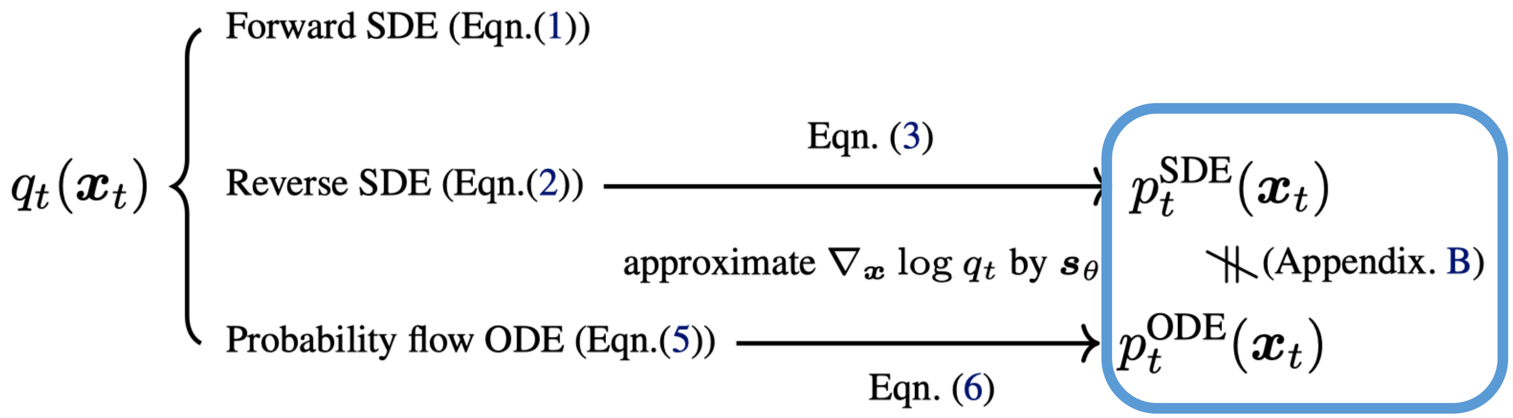
(Song, et al, 2021)

(Ours)

Part I.
Relationship between Score Matching and KL Divergence of ScoreODEs

Relationship between Data Distribution and Score-based Models

The three distributions are different



(a) Relationship between q_t, p_t^{SDE} and p_t^{ODE}.

Proposition 1. (ours, informal). Assume $f(x_t, t)$ is linear w.r.t. x_t , if $p_t^{SDE} = p_t^{ODE}$, then p_t^{SDE} is a Gaussian distribution for all $t \in [0, T]$.

For SGMs trained on the real data, p_t^{SDE} is **always different** from p_t^{ODE} (even if the score model achieves the optimum).

Motivation: Exact Likelihood Computation of ScoreODEs

First-order score matching is not enough for ScoreODEs

- **Theorem**. (Ricky T. Q. Chen et al., 2018) "Instantaneous Change of Variables":

$$\log p_0^{\mathrm{ODE}}(\boldsymbol{x}_0) = \log p_T^{\mathrm{ODE}}(\boldsymbol{x}_T) + \int_0^T \nabla_{\boldsymbol{x}} \cdot \left(\boldsymbol{f}(\boldsymbol{x}_t, t) - \frac{1}{2} g(t)^2 \boldsymbol{s}_\theta(\boldsymbol{x}_t, t) \right) \mathrm{d}t$$

- Score matching can only control $s_\theta(x_t, t)$, but **cannot control** $\boldsymbol{\nabla_x s_\theta(x_t, t)}$ **!**

- A straightforward way: Directly MLE by the above equation?

  No!

  Even for evaluation, computing the likelihood of a **single batch** needs **2~3 minutes**.

KL-Divergence of ScoreODEs

The score matching objective is part of KL-divergence

Theorem 1. (ours, informal) The KL-divergence between data distribution and ScoreODE distribution is:

$$D_{\mathrm{KL}}(q_0 \parallel p_0^{ODE}) = D_{\mathrm{KL}}(q_T \parallel p_T^{ODE}) + \mathcal{J}_{ODE}(\theta)$$

$$= \underbrace{D_{\mathrm{KL}}(q_T \parallel p_T^{ODE}) + \mathcal{J}_{SM}(\theta)}_{\text{upper bound of } D_{\mathrm{KL}}(q_0 \parallel p_0^{SDE}) \text{ in Eqn. (4)}} + \boxed{\mathcal{J}_{Diff}(\theta)}$$

Uncontrolled Error

where

$$\mathcal{J}_{\mathrm{ODE}}(\theta) := \frac{1}{2}\int_0^T g(t)^2 \mathbb{E}_{q_t(\boldsymbol{x}_t)}\left[(\boldsymbol{s}_\theta(\boldsymbol{x}_t,t) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}_t))^\top \boxed{(\nabla_{\boldsymbol{x}} \log p_t^{\mathrm{ODE}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}_t))}\right] \mathrm{d}t,$$

$$\mathcal{J}_{\mathrm{Diff}}(\theta) := \frac{1}{2}\int_0^T g(t)^2 \mathbb{E}_{q_t(\boldsymbol{x}_t)}\left[(\boldsymbol{s}_\theta(\boldsymbol{x}_t,t) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}_t))^\top \boxed{(\nabla_{\boldsymbol{x}} \log p_t^{\mathrm{ODE}}(\boldsymbol{x}_t) - \boldsymbol{s}_\theta(\boldsymbol{x}_t,t))}\right] \mathrm{d}t.$$

Bounding the KL-Divergence of ScoreODEs

Turn MLE to score matching

- By Cauchy–Schwarz inequality:

$$D_{\mathrm{KL}}(q_0 \parallel p_0^{\mathrm{ODE}}) = D_{\mathrm{KL}}(q_T \parallel p_T^{\mathrm{ODE}}) + \frac{1}{2}\int_0^T g(t)^2 \mathbb{E}_{q_t(\boldsymbol{x}_t)}\left[(\boldsymbol{s}_\theta - \nabla \log q_t)^\top (\nabla \log p_t^{\mathrm{ODE}} - \nabla \log q_t)\right] \mathrm{d}t$$

$$\leq D_{\mathrm{KL}}(q_T \parallel p_T^{\mathrm{ODE}}) + \frac{1}{2}\sqrt{\int_0^T g(t)^2 \mathbb{E}_{q_t(\boldsymbol{x}_t)} \|\boldsymbol{s}_\theta - \nabla \log q_t\|_2^2 \mathrm{d}t} \cdot \sqrt{\int_0^T g(t)^2 \mathbb{E}_{q_t(\boldsymbol{x}_t)} \|\nabla \log p_t^{\mathrm{ODE}} - \nabla \log q_t\|_2^2 \mathrm{d}t}$$

(First-Order) Score Matching
(Song, et al, 2021)

Fisher Divergence between ODEs
(**Uncontrolled error**)

Bounding Fisher Divergence by High-Order Score Matchings

First-order, second-order and third-order score matchings

Theorem 2. (ours, informal) Assume $\left\|\nabla_x \log p_t^{ODE}\right\|_2 < C$, then the Fisher divergence between q_t and p_t^{ODE} can be bounded by $U(t; \delta_1, \delta_2, \delta_3, C, q)$, where $\delta_1, \delta_2, \delta_3$ are first-order, second-order and third-order score matching errors:

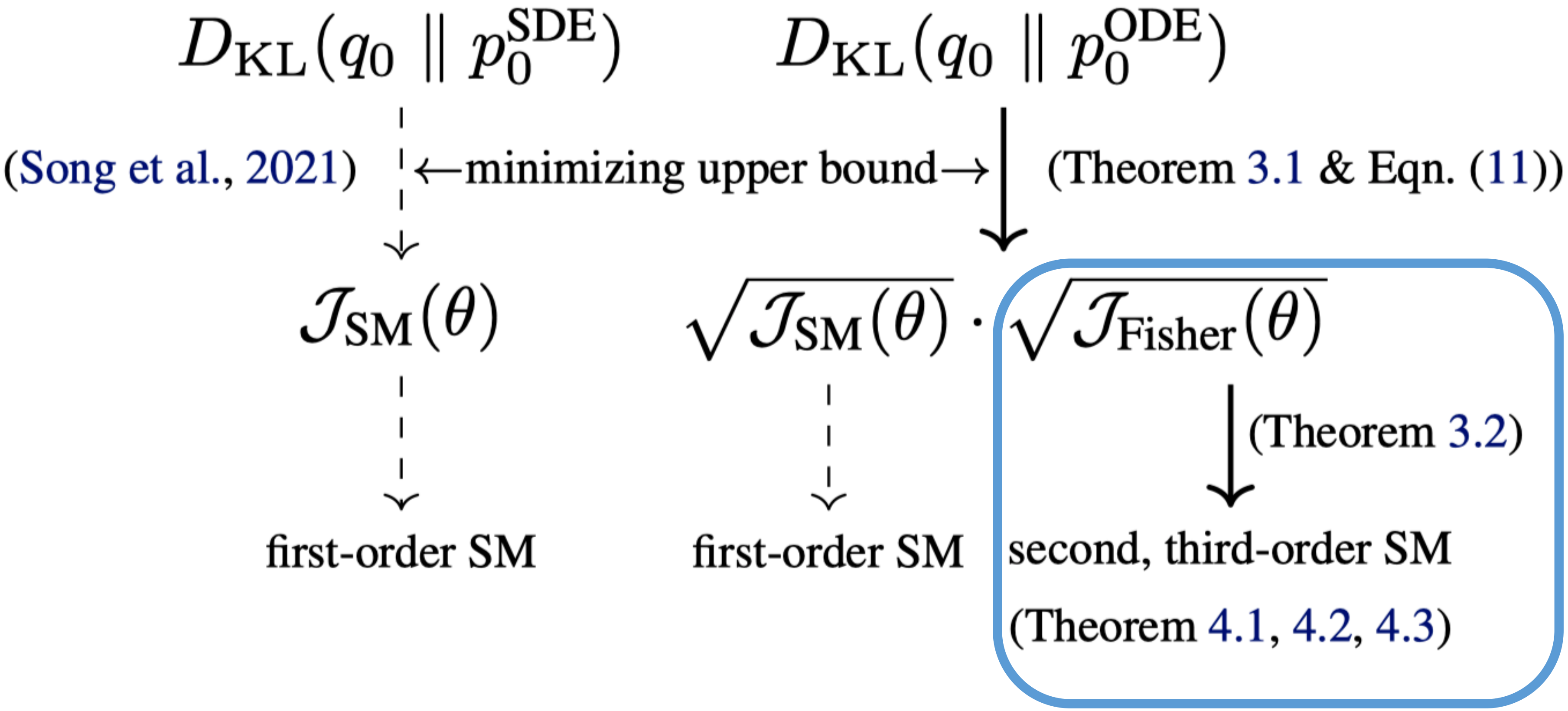
$$\|\boldsymbol{s}_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}_t)\|_2 \leq \delta_1,$$

$$\|\nabla_{\boldsymbol{x}} \boldsymbol{s}_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\|_F \leq \delta_2,$$

$$\|\nabla_{\boldsymbol{x}} \operatorname{tr}(\nabla_{\boldsymbol{x}} \boldsymbol{s}_\theta(\boldsymbol{x}_t, t)) - \nabla_{\boldsymbol{x}} \operatorname{tr}\left(\nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right)\|_2 \leq \delta_3$$

Summary: Relationship between Score Matching and KL Divergence

ScoreSDE and ScoreODE are different



Part II.
Error-Bounded High-Order Denoising Score Matching (DSM)

Second-Order Denoising Score Matching

Second-order score function

- The second-order score function includes the first-order score function:

$$\nabla^2_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t) = \mathbb{E}_{q_{t0}(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\nabla^2_{\boldsymbol{x}_t} \log q_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0) + \nabla_{\boldsymbol{x}_t} \log q_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0) \nabla_{\boldsymbol{x}_t} \log q_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0)^\top\right]$$

$$- \nabla_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t) \nabla_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t)^\top$$

"Second-order noise"
(**Can turn to Denoising**)

First-order score
(**Unkown**)

- A straightforward way (Meng, et al, 2021): replacing the first-order score function $\nabla_x \log q_t(x_t)$ by the approximated score network $\hat{s}_1(x_t, t)$.

Second-Order Denoising Score Matching

Straightforward way

- (Meng et al., 2021) uses the following objective for second-order DSM:

$$\theta^* = \operatorname*{argmin}_\theta \mathbb{E}_{q_t} \mathbb{E}_{q_{t0}} \left[\left\| \boldsymbol{s}_2(\theta) - \nabla^2 \log q_{0t} - \nabla \log q_{0t} \nabla \log q_{0t}^\top + \hat{\boldsymbol{s}}_1 \hat{\boldsymbol{s}}_1^\top \right\|_F^2 \right]$$

However, we show that this method has **unbounded score matching error, even if the training objective achieves the global optimal.**

Proposition 2. (ours, informal) Assume $\nabla_x^2 \log q_t$ is unbounded (e.g. Gaussian distribution), and there exists $\delta_1 > 0$ such that $\|\hat{s}_1 - \log q_t\|_2 > \delta_1$. Then for any $\delta_1 > 0$ and $C > 0$, there always exists x_t such that

$$\|s_2(x_t, t; \theta^*) - \nabla_x \log q_t\,(x_t)\|_F > C$$

Error-Bounded Second-Order Denoising Score Matching

Matrix form

Theorem 3. (ours, informal) Assume $\hat{s}_1$ is an estimation for $\nabla_x \log q_t$, then we can learn a second-order score model $s_2(\theta)$ which minimizes

$$\mathbb{E}_{q_t(\boldsymbol{x}_t)}\left[\left\|\boldsymbol{s}_2(\boldsymbol{x}_t, t; \theta) - \nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right\|_F^2\right]$$

by optimizing

$$\theta^* = \operatorname*{argmin}_\theta \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}}\left[\frac{1}{\sigma_t^4}\left\|\sigma_t^2 \boldsymbol{s}_2(\boldsymbol{x}_t, t; \theta) + \boldsymbol{I} - \boldsymbol{\ell}_1 \boldsymbol{\ell}_1^\top\right\|_F^2\right]$$

where

$$\boldsymbol{\ell}_1(\boldsymbol{\epsilon}, \boldsymbol{x}_0, t) := \sigma_t \hat{\boldsymbol{s}}_1(\boldsymbol{x}_t, t) + \boldsymbol{\epsilon}, \quad \boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

Moreover, the score matching error can be **bounded by the training error and the first-order score matching error:**

$$\left\|\boldsymbol{s}_2(\boldsymbol{x}_t, t; \theta) - \nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right\|_F \le \left\|\boldsymbol{s}_2(\boldsymbol{x}_t, \theta) - \boldsymbol{s}_2(\boldsymbol{x}_t, t; \theta^*)\right\|_F + \delta_1^2(\boldsymbol{x}_t, t)$$

$$\left(\delta_1(\boldsymbol{x}_t, t) := \left\|\hat{\boldsymbol{s}}_1(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}_t)\right\|_2 \right)$$

Error-Bounded Second-Order Denoising Score Matching

Scalar form (matching trace)

Corollary 1. (ours, informal) Assume $\hat{s}_1$ is an estimation for $\nabla_x \log q_t$, then we can learn a second-order score model $s_2^{trace}(\theta)$ which minimizes

$$\mathbb{E}_{q_t(\boldsymbol{x}_t)}\left[\left|\boldsymbol{s}_2^{trace}(\boldsymbol{x}_t, t; \theta) - \mathrm{tr}\left(\nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right)\right|^2\right]$$

by optimizing

$$\theta^* = \operatorname*{argmin}_{\theta} \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}}\left[\frac{1}{\sigma_t^4}\left|\sigma_t^2 \boldsymbol{s}_2^{trace}(\boldsymbol{x}_t, t; \theta) + d - \|\boldsymbol{\ell}_1\|_2^2\right|^2\right]$$

Moreover, the score matching error can be **bounded by the training error and the first-order score matching error:**

$$\left|\boldsymbol{s}_2^{trace}(\boldsymbol{x}_t, t; \theta) - \mathrm{tr}\left(\nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right)\right| \leq \left|\boldsymbol{s}_2^{trace}(\boldsymbol{x}_t, t; \theta) - \boldsymbol{s}_2^{trace}(\boldsymbol{x}_t, t; \theta^*)\right| + \delta_1^2(\boldsymbol{x}_t, t)$$

Error-Bounded Third-Order Denoising Score Matching

Vector form

Theorem 4. (ours, informal) Assume $\hat{s}_1$ is an estimation for $\nabla_x \log q_t$ and $\hat{s}_2$ is an estimation for $\nabla_x^2 \log q_t$, then we can learn a third score model $s_3(\theta)$ which minimizes

$$\mathbb{E}_{q_t(\boldsymbol{x}_t)}\left[\left\|\boldsymbol{s}_3(\boldsymbol{x}_t, t; \theta) - \nabla_{\boldsymbol{x}} \operatorname{tr}\left(\nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right)\right\|_2^2\right]$$

by optimizing

$$\theta^* = \operatorname*{argmin}_\theta \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}}\left[\frac{1}{\sigma_t^6}\left\|\sigma_t^3 \boldsymbol{s}_3(\boldsymbol{x}_t, t; \theta) + \boldsymbol{\ell}_3\right\|_2^2\right]$$

where

$$\boldsymbol{\ell}_1(\boldsymbol{\epsilon}, \boldsymbol{x}_0, t) := \sigma_t \hat{\boldsymbol{s}}_1(\boldsymbol{x}_t, t) + \boldsymbol{\epsilon}, \quad \boldsymbol{\ell}_2(\boldsymbol{\epsilon}, \boldsymbol{x}_0, t) := \sigma_t^2 \hat{\boldsymbol{s}}_2(\boldsymbol{x}_t, t) + \boldsymbol{I},$$

$$\boldsymbol{\ell}_3(\boldsymbol{\epsilon}, \boldsymbol{x}_0, t) := \left(\|\boldsymbol{\ell}_1\|_2^2 \boldsymbol{I} - \operatorname{tr}(\boldsymbol{\ell}_2)\boldsymbol{I} - 2\boldsymbol{\ell}_2\right)\boldsymbol{\ell}_1, \quad \boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

Moreover, the score matching error can be **bounded by the training error and the first-order and second-order score matching errors:**

$$\left\|\boldsymbol{s}_3(\boldsymbol{x}_t, t; \theta) - \nabla_{\boldsymbol{x}} \operatorname{tr}\left(\nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right)\right\|_2 \le \|\boldsymbol{s}_3(\boldsymbol{x}_t, t; \theta) - \boldsymbol{s}_3(\boldsymbol{x}_t, t; \theta^*)\|_2 + \left(\delta_1^2 + \delta_{2,tr} + 2\delta_2\right)\delta_1^2$$

Summary: Error-Bounded High-Order DSM

Bounded by training error and lower-order score matching errors

$$\left\|\boldsymbol{s}_2(\boldsymbol{x}_t, t; \theta) - \nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right\|_F \leq \left\|\boldsymbol{s}_2(\boldsymbol{x}_t, \theta) - \boldsymbol{s}_2(\boldsymbol{x}_t, t; \theta^*)\right\|_F + \delta_1^2(\boldsymbol{x}_t, t)$$

$$\left|\boldsymbol{s}_2^{trace}(\boldsymbol{x}_t, t; \theta) - \operatorname{tr}\left(\nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right)\right| \leq \left|\boldsymbol{s}_2^{trace}(\boldsymbol{x}_t, t; \theta) - \boldsymbol{s}_2^{trace}(\boldsymbol{x}_t, t; \theta^*)\right| + \delta_1^2(\boldsymbol{x}_t, t)$$

$$\left\|\boldsymbol{s}_3(\boldsymbol{x}_t, t; \theta) - \nabla_{\boldsymbol{x}} \operatorname{tr}\left(\nabla_{\boldsymbol{x}}^2 \log q_t(\boldsymbol{x}_t)\right)\right\|_2 \leq \left\|\boldsymbol{s}_3(\boldsymbol{x}_t, t; \theta) - \boldsymbol{s}_3(\boldsymbol{x}_t, t; \theta^*)\right\|_2 + \left(\delta_1^2 + \delta_{2,tr} + 2\delta_2\right)\delta_1^2$$

Part III.
Training Score Models by High-Order DSM

Variance Reduction by Time-Reweighting

The “noise-prediction” trick in (Ho et al. 2020)

- Our training objectives is:

$$\mathcal{J}_{\mathrm{DSM}}^{(1)}(\theta) \coloneqq \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\left\|\sigma_t \boldsymbol{s}_\theta(\boldsymbol{x}_t, t) + \boldsymbol{\epsilon}\right\|_2^2\right]$$

$$\mathcal{J}_{\mathrm{DSM}}^{(2)}(\theta) \coloneqq \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\left\|\sigma_t^2 \nabla_{\boldsymbol{x}} \boldsymbol{s}_\theta(\boldsymbol{x}_t, t) + \boldsymbol{I} - \boldsymbol{\ell}_1 \boldsymbol{\ell}_1^\top\right\|_F^2\right],$$

$$\mathcal{J}_{\mathrm{DSM}}^{(2,\mathrm{tr})}(\theta) \coloneqq \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\left|\sigma_t^2 \operatorname{tr}(\nabla_{\boldsymbol{x}} \boldsymbol{s}_\theta(\boldsymbol{x}_t, t)) + d - \|\boldsymbol{\ell}_1\|_2^2\right|^2\right],$$

$$\mathcal{J}_{\mathrm{DSM}}^{(3)}(\theta) \coloneqq \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\left\|\sigma_t^3 \nabla_{\boldsymbol{x}} \operatorname{tr}(\nabla_{\boldsymbol{x}} \boldsymbol{s}_\theta(\boldsymbol{x}_t, t)) + \boldsymbol{\ell}_3\right\|_2^2\right],$$

$\Rightarrow$

$$\min_\theta \mathcal{J}_{\mathrm{DSM}}^{(1)}(\theta) + \lambda_1 \left(\mathcal{J}_{\mathrm{DSM}}^{(2)}(\theta) + \mathcal{J}_{\mathrm{DSM}}^{(2,tr)}(\theta)\right) + \lambda_2 \mathcal{J}_{\mathrm{DSM}}^{(3)}(\theta),$$

Scale-up to High Dimension

By Hutchinson's trace estimator (Hutchinson, 1989)

- Our training objectives for high-dimensional data are:

$$\mathcal{J}^{(2)}_{\text{DSM,estimation}}(\theta) = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\mathbb{E}_{p(\boldsymbol{v})}\left[\left\|\sigma_t^2 \boldsymbol{s}_{jvp} + \boldsymbol{v} - (\sigma_t \hat{\boldsymbol{s}}_1 \cdot \boldsymbol{v} + \boldsymbol{\epsilon} \cdot \boldsymbol{v})(\sigma_t \hat{\boldsymbol{s}}_1 + \boldsymbol{\epsilon})\right\|_2^2\right],$$

$$\mathcal{J}^{(2,\text{tr})}_{\text{DSM,estimation}}(\theta) = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\mathbb{E}_{p(\boldsymbol{v})}\left[\left|\sigma_t^2 \boldsymbol{v}^\top \boldsymbol{s}_{jvp} + \|\boldsymbol{v}\|_2^2 - |\sigma_t \hat{\boldsymbol{s}}_1 \cdot \boldsymbol{v} + \boldsymbol{\epsilon} \cdot \boldsymbol{v}|^2\right|^2\right],$$

$$\begin{aligned}\mathcal{J}^{(3)}_{\text{DSM,estimation}}(\theta) = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\mathbb{E}_{p(\boldsymbol{v})}\Bigg[\Big\|&\sigma_t^3 \boldsymbol{v}^\top \nabla_{\boldsymbol{x}} \boldsymbol{s}_{jvp} + |\sigma_t \hat{\boldsymbol{s}}_1 \cdot \boldsymbol{v} + \boldsymbol{\epsilon} \cdot \boldsymbol{v}|^2 (\sigma_t \hat{\boldsymbol{s}}_1 + \boldsymbol{\epsilon}) - (\sigma_t^2 \boldsymbol{v}^\top \hat{\boldsymbol{s}}_{jvp} + \|\boldsymbol{v}\|_2^2)(\sigma_t \hat{\boldsymbol{s}}_1 + \boldsymbol{\epsilon}) \\ &- 2(\sigma_t \hat{\boldsymbol{s}}_1 \cdot \boldsymbol{v} + \boldsymbol{\epsilon} \cdot \boldsymbol{v})(\sigma_t^2 \hat{\boldsymbol{s}}_{jvp} + \boldsymbol{v})\Big\|_2^2\Bigg],\end{aligned}$$

Proposition 3. (ours, informal) The training objectives for high-dimensional data can upper bound the corresponding original objectives:

$$\mathcal{J}^{(2)}_{\text{DSM}}(\theta) = \mathcal{J}^{(2)}_{\text{DSM,estimation}}(\theta), \quad \mathcal{J}^{(2,\text{tr})}_{\text{DSM}}(\theta) \leq \mathcal{J}^{(2,\text{tr})}_{\text{DSM,estimation}}(\theta), \quad \mathcal{J}^{(3)}_{\text{DSM}}(\theta) \leq \mathcal{J}^{(3)}_{\text{DSM,estimation}}(\theta)$$

Experiments

Example: 1-D mixture-of-Gaussians

- Denote

$$\ell_{\text{Fisher}}(t) := \frac{1}{2} g(t)^2 D_{\text{F}}(q_t \parallel p_t^{\text{ODE}}),$$

$$\ell_{\text{SM}}(t) := \frac{1}{2} g(t)^2 \mathbb{E}_{q_t(\boldsymbol{x}_t)} \| \boldsymbol{s}_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}_t) \|_2^2,$$

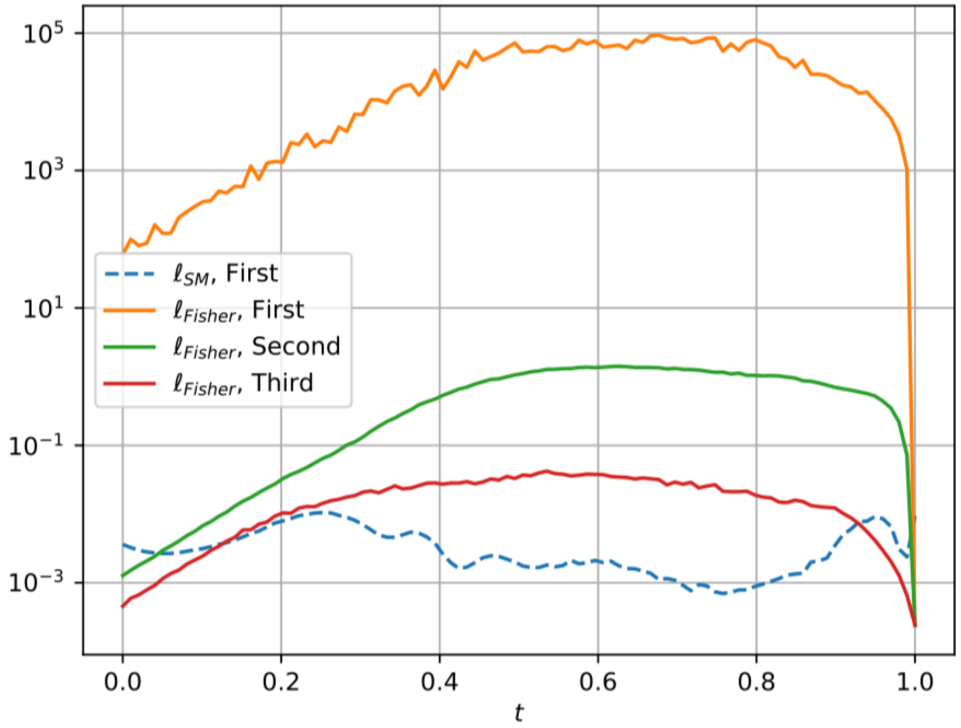
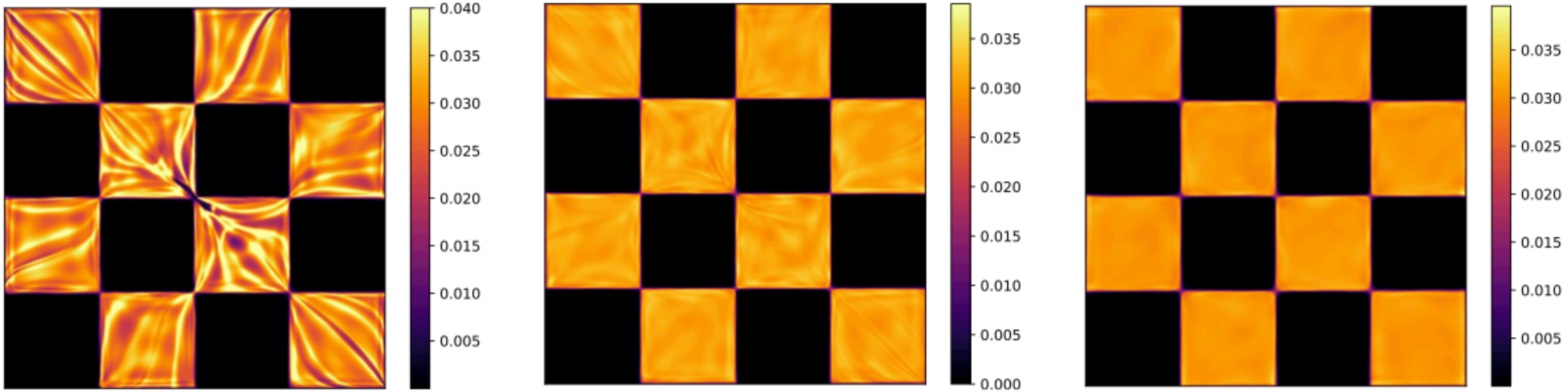


Figure 3. $\ell_{\text{Fisher}}(t)$ and $\ell_{\text{SM}}(t)$ of ScoreODEs (VE type) on 1-D mixture of Gaussians, trained by minimizing the first, second, third-order score matching objectives.

Experiments
Density modeling on 2-D checkerboard data



(a) First-order score matching (b) Second-order score matching (c) Third-order score matching

Figure 4. Model density of ScoreODEs (VE type) on 2-D checkerboard data.

Experiments

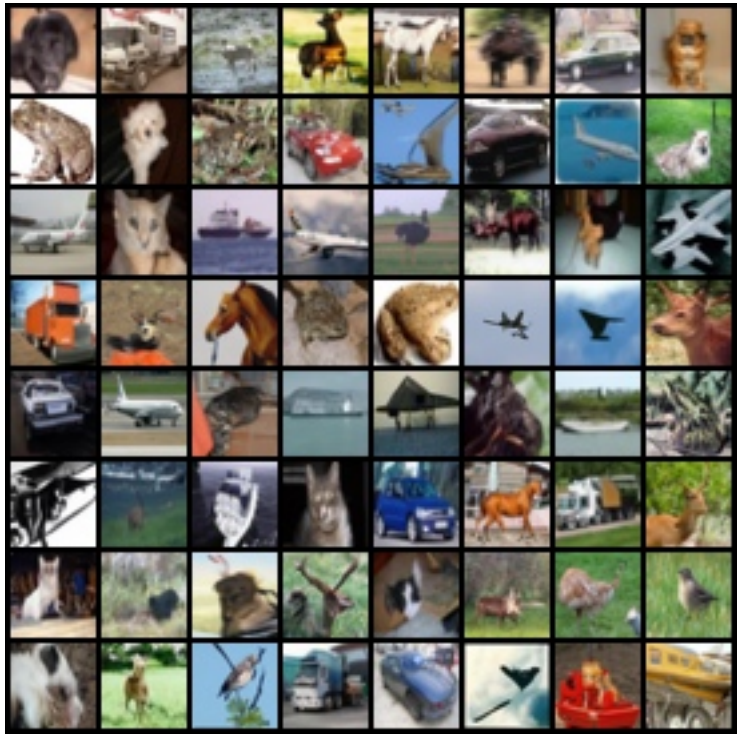
Density modeling on CIFAR-10

Table 1. Negative log-likelihood (NLL) in bits/dim (bpd) and sample quality (FID scores) on CIFAR-10 and ImageNet 32x32.

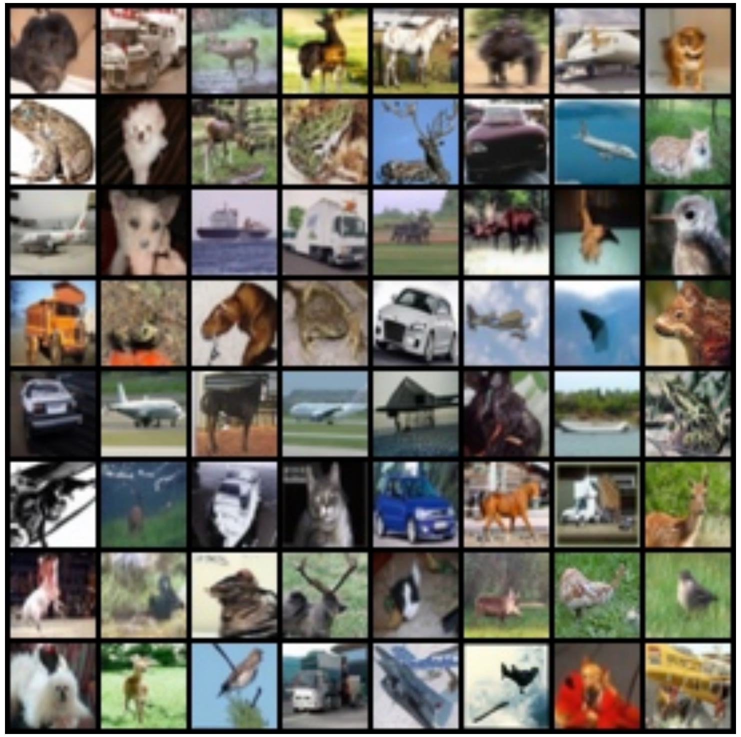
Model	CIFAR-10		ImageNet 32x32
	NLL ↓	FID ↓	NLL ↓
VE (Song et al., 2020)	3.66	2.42	4.21
VE (second) (**ours**)	3.44	**2.37**	4.06
VE (third) (**ours**)	**3.38**	2.95	**4.04**
VE (deep) (Song et al., 2020)	3.45	**2.19**	4.21
VE (deep, second) (**ours**)	3.35	2.43	4.05
VE (deep, third) (**ours**)	**3.27**	2.61	**4.03**

Experiments

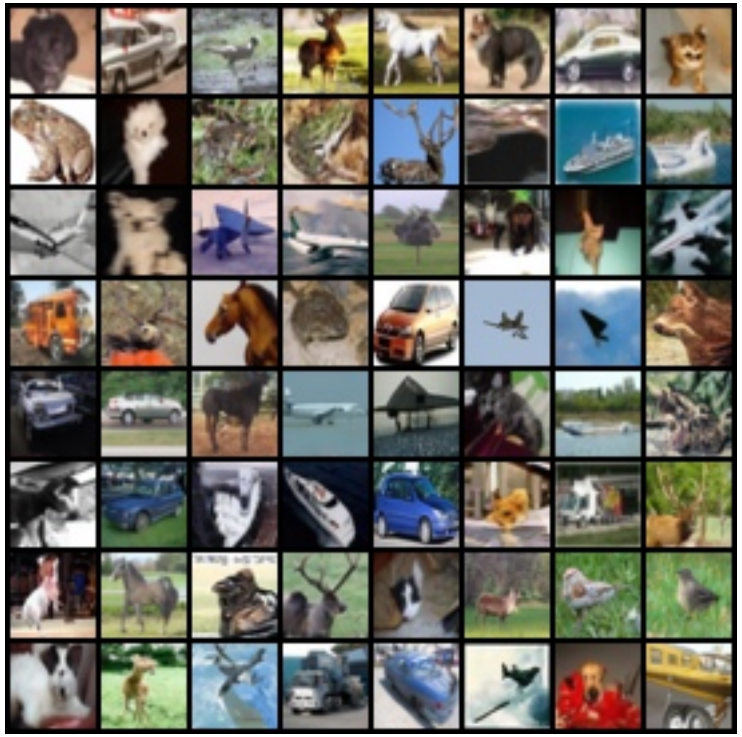
Random samples of SGMs by PC sampler (Song, et al., 2021)



First-Order DSM



Second-Order DSM



Third-Order DSM

Summary and Discussion

- We analyze the relationship between score matching and KL divergence of ScoreODEs, and give an upper bound of KL divergence by high-order score matchings.

- We propose a novel error-bounded high-order denoising score matching method.

- Our proposed method can improve the likelihood of ScoreODEs.