Local Curvature Smoothing with Stein's Identity for Efficient Score Matching

Genki Osada[1], Makoto Shing[2], and Takashi Nishide[3]
[1]LY Corporation [2]Sakana AI [3]Univ. of Tsukuba, Japan

NEURAL INFORMATION PROCESSING SYSTEMS

Background

Estimating the score, $\nabla_x \log p(x)$, enables sampling via Langevin/Hamiltonian Monte Carlo or stochastic/ordinary differential equations. To estimate the score of an unknown distribution $p(x)$, the score function $S_\theta(x)$ is optimized via score matching as:

$$\min_\theta \frac{1}{2}\mathbb{E}_p \| S_\theta(x) - \nabla_x \log p(x) \|^2.$$

Hyvärinen introduced the following as a trainable objective [1]:

$$J_{SM}(\theta) \coloneqq \mathbb{E}_{x\sim p}[J^s_{SM}(\theta, x)], \quad J^s_{SM}(\theta, x) \coloneqq \mathrm{Tr}(\nabla_x S_\theta(x)) + \frac{1}{2}\|S_\theta(x)\|^2.$$

Problem

Computing $\mathrm{Tr}(\nabla_x S_\theta(x))$ for high-dimensional data is computationally expensive, making learning with J_{SM} practically impossible.

Existing Methods

Sliced Score Matching (SSM) [2]

SSM approximates $\mathrm{Tr}(\nabla_x S_\theta(x))$ by the Skilling-Hutchinson trick. Finite Difference Sliced Score Matching (FD-SSM) [3] accelerates it further using the finite difference method.

$$J_{SSM}(\theta) = \mathbb{E}_{x\sim p}\left[\mathbb{E}_{v\sim p_v}[v^{\mathrm{T}} \nabla_x (S_\theta(x)\, v)] + \frac{1}{2}\|S_\theta(x)\|^2\right]$$

where p_v is $\mathcal{N}(0, \mathbb{I}_d)$ or Rademacher dist.

Denoising Score Matching (DSM) [4]

DSM bypasses $\mathrm{Tr}(\nabla_x S_\theta(x))$ computation by replacing $p(x)$ with $q(\tilde{x}) \coloneqq \int q_\sigma(\tilde{x}|x)p(x)dx$, Gaussian perturbed distribution.

$$J_{DSM}(\theta) = \frac{1}{2}\mathbb{E}_{\tilde{x}\sim q_\sigma(\tilde{x}|x)}\mathbb{E}_{x\sim p(x)}\|S_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x)\|^2$$

where $\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x) = \frac{1}{\sigma^2}(x - \tilde{x})$ as $\tilde{x} \sim \mathcal{N}(x, \sigma^2 \mathbb{I}_d)$

To express $\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x)$ in closed form, DSM models p_σ as Gaussian. This imposes a constraint that the drift and diffusion terms of SDE must be affine.

Our Method

We bypass $\mathrm{Tr}(\nabla_x S_\theta(x))$ by combining two lemmas.

Lemma 1. Local curvature smoothing [5]

$$J^s_{LCS}(\theta, x, \sigma) \coloneqq J^s_{SM}(\theta, x) + \frac{1}{2}\sigma^2 \|\nabla_x S_\theta(x)\|_F^2$$
$$= \mathbb{E}_{x'\sim\mathcal{N}(x, \sigma^2\mathbb{I}_d)}[J^s_{SM}(\theta, x')] + \mathcal{O}(\epsilon^2)$$

where $\epsilon \coloneqq \|x - x'\|_2$.

Lemma 2. Stein's Identity for Gaussian [6]

$$\mathbb{E}_{x'\sim Q}[\nabla_{x'} S(x') + S(x') \nabla_{x'} \log Q(x')^T] = 0$$

When $Q(x') = \mathcal{N}(x, \sigma^2 \mathbb{I}_d)$,

$$\mathbb{E}_{x'\sim Q}[\nabla_{x'} S(x')] = \mathbb{E}_{x'\sim Q}\left[\frac{x' - x}{\sigma^2} S(x')\right]$$

Bypassing Jacobian trace

From Lemma 2,

$$\mathbb{E}_{x'\sim\mathcal{N}(x, \sigma^2\mathbb{I}_d)}[\mathrm{Tr}(\nabla_x S_\theta(x'))] = \mathbb{E}_{x'\sim\mathcal{N}(x, \sigma^2\mathbb{I}_d)}\left[S(x')^T \frac{x' - x}{\sigma^2}\right]$$

Objective Function

Putting Lemma 1 into the above,

$$J^s_{LCSS}(\theta, x, \sigma) \coloneqq \mathbb{E}_{x'\sim\mathcal{N}(x, \sigma^2\mathbb{I}_d)}\left[S(x')^T \frac{x' - x}{\sigma^2} + \frac{1}{2}\|S_\theta(x')\|^2\right]$$

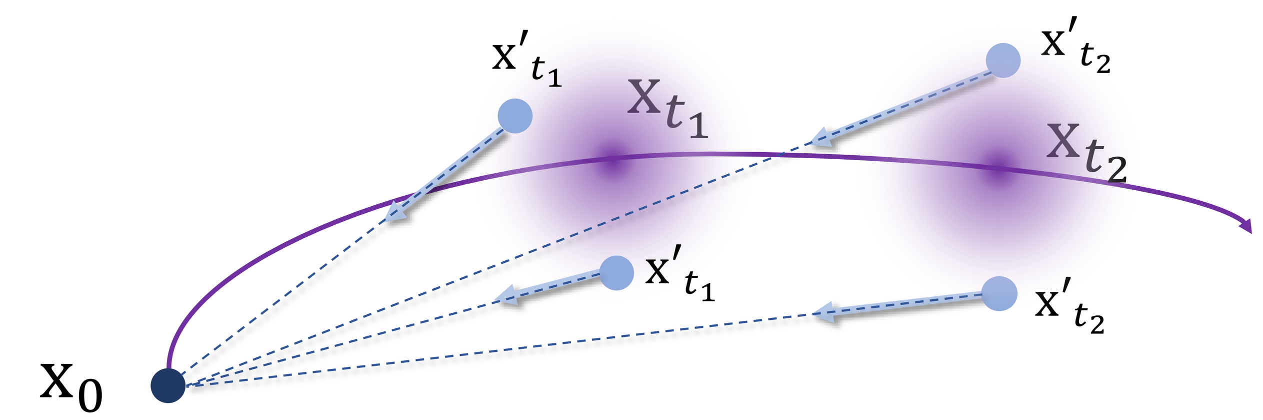
The time-conditional version is :

$$J^s_{LCSS}(\theta, x_0, t) \coloneqq \mathbb{E}_{x'_t\sim\mathcal{N}(x_0, \sigma_t^2\mathbb{I}_d)}\left[S(x'_t, t)^T \frac{x'_t - x_0}{\sigma_t^2} + \frac{1}{2}\|S_\theta(x'_t, t)\|^2\right].$$

The loss function of score-based diffusion model with LCSS is:

$$J_{LCSS}(\theta) \coloneqq \int_0^T \lambda(t) \mathbb{E}_{x_0\sim p_{data}}[J^s_{LCSS}(\theta, x_0, t)]dt.$$

We set $\lambda(t) = g(t)^2$, the drift term of SDE. (i.e., $\lambda(t) = \sigma_t^2$ for VE SDE.) [7]

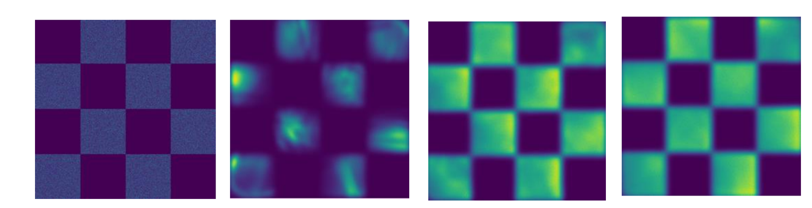


Experimental Results

Elapsed time for model training (ms)↓.

Dataset	Model	SSM	FD-SSM	DSM	LCSS
		Score matching method			
Checkerboard	MLP	497	445	430	**419**
FFHQ	NCSNv2	1838	1367	1381	**1075**

Density estimation.



Data SSM DSM LCSS (Ours)

Right: Samples generated from models trained on CelebA-HQ (1024 × 1024) with LCSS. Model: NCSN++ with VE SDE.

Ablation: the roles of each term by varying the weight γ:

$$\mathbb{E}_{x'_t\sim\mathcal{N}(x_0, \sigma_t^2\mathbb{I}_d)}\left[\gamma S(x'_t, t)^T \frac{x'_t - x_0}{\sigma_t^2} + \frac{1}{2}\|S_\theta(x'_t, t)\|^2\right]$$

$\gamma = 0.5$	$\gamma = 2$	$\gamma = 10$	$\gamma = 10$	$\gamma = 10$
Iter: 300k	Iter: 300k	Iter: 10k	Iter: 300k	Iter: 600k

References

1. Hyvärinen, Aapo, and Peter Dayan. "Estimation of non-normalized statistical models by score matching." Journal of Machine Learning Research 6.4 (2005).
2. Song, Yang, et al. "Sliced score matching: A scalable approach to density and score estimation." Uncertainty in Artificial Intelligence. PMLR, 2020.
3. Pang, Tianyu, et al. "Efficient learning of generative models via finite-difference score matching." Advances in Neural Information Processing Systems 33 (2020): 19175-19188.
4. Vincent, Pascal. "A connection between score matching and denoising autoencoders." Neural computation 23.7 (2011): 1661-1674.
5. Kingma, Durk P., and Yann Cun. "Regularized estimation of image statistics by score matching." *Advances in neural information processing systems* 23 (2010).
6. Gorham, Jackson, and Lester Mackey. "Measuring sample quality with Stein's method." Advances in neural information processing systems 28 (2015).
7. Song, Yang, et al. "Score-Based Generative Modeling through Stochastic Differential Equations." International Conference on Learning Representations. 2020.